# Book of the Short Papers

**Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni**

DRAFT

SIS Società Italiana di Statistica
UNIVERSITÀ POLITECNICA DELLE MARCHE
Dipartimento di Scienze Economiche e Sociali DISES
LIUC Università Cattaneo
LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB
Pearson

CHAIRS
Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, *Rossella Berni,* Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, *Regina Liu,* Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

DRAFT

Copyright © 2023

PUBLISHED BY PEARSON

WWW.PEARSON.COM

*ISBN da definire*

# Contents

**Preface** **XXI**

**1 Plenary Sessions** **1**

Inequality indices: accurate simulation-based inference 2
Maria-Pia Victoria-Feser

Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications 7
Christopher K. Wikle, Likun Zhang, Myungsoo Yoo and Xiaoyu Ma

Demographic change and sustainability: novel approaches from digital and computational demography n.a.
Emilio Zagheni

**2 Invited Sessions** **14**

Machine learning in the design, analysis and integration of sample surveys

Causal Discovery for complex survey data 15
Paola Vicard

Data Integration without conditional independence: a Bayesian Networks approach 21
Pier Luigi Conti, Paola Vicard and Vincenzina Vitale

Mass imputation through Machine Learning techniques in presence of multi-source data 27
Fabrizio De Fausti, Marco Di Zio, Romina Filippini and Simona Toti

Machine learning: different uses and perspectives

Evaluation of pollution containment policies in the US and the role of machine learning algorithms 32
Marco Di Cataldo, Margherita Gerolimetto, Stefano Magrini and Alessandro Spiganti

DRAFT

Machine Learning for Official Statistics: An Application on External Trade n.a.
Mauro Bruno, Maria Serena Causo, Alessio Guandalini, Francesco Ortame and Silvia Russo

## Statistical Machine Learning for environmental applications

Gaussian Processes and Deep Neural Networks for Spatial Prediction 38
Alex Cucco, Luigi Ippoliti, Nicola Pronello, Pasquale Valentini and Carlo Zaccardi

How can we explain Random Forests in a spatial framework? 42
Natalia Golini, Luca Patelli and Xavier Barber

Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns 48
Jorge Mateu and Abdollah Jalilian

## Statistical Process Monitoring for Complex Data in Industry 4.0

A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts 53
Konstantinos Bourazas, Konstantinos Fokianos, Christos Panayiotou and Marios Polycarpou

Anomaly Detection in Circular Data 58
Houyem Demni and Giovanni C. Porzio

## Advances in Data Science and Statistical Learning [IMS Invited Session]

Empirical Bayes approximation of Bayesian learning: understanding a common practice n.a.
Sonia Petrone

Generalized Fiducial Inference on Differentiable Manifolds - a geometric perspective n.a.
Jan Hannig

Model-free bootstrap and conformal prediction in regression n.a.
Dimitris Politis

## ENBIS Session: System Maintenance, Boosting algorithms for regression, and Research Excellence

Boosting Diversity in Regression Ensembles 64
Mathias Bourel, Jairo Cugliari, Yannig Goude and Jean-Michel Poggi

How ENBIS has contributed to the UK Universities Research Excellence Framework 66
Shirley Coleman

## Population Dynamics, Climate Change and Sustainability

Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data 70
Côme Cheritel, Roman Hoffmann and Raya Muttarak

Environmental and socio-demographic determinants of under-5 mortality in India: A survival analysis based on Indian DHS data 71
Joseph Vinod and Joseph Kannankeril

DRAFT

The impact of temperature on expressed sentiment by migration status: Evidence from geo-located Twitter data 76
Risto Conte Keivabu and Jisu Kim

## Statistical Learning for health research and omics data

An alternative to the Dirichlet-multinomial regression model for microbiome data analysis 87
Roberto Ascari, Sonia Migliorati and Andrea Ongaro

Modelling ordinal response to treatment in a real-world cohort study 93
Marco Alfò, Maria Francesca Marino and Silvia D'Elia

On the application of the symmetric graphical lasso for paired data 97
Saverio Ranciati and Alberto Roverato

## The Economic behaviour of Sustainability

Airports performances and sustainable practices. An empirical study on Italian data 102
Riccardo Gianluigi Serio, Maria Michela Dickson, Diego Giuliani and Giuseppe Espa

Sustainability: still an undefined concept for Italians n.a.
Raffaele Angelone and Andrea Marletta

Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability 108
Beatrice Biondi and Mario Mazzocchi

## Advances in statistical methods for complex problems

Inferring multiple treatment effects from observational studies using confounder importance learning n.a.
Omiros Papaspiliopoulos

Path analysis in Ising models: an application to cyber-security risk assessment 113
Monia Lupparelli and Giovanni M. Marchetti

Causal Regularization n.a.
Lucas Kania and Ernst Wit

## Explainable machine learning models

Enhancing Markowitz model: inspection of correlations and tail covariances 119
Gloria Polinesi

Objective and subjective dimension of economic well-being: an approach based on statistical matching 125
Daniela Marella, Vincenzina Vitale and Pierpaolo D'Urso

Sustainable, Accurate, Fair and Explainable Machine Learning Models n.a.
Paolo Giudici and Emanuela Raffinetti

## Flexible Learning for Environmental Sustainability

Comparison of traffic flow data sources for air pollution modelling 131
Theresa Smith and Nick McCullen

DRAFT

Data analysis of photogrammetry-based mapping: the sea cucumbers in the Giglio Island as a case-study 136
Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio and Alessio Pollice

Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive 142
Nicole Augustin, Heike Puhlmann and Simon Trust

## Inequalities in higher education outcomes: learning from data

Inequalities in international students mobility 149
Kristijan Breznik, Giancarlo Ragozini and Marialuisa Restaino

Uncovering the interplay of territorial, socioeconomic, and demographic factors in high school to university transition n.a.
Vincenzo Giuseppe Genova, Andrea Priulla and Martina Vittorietti

## Statistical Learning of demographic and health dynamics

Estimating the impact of a vaccine mandate: the case of measles in Italy n.a.
Chiara Chiavenna

Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth n.a.
Andrea Nigri

Nowcasting Daily Population Displacement in Ukraine through Social Media Advertising Data n.a.
Claire Dooley, Ridhi Kashyap, Douglas Leasure and Francesco Rampazzo

## Challenges towards Fairness and Transparency for Data Processes, Algorithms and Decision-Support Models

A new measure of discrimination in machine learning algorithms 155
Roberta Pappadà and Francesco Pauli

Challenges on Ethics, and Privacy in AI Applications to Fintech 161
Catarina Silva, Joana Matos Dias and Bernardete Ribeiro

Uncertainty and fairness metrics 166
Anna Gottard

## Educational Data mining: methods for complex data in students' assessment

Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring 172
Michela Battauz

Predicting high schools' students performances with registry's data: a machine learning approach 177
Lidia Rossi, Marta Cannistrà and Tommaso Agasisti

Using response times to identify cheaters in CAT: A simulation study 181
Luca Bungaro, Bernard P. Veldkamp and Mariagiulia Matteucci

DRAFT

## Spatial and Spatio-Temporal Modeling: Theory and Applications

A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy 186
Paolo Maranzano, Kelly McConville , Philipp Otto and Felicetta Carillo

Minimum contrast for point processes' first-order intensity estimation 192
Nicoletta D'Angelo and Giada Adelfio

## Statistical Framework for Measuring the Sustainability of Tourism

Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model 197
Claudia Cappello, Sabrina Maggio and Sandra De Iaco

Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data 201
Ida Camminatiello, Rosaria Lombardo, Jean-Francois Durand and Leonardo S. Alaimo

## Statistical learning for well-being analysis

Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach 207
Mariateresa Ciommi, Chiara Gigliarano, Francesca Mariani and Gloria Polinesi

The fuzzy set approach as statistical learning for the analysis of multidimensional well-being 213
Gianni Betti, Federico Crescenzi, Antonella D'Agostino and Laura Neri

What Makes a Satisfying Life? Prediction and Interpretation with Machine-Learning Algorithms n.a.
Conchita D'Ambrosio

## Bayesian contributions to Statistical Learning

A Bayesian framework for early cancer screening 218
Sally Paganin and Jeff Miller

Imputing Synthetic Pseudo Data from Aggregate Data: Development and Validation for Precision Medicine n.a.
Cecilia Balocchi

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach 223
Filippo Ascolani and Valentina Ghidini

## Data Visualization for Smart Insights and Advanced Predictive Analytics

Applications of data visualization for industry 228
Martina Dossi, Stefano Sangaletti, Marilena Di Bari and Federica Bruschini

Some Notes on the Use of the Circular Boxplot 234
Giovanni Camillo Porzio and Davide Buttarazzi

TERRA: a smart visualization tool for international trade in goods statistics 235
Francesco Amato, Mauro Bruno and Maria Serena Causo

DRAFT

Methods for the analysis of distributional data

Clustering of Distributional Data based on LDQ transformation 241
Gianmarco Borrata and Rosanna Verde

Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis 246
Francesca Fortuna, Fabrizio Maturo and Tonio Di Battista

Multivariate Parametric Analysis of Distributional Data n.a.
Paula Brito

Migrants and Refugees in Europe: social, economic and health-related issues

Social networks and loneliness among older migrants in Italy 252
Viviana Amati, Eralba Cela and Elisa Barbiano di Belgiojoso

The Italian Decree on Security: An Analysis of the Impact on Asylum Applications 257
Giorgio Piccitto

Modelling and Forecasting High-dimensional time series

Adaptive combinations of tail-risk forecasts 263
Alessandra Amendola, Vincenzo Candila, Antonio Naimoli and Giuseppe Storti

Are Monetary Policy Announcements related to Volatility Jumps? 269
Giampiero Gallo, Demetrio Lacava and Edoardo Otranto

Regularized Estimation and Prediction of the El Nino/Southern Oscillation Cycle 275
Alessandro Giovannelli and Tommaso Proietti

**3 Contributed Sessions 276**

Bayesian nonparametric methods

Bayesian density estimation for modeling age-at-death distribution 277
Davide Agnoletto, Tommaso Rigon and Bruno Scarpa

Bayesian mixing distribution estimation in the Gaussian-smoothed 1-Wasserstein distance 282
Catia Scricciolo

Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models 287
Francesco Denti, Antonio Di Noia and Antonietta Mira

Bayesian nonparametric multiple change point detection for time series of compositional data 293
Edoardo Marchionni and Riccardo Corradin

Galton-Watson process: a non parametric prior for the offspring distribution through generalised linear mixed models 299
Massimo Cannas, Michele Guindani and Nicola Piras

DRAFT

Hierarchical processes in survival analysis 304
Riccardo Cogo, Federico Camerlenghi and Tommaso Rigon

## Economics and Statistics

Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach 310
Yuri Calleo, Simone Di Zio and Francesco Pilla

Stocks price forecasts using Stochastic Differential Equations: an empirical assessment 316
Dario Frisardi and Matteo Spuri

The Added-Worker Effect within Italian Households 322
Donata Favaro and Anna Giraldo

## Health statistics 1

A model for the natural history of breast cancer: application to a Norwegian screening dataset 326
Laura Bondi, Marco Bonetti and Solveig Hofvind

Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data 331
Elena Ballante

Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach 336
Chiara Masci, Marta Spreafico and Francesca Ieva

Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data 342
Ilaria Bombelli and Daniele De Rocchi

Tracking attitudes towards COVID vaccines: A text mining analysis 348
Leonardo Scarso, Marco Novelli and Francesco Saverio Violante

Treatment effect assessment in observational studies with multi-level treatment and outcome 354
Federica Cugnata, Paola Vicard, Paola M.V. Rancoita, Fulvia Mecatti, Clelia Di Serio and Pier Luigi Conti

## Indicators: composition, uses and limitations

Are European consumers willing to pay the true price for sustainable food? 360
Luca Secondi and Mengting Yu

Can the reliability of composite indexes be impacted by uncertainty of individual indicators? 367
Caterina Giusti, Stefano Marchetti and Vincenzo Mauro

Initial Coin Offerings and ESG: allies or enemies? 372
Alessandro Bitetto and Paola Cerchiello

DRAFT

On the impact of intraclass correlation in the ANVUR evaluation of academic departments 378
Giorgio Edoardo Montanari and Marco Doretti

Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes 383
Luigi Biggeri, Stefano Marchetti, Caterina Giusti, Monica Pratesi, Francesco Schirripa Spagnolo and Gaia Bertarelli

Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis 389
Camilla Salvatore, Annamaria Bianchi and Silvia Biffignandi

## Multivariate data analysis 1

Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data 395
Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data 401
Farah Naz and Elena Ballante

ROBOUT: a multi-step methodology for conditional outlier detection 406
Matteo Farnè and Angelos Vouldis

Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments 412
Rosamarie Frieri, Alessandro Baldi Antognini, Maroussa Zagoraiou, and Marco Novelli

Separation scores: a new statistical tool for scoring and ranking partially ordered data 418
Marco Fattore

## Statistics in Society 1

Community detection analysis with robin on hashtag network 424
Valeria Policastro, Francesco Santelli and Giancarlo Ragozini

Film Tourism Motivation through the lens of Trip Advisor data 430
Nicolò Biasetton, Marta Disegna, Girish Prayag and Elena Barzizza

Life satisfaction and social activities in later life in Italy: a focus on the Internet use 436
Claudia Furlan and Silvia Meggiolaro

Social capital endowment's role in the intergenerational transmission of education 441
Alessandra Trimarchi, Maria Gabriella Campolo and Antonino Di Pino Incognito

Streaming Data from Social Networks to Track Political Trends 446
Emiliano del Gobbo and Barbara Cafarelli

DRAFT

The scientific production on gender dysphoria: a bibliometric analysis 451
Maria Gabriella Grassia, Marina Marino, Massimo Aria, Rocco Mazza, Luca D'Aniello and Agostino Stavolo

## Assessment and Education

A latent variable approach to Millennials' knowledge of green finance 456
Maria Iannario, Alessandra Tanda and Claudia Tarantola

Archetypal analysis and latent Markov models: A step-wise approach 462
Lucio Palazzo, Rosa Fabbricatore and Francesco Palumbo

From high school to university: academic intentions and enrolment of foreign students in Italy 468
Francesca Di Patrizio, Eleonora Trappolini and Cristina Giudici

Growth models for the progress test in Italian dentistry degree program 473
Giulio Biscardi, Leonardo Grilli, Carla Rampichini, Laura Antonucci and Corrado Crocetta

The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions 477
Francesco Santelli, Teresa Gentile, Davide Bizjak and Lorenzo Fattori

Working Students and job market outcomes: Insights from the University of Florence 482
Gabriele Lombardi, Valentina Tocchioni and Alessandra Petrucci

## Bayesian methods and applications 1

Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation 488
Elena Sabbioni, Enrico Bibbona, Gianluca Mastrantonio and Guido Sanguinetti

Approximate Bayesian Computation for Probabilistic Damage Identification 494
Cecilia Viscardi, Silvia Monchetti, Luisa Collodi, Gianni Bartoli, Michele Betti, Michele Boreale and Fabio Corradi

Estimation of scientific productivity with a hierarchical Bayesian model 500
Maura Mezzetti and Ilia Negri

Heat waves and free-knots splines 505
Gioia Di Credico and Francesco Pauli

The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector 510
Marco Dalla Pria and Silvia Montagna

## Health and mortality

A novel definition of comorbidity based on the Global Burden of Diseases project weights 516
Angela Andreella, Lorenzo Monasta and Stefano Campostrini

DRAFT

An Age-Period-Cohort model of gender gap in youth mortality 522
Giacomo Lanfiuti Baldi and Andrea Nigri

Kinlessness in adult and old age across Europe 528
Marta Pittavino, Bruno Arpino and Elena Pirani

Parameter orthogonalization for Siler mortality model 534
Claudia Di Caterina and Lucia Zanotto

Pseudo-observations in survival analysis 540
Marta Cipriani, Alfonso Piciocchi, Valentina Arena and Marco Alfò

Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity 545
Alessandro Feraldi, Cristina Giudici and Nicolas Brouard

Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence 549
Micaela Arcaio and Anna Maria Parroco

## Mixture Models

An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks 555
Dalila Failli, Maria Francesca Marino and Francesca Martella

Constrained Mixtures of Generalized Normal Distributions 561
Pierdomenico Duttilo, Alfred Kume and Stefano Antonio Gattone

Mixture-based clustering with covariates for ordinal responses 567
Kemmawadee Preedalikit, Daniel Fernàndez, Ivy Liuc, Louise McMillan, Marta Nai Ruscone and Roy Costilla

Partial membership models for soft clustering of multivariate count data 573
Emiliano Seri, Thomas Brendan Murphy and Roberto Rocci

Regression for mixture models for extremes 579
Viviana Carcaiso, Ilaria Prosdocimi and Isadora Antoniano-Villalobos

Robust matrix-variate mixtures of regressions 585
Salvatore Daniele Tomarchio and Michael P. B. Gallaugher

## Sampling methods and analysis of survey data

On the use of auxiliary information to define the sampling design for large-scale geospatial data 591
Chiara Bocci and Emilia Rocco

Optimal joint inclusion probabilities for spatial sampling n.a.
Giuseppe Arbia, Piero Demetrio Falorsi and Vincenzo Nardelli

Robustness and Balance of Sampling or Experimental Designs and Mixture of Designs 597
Yves Tillé and Ejub Talovic

Robustness Bounds for Sampling and Experimental Designs 604
Ejub Talovic and Yves Tillé

Statistical Matching: Hotdeck or Propensity Score? 611
Elena Dalla Chiara, Marcello D'Orazio and Federico Perali

DRAFT

## Space-time statistics

A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis 617
Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Daria Valente and Marco Gamba

An integrated space-time model to evaluate the innovation drivers in Italy 623
Emma Bruno, Rosalia Castellano and Gennaro Punzo

Revealing the dynamic relations between traffic and crowding using big data from mobile phone network 629
Selene Perazzini, Rodolfo Metulini and Maurizio Carpita

SMaC: Spatial Matrix Completion method 635
Giulio Grossi, Alessandra Mattei and Georgia Papadogeorgou

The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks 640
Andrea Gilardi and Riccardo Borgoni

## Clustering and classification 1

A clustering model for flow data: an application to international student mobility 646
Cinzia Di Nuzzo and Donatella Vicari

Contingency tables with structural zeros and discrete copulas 651
Roberto Fontana, Elisa Perrone and Fabio Rapallo

Levels Merging in the Latent Class Model 657
Christophe Biernacki

Model-based clustering of count processes with multiple change 663
Shuchismita Sarkar and Xuwen Zhu

Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data 667
Jana Cibulková, Zdeněk Šulc, Hana Řezanková and Jaroslav Horníček

Spectral clustering of mixed data via association-based distance n.a
Alfonso Iodice D'Enza, Francesco Palumbo and Cristina Tortora

## Dynamic models and time series

A multivariate hidden semi-Markov model for the analysis of multiple air pollutants 673
Marco Mingione, Pierfrancesco Alaimo Di Loro, Francesco Lagona and Antonello Maruotti

A smooth transition autoregressive model for matrix-variate time series 679
Andrea Bucci

Dynamic network models with time-varying nodes 685
Luca Gherardini, Mauro Bernardi and Monia Lupparelli

Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling 691
Ivan Sciascia, Andrea Crosino and Andrea Genre

DRAFT

Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models 696
Mirko Armillotta

## Environmental learning and indicators

Assessing the performance of nuclear norm-based matrix completion methods on $CO_2$ emissions data 702
Rodolfo Metulini, Francesco Biancalani, Giorgio Gnecco and Massimo Riccaboni

Deep Learning for smart and sustainable agriculture 708
Amalia Vanacore, Armando Ciardiello, Annalisa Izzo, Pierdomenico Zaffino, Carolina Vecchio, Gennaro Pio Auricchio and Luigi Uccelli

Do green transition, environmental taxes and renew-able energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression 714
Aamir Javed, Agnese Rapposelli and Asif Javed

Doubly Robust DID for National Parks evaluation: "just" environmental benefits, or socioeconomics impacts as well? 721
Riccardo D'Alberto, Francesco Pagliacci and Matteo Zavalloni

On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us? 727
Lorenzo Mori and Maria Rosaria Ferrante

Small scale analysis of energy vulnerability in the municipality of Palermo 732
Giuliana La Mantia

## Health statistics 2

A test for non-differential misclassification error in database epidemiological studies 738
Giorgio Limoncella, Leonardo Grilli, Emanuela Dreassi, Carla Rampichini, Robert Platt and Rosa Gini

Is the COVID-19 'color code' of Italian regions subjected to political manipulation? 742
Giovanni Busetta and Fabio Fiorillo

Modelling multilevel ordinal response under endogeneity: application to DTC patients' outcome 748
Silvia D'Elia

Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques 753
Nicole Fontana, Laura Savaré and Francesca Ieva

Optimal two-stage design based on error rates under a Bayesian perspective 759
Susanna Gentile and Valeria Sambucini

## Migrants in Italy and return migration

Comparing migrant and "native" Italian adolescents in risky behaviours from FSS and SHARE Corona surveys n.a.
Daniela Foresta

DRAFT

EU-Border crisis: narratives, sentiments and misinformation about migration-related events on Twitter 765
Elena Ambrosetti, Cecilia Fortunato and Sara Miccoli

Graduates' interregional migration in times of crisis: the Italian case 769
Thaís García-Pereiro, Ivano Dileo and Anna Paterno

Intentions to stay: The experience of return migrants in Albania 774
Maria Carella, Thaís García-Pereiro, Roberta Pace and Anna Paterno

Return migration to home country: a systematic literature review with text mining and topic modelling 779
Cecilia Fortunato, Andrea Iacobucci and Elena Ambrosetti

The allocation of time within native and foreign couples living in Italy 786
Giovanni Busetta, Maria Gabriella Campolo and Antonino Di Pino Incognito

The foreigners' contribution to fertility by Italian provinces 792
Eleonora Miaci, Cristina Giudici, Eleonora Trappolini, Marina Attili, Cinzia Castagnaro and Antonella Guarneri

## Sustainability assessment

ESG, sustainability and stock market risk 797
Michele Costa

Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment 801
Gloria Solano-Hermosilla, Jesus Barreiro-Hurle and Ilaria Amerise

Sustainability explained by ChatGPT artificialintelligence in a HITL perspective: innovative approaches 807
Vito Santarcangelo, Angelo Lamacchia, Emilio Massa, Saverio Gianluca Crisafulli, Massimiliano Giacalone and Vincenzo Basile

Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques 813
Massimo Gastaldi, Ginevra Virginia Lombardi, Agnese Rapposelli and Giulia Romano

Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism 819
Francesca Bassi, Josè Fernando Vera and Juan Antonio Marmolejo Martìn

Sustainable tourism: a survey on the propensity towards eco-friendly accommodations 825
Claudia Furlan and Giovanni Finocchiaro

## Bayesian methods and applications 2

A comparison of computational approaches for posterior inference in Bayesian Poisson regression 829
Laura D'Angelo

Bias-reduction methods for Poisson regression models 834
Luca Presicce, Tommaso Rigon and Emanuele Aliverti

DRAFT

Finite Mixture Model for Multiple Sample Data 839
Alessandro Colombi, Raffaele Argiento, Federico Camerlenghi and Lucia Paci

On Bayesian power analysis in reliability 844
Fulvio De Santis, Stefania Gubbiotti and Francesco Mariani

Power priors elicitation through Bayes factors 849
Roberto Macrì Demartino, Leonardo Egidi and Nicola Torelli

Predictive Bayes factors 855
Leonardo Egidi and Ioannis Ntzoufras

## Clustering and classification 2

A Clusterwise Regression Method for Distributional-Valued Data 861
Antonio Balzanella, Rosanna Verde and Francisco de A.T. de Carvalho

A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels 865
Alessandra Ragni, Chiara Masci, Francesca Ieva and Anna Maria Paganoni

Introducing a novel directional distribution depth function for supervised classification 871
Edoardo Redivo and Cinzia Viroli

Clustering alternatives in the preference-approval context 876
Alessandro Albano, José Luis Garcia-Lapresta , Mariangela Sciandra and Antonella Plaia

Computational assessment of k-means clustering on a Structural Equation Model based index 881
Mariaelena Bottazzi Schenone, Elena Grimaccia and Maurizio Vichi

Handling missing data in complex phenomena: an ultrametric model-based approach for clustering 887
Francesca Greselin and Giorgia Zaccaria

## Economics and labour markets

A multivariate ranking analysis on the employability of young adults 893
Rosa Arboretti, Elena Barzizza, Nicolo Biasetton, Riccardo Ceccato, Monica Fedeli and Concetta Tino

Analysis of the Gender Pay Gap in the Italian Labour Market 899
Giulia Cappelletti and Daniele Toninelli

Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes 905
Lorenzo Giammei

Patterns of flexible employment careers. Does measurement error matter? 911
Mauricio Garnier-Villarreal, Dimitris Pavlopoulos and Roberta Varriale

Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention 917
Ulpiani Kocollari, Fabio Demaria and Maddalena Cavicchioli

DRAFT

The CAP instruments impact on GVA and employment: a multivalued treatment approach 923
Montezuma Dumangane and Marzia Freo

The determinants of leaving the parental home in Italy: 2012-18 929
Ilaria Rocco and Gianpiero Dalla Zuanna

## Environmental modeling

A Bayesian weather-driven spatio-temporal model for PM10 in Lombardy 935
Michela Frigeri, Alessandra Guglielmi and Giovanni Lonati

A preliminary study on shape descriptors for the characterization of microplastics ingested by fish 941
Greta Panunzi, Tommaso Valente, Marco Matiddi and Giovanna Jona Lasinio

Artificial neural network in predicting odour concentrations: a case study 947
Veronica Distefano and Gideon Mazuruse

Bayesian analysis of PM10 concentration by spatio-temporal ARIMA and STS models 952
Michela Frigeri and Ilenia Epifani

Functional ANOVA to monitor yearly Adriatic sea temperature variations 958
Annalina Sarra, Adelia Evangelista, Tonio Di Battista and Nicola Di Deo

New perspectives in the measurement of biodiversity 964
Linda Altieri, Daniela Cocchi and Massimo Ventrucci

## Multivariate data analysis 2

Feature Selection via anomaly detection autoencoders in radiogenomics studies 970
Alessia Mapelli, Michela Carlotta Massi, Nicola Rares Franco, Francesca Ieva, Catharine West, Petra Seibold, Jenny Chang-Claude and the REQUITE and RADprecise Consortia

Further considerations on the Spectral Information Criterion 976
Luca Martino

How to increase the power of the test in sparse contingency tables: a simulation study 983
Federica Nicolussi and Manuela Cazzaro

Latent event history models for quasi-reaction systems 989
Matteo Framba, VeronicaVinciotti and Ernst Wit

Quantile-based graphical models for continuous and discrete variables 995
Luca Merlo, Marco Geraci and Lea Petrella

The logratio Student t distribute 1001
Gianna Monti and Gloria Mateu-Figueras

## Statistics in Society 2

A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period 1005
Mauro Mussini

DRAFT

Bayesian networks as a territorial gender impact assessment tool 1010
Flaminia Musella, Lorenzo Giammei, Fulvia Mecatti and Paola Vicar

Can statistics be helpful in detecting electoral fraud? 1014
Massimo Attanasio, Vincenzo G. Genova and Michele Tumminello

Companies' sustainability disclosure and contrast to hunger: the role of social inclusion 1019
Chiara Di Maria and Rodolfo Damiano

Passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017 1025
Riccardo Ievoli, Lucio Palazzo and Giancarlo Ragozini

## Bayesian methods and applications 3

An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior 1031
Paolo Onorati and Brunero Liseo

Binomial Extended Stochastic Block Model for Brain Networks 1037
Valentina Ghidini, Sirio Legramanti and Raffaele Argiento

Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures 1043
Giulia Capitoli, Simone Colombara, Alessia Cotroneo, Francesco De Caro, Riccardo Morandi, Chiara Schembri, Alfredo G. Zapiola and Francesco Denti

Efficient expectation propagation for high-dimensional probit models 1049
Augusto Fasano, Niccolo Anceschi, Beatrice Franzolini and Giovanni Rebaudo

Model-based clustering of non-stationary time series with common historical change times 1055
Riccardo Corradin, Luca Danese, Wasiur KhudaBukhsh and Andrea Ongaro

## Functional Data Analysis

A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves 1061
Teresa Bortolotti, Riccardo Peli, Giovanni Lanzano, Sara Sgobba and Alessandra Menafoglio

Conditional Gaussian Graphical Models for Functional Variables whit Partial Separable Operators 1065
Rita Fici, Gianluca Sottile and Luigi Augugliaro

Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot 1071
Annachiara Rossi, Andrea Cappozzo and Francesca Ieva

Functional Graphical Models to map Brexit debate on Twitter 1076
Nicola Pronello, Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Sara Fontanella

Measuring Dependence in Multivariate Functional Datasets 1082
Francesca Ieva, Michael Ronzulli and Anna Maria Paganoni

Robust Statistical Process Monitoring of Multivariate Functional Data 1089
Christian Capezza, Fabio Centofanti, Antonio Lepore and Biagio Palumbo

DRAFT

The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021 1095
Veronica Mazzola, Giovanni Bonaccorsi, Piercesare Secchi and Francesca Ieva

## Machine Learning and text mining

A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement 1101
Giulio Giacomo Cantone, Simone Del Sarto and Michela Gnaldi

Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study 1107
Giovanni Bocchi, Alessandra Micheletti, Patrizio Frosini, Alessandro Pedretti, Andrea R. Beccari, Filippo Lunghini, Carmine Talarico and Carmen Gratteri

Hedging global currency risk with factorial machine learning models 1113
Paolo Pagnottoni and Alessandro Spelta

InstanceSHAP: An instance-based estimation approach for Shapley values 1119
Golnoosh Babaei and Paolo Giudici

Networks & Nature Based Solutions: an application for Milan hydric resources 1125
Alessia Forciniti and Emma Zavarrone

The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization 1131
Maria Gabriella Grassia, Marina Marino, Rocco Mazza and Agostino Stavolo

## Multivariate data analysis 3

A comparison of different techniques for handling missing covariate values in propensity score methods 1135
Anna Zanovello, Alessandra R. Brazzale and Omar Paccagnella

A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach 1140
Daniele Cuntrera, Vito M.R. Muggeo and Luigi Augugliaro

A tool for assessing weak identifiability of statistical models 1146
Antonio Di Noia, Francesco Denti and Antonietta Mira

Computing Highest Density Regions with Copulae 1151
Nina Deliu and Brunero Liseo

Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions 1157
Francesca Labanca and Anna Gottard

## Nonparametric statistical methods

A Comparison of Distribution-Free Control Charts 1162
Michele Scagliarini

Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida 1167
Dafne Zorzetto

Comparing three robust procedures for CANDECOMP/PARAFAC estimation 1172
Valentin Todorov, Violetta Simonacci, Michele Gallo and Nikolay Trendafilov

DRAFT

How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods 1178
Anna Vesely and Angela Andreella

Non Parametric Combination methodology: a literature review on recent developments 1184
Elena Barzizza, Nicolò Biasetton and Riccardo Ceccato

## Regression modeling

A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law 1190
Carlo Cusatelli, Massimiliano Giacalone and Eugenia Nissi

A variable selection procedure based on predictive ability: a preliminary study on logistic regression 1195
Rosaria Simone and Mariarosaria Coppola

Comparison of binary regressions with asymmetric link function for imbalanced data 1201
Michele La Rocca, Marcella Niglio and Marialuisa Restaino

New advances in Regression Forests 1207
Mila Andreani, Lea Petrella and Nicola Salvati

Using expectile regression with latent variables for digital assets 1213
Beatrice Foroni, Luca Merlo and Lea Petrella

DRAFT

# Preface

DRAFT

This book includes the contributions presented at the Intermediate Meeting of the Italian Statistical Society (SIS) "SIS 2023 - Statistical Learning, Sustanaibility and Impact Evolution" held in Ancona at the Università Politecnica delle Marche, from June 21th to 23th of 2023.

The new challenges of digitalization, innovation and sustainability are showing the crucial role of data-driven approaches in supporting decision-making processes. Methodologies resulting from the integration of different know-how seem to be a reliable way to deal with the increasing need to measure the impact of the policies and to forecast scenarios. This meeting welcomes any attempt to face new challenges.

The conference has registered more than 250 presentations, including 3 keynote speakers in 3 plenary sessions and 72 presentations in 24 invited sessions, all dealing with specific themes in methodological and/or applied statistics and demography. Furthermore, more than 180 contributions, with one or more authors, have been spontaneously submitted to the Program Committee and arranged in 30 contributed sessions.

DRAFT

# 1 Plenary Sessions

DRAFT

# Inequality Indices: Accurate Simulation-Based Inference

Maria-Pia Victoria-Feser[a]

[a]Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Switzerland; maria-pia.victoriafeser@unige.ch

**Abstract**

In this presentation, we consider the problem of inference for general classes of inequality indices. Indeed, obtaining accurate inference, for example confidence intervals with a coverage that is approximately equal to the confidence level, is not an obvious task, as attested by numerous researchers. The inaccuracy of standard, and some times improved, methods, is particularly important when the income distribution has heavy tails. Therefore, we propose an alternative simulation-based method, based on parametric models, that guaranties theoretically a better accuracy (second order) and is found to provide very accurate inference, even in the presence of very heavy tails, through simulation studies.

Income distribution, indirect inference, bootstrap, heavy tails, second order accuracy.

DRAFT

## 1. Introduction

In this presentation, we consider the problem of inference for an inequality index functional $T$, i.e., quantities of interest that can be written as a function of the data generating model. Without loss of generality, we anchor the discussion in the framework of inequality indices (see e.g., Cowell, 2011). Given a sample $x_i, i = 1, \ldots, n$ and an associated distribution $F$ such that one can assume that $X_i \sim F,\ i = 1, \ldots, n$, we are interested in computing Confidence Intervals (CIs) for $T(F)$. For that, there exists many different approaches that are based on either $T(F^{(n)})$ or $T(F_\theta)$, where $F^{(n)}$ is the empirical distribution (hence leading to a nonparametric approach) and $F_\theta,\ \theta \in \Theta \subset \mathbb{R}^p$ is a parametric model for which $\theta$ needs to be estimated from the sample (hence leading to a parametric approach).

## 2. State of the Art

Inference on the population quantity $T(F)$, can be done in the following manners.

1. The (nonparametric) jackknife produces an estimate of the sampling distribution of $T(F^{(n)})$, which is obtained by constructing samples made of all the observations minus one; for applications to inequality indices, see e.g., Giles (2004); Modarres and Gastwirth (2006).
2. The (nonparametric) bootstrap produces an estimate of the sampling distribution of $T(F^{(n)})$, which is obtained by constructing samples that are a drawn with replacement of the original sample; for application to inequality indices, see e.g., Mills and Zandvakili (1997) and Biewen (2002).
3. Another distribution-free approach consists in deriving the asymptotic variance of the inequality index using the Influence Function ($IF$) of Hampel (1974) as is done in Cowell and Victoria-Feser (2003) (for different types of data features such as censoring and truncating) and estimate it directly from the sample (see also Victoria-Feser, 1999; Cowell and Flachaire, 2015).

4. A parametric approach, given a chosen parametric model $F_{\boldsymbol{\theta}}$ for the data generating model, consists in first consistently estimating $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, say $\hat{\boldsymbol{\theta}}$, then, considering its asymptotic properties, such as its variance var $\left(\hat{\boldsymbol{\theta}}\right)$ and derive the corresponding asymptotic variance of $T\left(F_{\hat{\boldsymbol{\theta}}}\right)$ using e.g., the delta method (based on a first order Taylor series expansion). The later is estimated by plugging in the value of $\hat{\boldsymbol{\theta}}$.
5. The (parametric) bootstrap, produces an estimate of the sampling distribution of $T\left(F_{\hat{\boldsymbol{\theta}}}\right)$, which is obtained by constructing samples that are simulated from $F_{\hat{\boldsymbol{\theta}}}$. This approach is often referred to as the percentile parametric bootstrap.
6. Refinements and combinations of these approaches, such as the studentized bootstrap that is based on a (asymptotic) pivotal statistics.

While most would agree that the fully parametric and asymptotic approach based on the delta method cannot provide as accurate inference as the simulation-based methods, it is not clear that avoiding the specification of a parametric model is the way to go. Indeed, for example, Davidson (2009) note that the jackknife does not yield a reliable estimate of the standard error (and does not correct bias), and Cowell and Flachaire (2015) notice that nonparametric bootstrap inference on inequality indices is sensitive to the exact nature of the upper tail of the distribution, in that nonparametric bootstrap inference is expected to perform reasonably well in moderate and large samples, unless the tails are quite heavy. Similar conclusions are also drawn in Davidson and Flachaire (2007); Schluter and van Garderen (2009) and Davidson (2010).

These findings have motivated the construction of more refined inferential methods, based on second order accuracy, such as, for example, Schluter and van Garderen (2009) and Schluter (2012) who, using the results of Hall (1992), propose normalizing transformations of inequality indices using Edgeworth expansions, to adjust asymptotic Gaussian approximations. Another example is provided by Dufour et al. (2018) who propose an adjustment based on a Fieller-type method (Fieller, 1940, 1954).

Alternatively, Davidson and Flachaire (2007), Cowell and Flachaire (2007) and Davidson (2012) consider a semi-parametric bootstrap, where bootstrap samples are generated from a distribution which combines a parametric estimate of the upper tail, namely the Pareto distribution, with a nonparametric estimate for the other part of the distribution. Finally, Guerrier et al. (2018) propose a percentile parametric bootstrap based on a Generalized Method of Moment (GMM) estimator, with the inequality index as one of the moments, which allows to reduce the biases in the CIs coverage.

DRAFT

## 3. Motivation

Through a large simulation study, Cowell and Flachaire (2015) compare the actual coverage probabilities of 95% CIs for the Theil index, using, as data generating models, different distribution with varying parameters to increase the heaviness of the tail. Most of the different methods cited above are compared. Cowell and Flachaire (2015) conclude that, in the presence of (very) heavy-tailed distributions, even if significant improvements can be obtained on the fully asymptotic and the percentile (parametric) bootstrap methods, none of the alternative methods provides very good results overall.

This result is not surprising, and a simple rationale is given by the fact that any inferential method derived from a non linear function of random variables, specifically an inequality index as a function of estimators, if not directly targeted, adds (uncontrolled) uncertainty to the resulting inferential method. In other words, even if we can have an unbiased estimator for the parameters of a model, say $\hat{\boldsymbol{\theta}}$, so that $\mathbb{E}\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$, we cannot guaranty that $\mathbb{E}\left[T\left(F_{\hat{\boldsymbol{\theta}}}\right)\right]$ equals $T\left(F_{\boldsymbol{\theta}}\right)$. The parametric bootstrap corrects for the bias by improving the bootstrap estimator from an accuracy of $n^{-1/2}$ to $n^{-1}$, but, as has been stated in e.g. Andrews (2000), using as example the sample mean, the bootstrap is not consistent when the parameter's value is near a boundary of the parameter space. This is a typical feature of income distributions, in the sense that most parameters, especially the ones governing the thickness of the tail,

are typically positive. Moreover, because inequality indices are functions of estimators (in the parametric framework), a valid resampling method should be based on percentiles, which guaranties the *Parameter-Transformation* (PT) property, which is not the case, for example, with the studentized bootstrap or other refinements of the percentile (parametric) bootstrap.

## 4. Results

In this presentation, we propose a simulation-based parametric approach for inference about an inequality index $T(F_{\boldsymbol{\theta}})$ that has the following properties:

1. it satisfies the PT property,
2. it is not affected by the boundary problem like the bootstrap methods,
3. it is second order accurate.

Although specifying a parametric distribution for the data generating process can be considered as an additional risk of introducing "error" in the inferential procedure, common wisdom however suggests, at least with income distributions, that some parametric models are sufficiently flexible to encompass most of the data generating processes observed with real data. For example, the four parameters generalized beta distribution of second kind (GB2) proposed by McDonald (1984), which encompasses the generalized gamma, the Singh-Maddala (SM) (Singh and Maddala, 1976) and Dagum distribution (Dagum, 1977) (see also McDonald and Xu, 1995), can be considered as sufficiently general to model income data. If this is not the case, then one would wonder if the lack of flexibility of a general four parameter model is not due to a spurious amount of observations, and hence consider a robust estimation approach as proposed and motivated in Cowell and Victoria-Feser (1996) (see also Cowell and Victoria-Feser, 2000).

The proposed method is based on matching estimators in the spirit of indirect inference (Gouriéroux et al., 1993; Smith, 1993), which are simulated to provide their sample distribution. In Guerrier et al. (2018), it is shown that the estimation bias of such estimators is of order $n^{-1}$, which indicates that a properly constructed simulation-based inferential procedure can lead to more accurate inference. We show that we can build a simulation-based procedure that does not involve a double bootstrap such as is the case for the Studentized bootstrap, and the resulting method satisfies the properties stated above. Moreover, through simulation studies, we can confirm our theoretical results in that CIs computed for some inequality indices, under different parametric income distributions, also with very heavy tails, have almost exact coverage.

DRAFT

## 5. Inequality Indices Functionals

As a leading example, we consider $T$ to be an inequality index and $F$ an income distribution. Inequality indices are welfare indices which can be very generally written in the following quasi-additively decomposable form (see Cowell and Victoria-Feser (2002, 2003) for the original formal setting)

$$W_{\text{QAD}}(F) = \int \varphi\left(x, \mu(F)\right) dF(x), \tag{1}$$

where $\varphi$ is piecewise differentiable in $\mathbb{R}$. The generalized entropy family of inequality indices given by

$$I^{\xi}_{\text{GE}}(F) = \frac{1}{\xi^2 - \xi}\left[\int \left[\frac{x}{\mu(F)}\right]^{\xi} dF(x) - 1\right], \tag{2}$$

is obviously obtained by setting

$$\varphi\left(x,\mu(F)\right)=\frac{1}{\xi^2-\xi}\left[\left[\frac{x}{\mu(F)}\right]^{\xi}-1\right]. \tag{3}$$

For example, the cases $\xi=0$ and $\xi=1$ are given by

$$\begin{aligned} I^0_{\mathrm{GE}}(F) &= -\int \log\left(\frac{x}{\mu(F)}\right) dF(x), \\ I^1_{\mathrm{GE}}(F) &= \int \frac{x}{\mu(F)} \log\left(\frac{x}{\mu(F)}\right) dF(x), \end{aligned} \tag{4}$$

with $I^0_{\mathrm{GE}}(F)$ being the Mean Logarithmic Deviation (see Cowell and Flachaire 2015) and $I^1_{\mathrm{GE}}(F)$ being the Theil index. A notable exception to the class in (1) is the Gini coefficient which can be expressed in several forms, such as

$$I_{\mathrm{Gini}}(F)=1-2\int_0^1 \frac{C(F;q)}{\mu(F)}dq, \tag{5}$$

with $C(F;q)=\int^{F^{-1}(q)} xdF(x)$, the cumulative income functional.

DRAFT

## REFERENCES


Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica 68, 399–405.

Biewen, M. (2002). Bootstrap inference for inequality, mobility and poverty measurement. Journal of Econometrics 108, 317–342.

Cowell, F. (2011). Measuring Inequality. Oxford University Press. Third Edition.

Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. Journal of Econometrics 141, 1044–1072.

Cowell, F. A. and E. Flachaire (2015). Statistical Methods for Distributional Analysis. In F. Bourguignon and A. B. Atkinson (Eds.), Handbook of Income Distribution, Volume 2, pp. 359–465. Elsevier.

Cowell, F. A. and M.-P. Victoria-Feser (1996). Robustness properties of inequality measures. Econometrica 64, 77–101.

Cowell, F. A. and M.-P. Victoria-Feser (2000). Distributional analysis: A robust approach. In A. B. Atkinson, H. Glennerster, and N. Stern (Eds.), Putting Economics To Work, volume in honour of Michio Morishima. London, UK.

Cowell, F. A. and M.-P. Victoria-Feser (2002). Welfare rankings in the presence of contaminated data. Econometrica 70, 1221–1233.

Cowell, F. A. and M.-P. Victoria-Feser (2003). Distribution-free inference for welfare indices under complete and incomplete information. Journal of Economic Inequality 1/3, 1–29.

Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. Economie Appliquée 30, 413–436.

Davidson, R. (2009). Reliable inference for the Gini index. Journal of Econometrics 150, 30–40.

Davidson, R. (2010). Innis lecture: Inference on income distributions. Canadian Journal of Economics 43, 1122–1148.

Davidson, R. (2012). Statistical inference in the presence of heavy tails. Econometrics Journal 15, 31–53.

Davidson, R. and E. Flachaire (2007). Asymptotic and bootstrap inference for inequality and poverty measures. Journal of Econometrics 141, 141–166.

Dufour, J.-M., E. Flachaire, L. Khalaf, and A. Zalghout (2018). Confidence sets for inequality measures: Fieller-type methods. In W. H. Greene, L. Khalaf, P. Makdissi, R. C. Sickles, M. Veall, and M.-C. Voia (Eds.), Productivity and Inequality, Cham, pp. 143–155. Springer International Publishing.

Fieller, E. C. (1940). The biological standardization of insulin. Journal of the Royal Statistical Society (Supplement) 7, 1–64.
Fieller, E. C. (1954). Some problems in interval estimation. Journal of the Royal Statistical Society, Series B 16, 175–185.
Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: some further results. Oxford Bulletin of Economics and Statistics 66, 425–433.
Gouriéroux, C., A. Monfort, and A. E. Renault (1993). Indirect inference. Journal of Applied Econometrics 8 (supplement), S85–S118.
Guerrier, S., E. Dupuis, Y. Ma, and M.-P. Victoria-Feser (2018). Simulation based bias correction methods for complex models. Journal of the American Statistical Association (Theory & Methods), *in press*.
Guerrier, S., S. Orso, and M.-P. Victoria-Feser (2018). Parametric inference for index functionals. Econometrics 6.
Hall, P. (1992). The Bootstrap and Edgeworth Expansions. Springer Verlag.
Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association 69, 383–393.
McDonald, J. B. (1984). Some generalized functions for the size distribution of income. Econometrica 52, 647–664.
McDonald, J. B. and Y. J. Xu (1995). A generalization of the beta distribution with applications. Journal of Econometrics 66, 133–152.
Mills, J. and S. Zandvakili (1997). Statistical inference via bootstrapping for measures of inequality. Journal of Applied Econometrics 12, 133–150.
Modarres, R. and J. L. Gastwirth (2006). A cautionary note on estimating the standard error of the Gini index of inequality. Oxford Bulletin of Economics and Statistics 68.
Schluter, C. (2012). On the problem of inference for inequality measures for heavy-tailed distributions. The Econometrics Journal 15, 125–153.
Schluter, C. and K. J. van Garderen (2009). Edgeworth expansions and normalizing transforms for inequality measures. Journal of Econometrics 150, 16–29.
Singh, S. K. and G. S. Maddala (1976). A function for the size distribution of income. Econometrica 44, 963–970.
Smith, J. A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. Journal of Applied Econometrics 8, S63–S84.
Victoria-Feser, M.-P. (1999). Comment on Giorgi's chapter: The sampling properties of inequality indices. In J. Silber (Ed.), Income Inequality Measurement: From Theory to Practice, pp. 260–267. Boston: Kluwer Academic Publisher.

DRAFT

# Examples from the Interface of Neural Models and Spatio-Temporal Statistics in Environmental Applications

Christopher K. Wikle[a], Likun Zhang[a], Myungsoo Yoo[a], and Xiaoyu Ma[a]

[a]146 Middlebush Hall, University of Missouri, Columbia, MO 65203 USA;
wiklec@missouri.edu, likun.zhang@missouri.edu,
myungsoo.yoo@mail.missouri.edu, xm3cf@mail.missouri.edu

DRAFT

**Abstract**

Hybrid models that combine neural network structures within statistical models are increasingly useful to predict complex processes. Here, we demonstrate two such approaches for spatio-temporal data motivated by enviornmental problems. The first example combines an echo state network (ESN) within a spatio-temporal statistical level-set model to forecast the movement of a wildfire boundary. The ESN is used to model the speed of the fire front in the normal direction to the front via an implicit signed-distance function. The second problem is concerned with emulating complex environmental model output that can have dependent extremes. This is accomplished by combining a flexible statistical model for spatial extremes within a variational autoencoder. Both examples have the benefit of modeling complex processes while being computationally efficient.

***Keywords:*** basis expansion, echo state network, ELBO, emulation, level set, variational autoencoder, wildfire

## 1. Introduction

Neural models and "deep learning" have dramatically altered the modeling landscape for dependent data. Indeed, for prediction and classification problems that have data with spatial and/or temporal dependence, these models (particularly convolutional neural networks and recurrent neural networks of various types) are often more successful than traditional statistical models, especially when there are large amounts of training data available. In statistics, deep models (i.e., Bayesian hierarchical models) have been used since the late 1990s to model spatial and spatio-temporal data in environmental statistical applications (see the overview in 2). In many ways, these models are not unlike the deep neural models in that they both learn handle complex dependence structure by a series of linked multi-level (or conditional) formulations (see the discussion in 13).

In recent years, there has been a great deal of interest in building hybrid neural-statistical (or machine learning-statistical) modeling methodologies for spatial and spatio-temporal data, particularly in the environmental sciences. For a recent overview, see (14). Such approaches borrow the black-box predictive strengths of the deep neural approaches and yet accommodate uncertainty quantification and other explainability approaches.

Here, we present two brief examples from our work at the interface of statistics and neural modeling. The first example, presented in Section 2, is a spatio-temporal model for modeling the spread of wildfire fronts that models the speed of boundary propagation with an echo state network (ESN) within a

statistical framework that is motivated by level-set dynamics. The second example, presented in Section 3, presents a novel variational autoencoder to emulate spatio-temporal processes that exhibit dependent extremes. This paper presents a brief conclusion in Section 4.

## 2. Modeling Wildfire Spread with Hybrid Level-Set/Echo State Network Models

Millions of acres of land are destroyed by wildfires every year, and they pose a significant threat to humans both in terms of property damage and potential loss of life, as well as significantly impacting the ecosystem. It is important to develop trustworthy models that can be used to manage and mitigate large wildfires. The majority of operational wildfire spread models are based on semi-empirical formulas and frequently assume that the environment is homogenous and well-known. That is, these models do well when the fire is burning over homogeneous fuel sources with relatively steady winds and relatively level terrain. They do not do so well in more complex real-world settings (e.g., in steep complex terrain with multiple fuel sources and where the fire itself generates winds). Recently (15) showed that one could use a level-set dynamic model (which is common in fire modeling) within a spatio-temporal statistical framework, and where the propagation speed was given by an echo state network (ESN), which is an effective and easily trained neural model. We briefly summarize that model here.

### 2.1 Level-Set Dynamics

A level set of a real-valued function $\phi$ of $n$ real variables is a set where the function takes on a given constant value $c$:

$$L_c(\phi)\{(x_1, \ldots, x_n)|\phi(x_1, \ldots, x_n) = c\}. \tag{1}$$

A closed-contour fire front can be defined as the intersection between the level set function and the zero-plane (a zero level-set). The evolution of such an iimplicit function that depends on space and time, $\phi(\mathbf{s}, t)$, was first introduced by (12) in terms of an advection equation of the form:

$$\frac{\partial \phi(\mathbf{s}, t)}{\partial t} + \mathbf{v}(\mathbf{s}, t) \cdot \nabla \phi(\mathbf{s}, t) = 0, \tag{2}$$

where $\nabla\phi$ is spatial gradient of $\phi$, and $\mathbf{v}(\mathbf{s}, t) = (v_x(\mathbf{s}, t), v_y(\mathbf{s}, t))^\top$ denotes the velocity vector, which also depends on time $t$ and spatial location $\mathbf{s}$. If evolution is assumed to be in normal direction to the boundary (a reasonable assumption for fire spread), then

$$\frac{\partial \phi(\mathbf{s}, t)}{\partial t} + v_n(\mathbf{s}, t)\ ||\nabla \phi(\mathbf{s}, t)|| = 0, \tag{3}$$

where $v_n$ is the scalar velocity (speed) in the normal direction (positive is outward and negative is inward), and $||\cdot||$ denotes $\ell_2$ norm operation (magnitude).

It is most useful to consider the implicit function $\phi(\mathbf{s})$ for $\forall \mathbf{s}$ to be a signed-distance function, which is defined as:

$$\phi(\mathbf{s}) = \begin{cases} -d(\mathbf{s}), & \mathbf{s} \in \Omega^- \\ 0, & \mathbf{s} \in \partial\Omega \\ d(\mathbf{s}), & \mathbf{s} \in \Omega^+ \end{cases}, \tag{4}$$

where the distance function $d(\mathbf{s})$ is defined as $d(\mathbf{s}) = min(\mathbf{s} - \mathbf{s_I})$, for $\mathbf{s_I} \in \partial\Omega$, and $\partial\Omega, \Omega^+$ and $\Omega^-$ correspond to the boundary, exterior, and interior of interest, respectively.

A signed distance function also has the important property that $||\nabla\phi|| = 1$. Thus, in the case where $\phi$ is a signed distance function with flow in the normal direction, the advection equation can be written:

$$\frac{\partial \phi}{\partial t} + v_n(\mathbf{s}, t) = 0, \tag{5}$$

DRAFT

which can be approximated very simply by a forward Euler discretization that gives:

$$\phi(\mathbf{s}, t + \triangle t) = \phi(\mathbf{s}, t) - (\triangle t) v_n(\mathbf{s}, t), \tag{6}$$

where $\triangle t$ is time increment.

So, if we know the normal component of the velocity along the boundary, we can simply evolve the signed-distance boundary using this approximation. The challenge is that we do not know $v_n$ for all spatial locations and time points; also, this simplification introduces some error. Traditionally, $v_n$ is parameterized in fire models with a quasi-empirical relationship, which is not often appropriate for complex real-world fires. We wish to use the data to learn the normal velocity more generally and account for uncertainty in the observations and the simple dynamical representations.

## 2.2 Hybrid Statistical Model

Consider the vector representation of the physical-statistical Euler advection equation for a normal flow signed-distance function:

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_{t-\triangle t} - \mathbf{v}_{t-\triangle t} \triangle t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim Gau(\mathbf{0}, \sigma_\eta^2 \mathbf{I}), \tag{7}$$

where $\mathbf{v}_{t-\triangle t} = (v_n(\mathbf{s}_1, t - \triangle t), \ldots, v_n(\mathbf{s}_N, t - \triangle t))^\top$ is vector of speed in normal direction at time $t - \triangle t$ on a spatial grid $\{\mathbf{s}_i : i = 1, \ldots, N\}$, and $\sigma_\eta^2$ is common variance for process error $\boldsymbol{\eta}_t$. We model the $v_n$ process in terms of an ESN.

ESNs are a type of reservoir computing in which an input signal goes into a hidden fixed "reservoir" that allows sparse neural linkages. Importantly, the fixed weights within the reservoir are selected randomly and are not trained, and only the output layer is trained – with a regularization penalty (e.g., ridge regression). (10; 11) introduced spatio-temporal ESNs to the statistics community and showed that they were excellent and forecasting, especially in situations where there was more limited training data. Embedding the ESN within the level-set model gives the following deep model.

$$\begin{aligned}
\text{output response}: \quad & \boldsymbol{\phi}_t = \boldsymbol{\phi}_{t-\triangle t} - \mathbf{v}_{t-\triangle t} \triangle t + \boldsymbol{\eta}_t \\
\text{spread speed}: \quad & \mathbf{v}_{t-\triangle t} = \mathbf{W}^{out} \mathbf{h}_{t-\triangle t} \\
\text{hidden states}: \quad & \mathbf{h}_{t-\triangle t} = (1 - \alpha_\ell)\mathbf{h}_{t-2\triangle t} + \alpha_\ell \tilde{\mathbf{h}}_{t-\triangle t} \\
& \tilde{\mathbf{h}}_{t-\triangle t} = g_h\left(\frac{\nu}{|\lambda_w|}\mathbf{W}^{in}\mathbf{h}_{t-2\triangle t} + \mathbf{U}\mathbf{x}_{t-\triangle t}\right) \\
\text{parameters}: \quad & \mathbf{W}^{in} = [w_{i,\ell}]_{i,\ell} : \gamma^w_{i,\ell} Unif(-a_w, a_w) + (1 - \gamma^w_{i,\ell})\delta_0 \\
& \mathbf{U} = [u_{i,j}]_{i,j} : \gamma^u_{i,j} Unif(-a_u, a_u) + (1 - \gamma^u_{i,j})\delta_0 \\
& \gamma^w_{i,\ell} \sim Bern(\pi_w) \\
& \gamma^u_{i,j} \sim Bern(\pi_u) \\
\text{hyperparameters}: \quad & \alpha_\ell, \nu, a_w, a_u, \gamma^w_{i,\ell}, \gamma^u_{i,j}, \pi_w, \pi_u, n_h
\end{aligned}$$

Note that ESN models, like most neural models, do not include an uncertainty quantification component. It is customary with ESNs to consider an ensemble of reservoirs because it is so inexpensive to fit these models (they are just regularized regression models with nonlinearly transformed inputs). As demonstrated in (15), we sample different ensembles by sampling hyperparameters from their prior distributions, and then calibrate the ensembles to obtain a prediction interval analogous to HPD intervals in Bayesian models.

## 2.3 Wildfire Prediction Example

As an example, (15) considered multiple simulations and two real-world fires. Here, we present a result on one of those fires, with inputs given by $\mathbf{x}_t = [\boldsymbol{\phi}_t^T, \boldsymbol{\phi}_{t-1}^T]^T$. In particular, we consider the Haypress

DRAFT

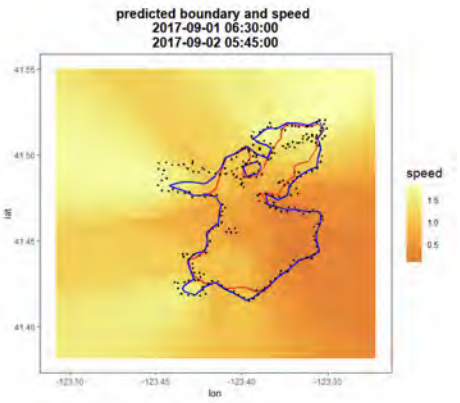



Figure 1: Forecast median and normal direction speed for the Haypress fire at time 05:45:00 on 2 September 2017. The dashed black line corresponds to the truth at the forecast time $T$, the solid blue line corresponds to the corresponding forecast median, and the solid red line corresponds to the true boundary at time $T-1$. The background speeds correspond to the forecast from time $T-1$.

fire, which burned over 27,000 acres in Siskiyou County, California (USA). The fire started in July 2017 and was contained in January 2018. We consider the first 10 observations of the fire boundary from the GeoMac database as training data (converted to signed-distance functions), and forecast the next observation. Figure 1 gives the forecast median boundary and corresponding forecast speed. Note that although the forecast boundary does not quite resolve the "finger" that extended to the west, the forecast interval (not shown) does cover it. More importantly, one can see that the boundary is responding to the forecast speeds by extending in the directions of maximum speed.

DRAFT

## 3. Emulating Spatial Extremes with Variational Autoencoders

Large mechanistic-based simulation models are crucial for understanding complex problems related to the environment. These models are ubiquitous in science and engineering. For example, earth system models couple complex climate models with physical, chemical, and biological models that include interactions between ocean, atmosphere, sea ice, the biosphere along with human population effects. Each component model in such systems is typically quite computationally expensive and, when combined, the computational and uncertainty quantification (UQ) challenges can be quite formidable. Examples of such challenges include model calibration, use of such models for inverse problems, UQ for forward simulations, and the specification of realistic model parameterizations for coupled processes. Surrogate model emulators (e.g., see the overview in (5)) have proven to be useful in recent years to facilitate UQ in these contexts, particularly when combined with Bayesian inference (7). In many cases, the interest in a mechanistic simulator is to quantify extreme events (e.g., mega-wildfires, drought, floods, pandemics, anomaly detection, etc.). In most cases these extreme events are dependent in space and time (e.g., drought in the desert southwest that lasts multiple years). Traditional methods for model emulation such as Gaussian processes, polynomial chaos expansions, and more recently, deep neural networks and generative models (generative adversarial networks (GANs) and variational autoencoders or VAEs) do not naturally accommodate extreme values, and certainly not dependent extreme values. Our interest is to demonstrate that one can modify a traditional VAE to model dependent spatial extremes.

### 3.1 Variational Autoencoders (VAEs)

Say we wish to generate realizations of a process, say $\mathbf{X}$. VAEs work by training a latent process, say $\mathbf{Z}$, given real-world examples of $\mathbf{X}$. This problem fits naturally into a Bayesian framework, where the pos-

terior distribution for $\mathbf{Z}$ is given by: $p(\mathbf{Z}|\mathbf{X}) \propto p_\theta(\mathbf{X}|\mathbf{Z})p_\alpha(\mathbf{Z})$, where we say the distribution $p_\theta(\mathbf{X}|\mathbf{Z})$, which depends on parameters $\boldsymbol{\theta}$ is the *decoder*. The posterior distribution $p(\mathbf{Z}|\mathbf{X})$ learns the appropriate distribution for the $\mathbf{Z}$ process, which has a prior distribution $p_\alpha(\mathbf{Z})$ that depends on parameters $\boldsymbol{\alpha}$. In practice, obtaining the normalizing constant that gives the posterior distribution is not available in closed form, yet samples from the posterior distribution can be obtained by computationally expensive Markov chain Monte Carlo (MCMC) methods. Alternatively, the VAE literature seeks a tractable approximation to the posterior, $q_\phi(\mathbf{Z}|\mathbf{X})$, which depends on parameters $\boldsymbol{\phi}$ and we call this the *encoder*. The variational approach then allows us to learn the parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ by maximizing the evidence lower bound (ELBO):

$$ELBO = -D_{KL}[q_\phi(\mathbf{Z}|\mathbf{X})||p_\alpha(\mathbf{Z})] + E_{q_\phi(\mathbf{Z}|\mathbf{X})}[p_\theta(\mathbf{X}|\mathbf{Z})],$$

where $D_{KL}$ corresponds to the Kullback-Leibler divergence between the distribution on the left side and right sides of $||$. In practice, we can easily estimate the second term on the right-hand side via Monte Carlo, utilizing samples of the latent $\mathbf{Z}$ process from the encoder for different values of $\mathbf{X}$, and where a *reparameterization trick* is used to get these samples from $q_\phi(\mathbf{Z}|\mathbf{X})$. Thus, after we have optimized the ELBO to obtain the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, we can easily generate realistic realizations of $\mathbf{X}$ from the decoder after sampling a $\mathbf{Z}$. In practice, deep neural models are often used to specify the encoder and decoder, providing very expressive models. However, despite the effectiveness of these methods for most emulation tasks, none of these methods explicitly accounts nor is provably robust to dependent extreme values generated in the simulator.

## 3.2 Modeling Dependent Extremes

Classic extreme value theory has been widely used to study the tail behaviors of univariate climate and finance observations (3, Ch. 1–5). For multivariate observations over space and time, it is of primary interest to characterize the spatial and multivariate dependence in the joint tail but it is challenging to develop models that have flexible dependence properties. Under the spatial setting, well-known models like max-stable or generalized Pareto processes directly assume asymptotic results from multivariate extreme value theory that are too rigid to be applied on environmental data (4). In contrast, Gaussian processes always underestimate the risk associated with simultaneous occurrences of extremes due to its light joint tail. To better describe the nuanced tail behaviors such as those commonly seen in extreme weather events, we briefly introduce the following tail dependence measure.

Similar to variograms in spatial statistics, the extremal dependence structure is commonly described by the bivariate measure

$$\chi(u) = P\{F_2(X_2) > u | F_1(X_1) > u\} = \frac{P\{F_2(X_2) > u, F_1(X_1) > u\}}{P\{F_1(X_1) > u\}}, \tag{8}$$

in which $u \in (0, 1)$ and $X_1$ and $X_2$ are two variables with marginal distribution functions $F_1$ and $F_2$. In the spatial case, $X_1 = X(\mathbf{s}_1)$ and $X_2 = X(\mathbf{s}_2)$ are two observations from the process $\{X(\mathbf{s}); \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2\}$ at spatial locations $\mathbf{s}_1$ and $\mathbf{s}_2$. In the multivariate case, $X_1$ and $X_2$ are components of the $d$-variate random vector $\mathbf{X}$. When $u$ is close to one, $\chi(u)$ quantifies the probability of simultaneous occurrences of events exceeding the high extremity level $u$ given one variable is extreme. If $\lim_{u\to 1} \chi(u) = 0$, $X_1$ and $X_2$ are said to be *asymptotically independent* (AI), and if $\lim_{u\to 1} \chi(u) > 0$, $X_1$ and $X_2$ are *asymptotically dependent* (AD). The aforementioned max-stable or generalized Pareto processes have the property that $\chi(u)$ becomes independent of $u$ for increasingly extreme events, resulting in a positive limit (6; 8). On the other hand, Gaussian processes (or multivariate Gaussian distributions) have the property that $\chi(u)$ will always converge to 0, no matter how close $\mathbf{s}_1$ and $\mathbf{s}_2$ are, or how similar the two components $X_1$ and $X_2$ are. Observed tail dependence in environmental processes often decays as events get more extreme and rare events often tend to be more spatially localized as the intensity increases. In this case, the stability property of max-stable and generalized Pareto models is a physically inappropriate restriction on the joint tail.

DRAFT

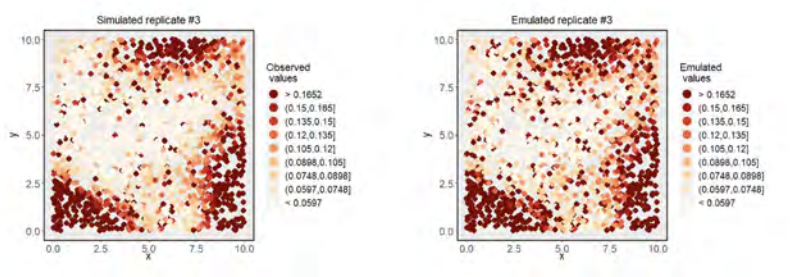



Figure 2: The left panel shows a simulated realization from a spatial extremes process. The right panel shows an emulated replicate corresponding to this process that was obtained from our extremes VAE.

## 3.3 Extreme Variational Autoencoder

As described in (9), we combine a flexible spatial extremes model with a VAE that has non-stationary and space-scale aware dependence flexibility. We base our extremes model on the max-infinitely divisible process proposed by (1), which allows for both short-range asymptotic independence and dependence along with long-range aysymptotic independence:

$$X_t(\mathbf{s}) = \epsilon_t(\mathbf{s}) Y_t(\mathbf{s}), \ \mathbf{s} \in \mathcal{D}, \tag{9}$$

where $X_t(\mathbf{s})$ is a spatio-temporal output from a simulator (e.g., high-resolution climate model), $\epsilon_t(\mathbf{s})$ is a white noise process with an independent Fréchet$(m, \tau, 1)$ marginal distribution, where $Y_t(\mathbf{s})$ is described by a low-rank basis expansion:

$$Y_t(\mathbf{s}) = \sum_{k=1}^{K} \omega_k(\mathbf{s}, r_k)^{1/\alpha} Z_{kt}, \tag{10}$$

where $\omega_k(\mathbf{s}, r_k)$, $k = 1, \ldots, K$ are compactly-supported basis functions centered at $K$ knots, and $Z_{kt}$ are latent basis expansion coefficients. The latent variables are lighter-tailed, exponentially tilted, positive-stable variables:

$$Z_{kt} \sim H(\alpha, \alpha, \theta_k), \ k = 1, \ldots, K, \tag{11}$$

where $\alpha \in (0, 1)$ is the concentration parameter, and lighter-tailed distributions of $Z_{kt}$ are given by larger values of the tail index parameters $\theta_k \geq 0$.

Critically, the latent variables $Z_{kt}$, $k = 1, \ldots, K$ are used in the encoded space of a VAE. So, spatio-temporal inputs are compressed (encoded) into low $(K)$ dimensional latent space, yet we keep the distributional assumptions given above when decoding the latent variables back to the initial space. In our case, the encoder and decoder are both deep neural networks with weights updated iteratively to minimize the ELBO. Reiterating, the goal here is to emulate the field $\{X_t(\mathbf{s})\}$. Because the latent variables enable a generative process, once trained, we can obtain many copies of the original field $\{X_t(\mathbf{s})\}$ that have similar properties to the original data, even in the joint tail.

Figure 2 shows a realization from a simulated dependent extremes process in the left panel, and the right panel shows an emulated realization from our extremes VAE. Although not shown here, estimates of the dependence measure, $\chi(u)$ shows agreement in the extremal dependence, whereas a Gaussian process emulator produced realizations that underestimate the dependence in the tails.

# 4. Conclusion

As machine learning and "AI" neural models become more prevalent, there are increasing opportunities to combine those models into a statistical framework to facilitate prediction and classification, yet with uncertainty quantification. Here, we present two approaches that illustrate such hybrid models. The first considers a model for wildfire spread that uses ESNs to estimate the rate of spread in the normal

DRAFT

direction to the fire front. This approach takes advantage of the fact that ESNs produce nonlinear evolution yet are simple to train do not require a large amount of training data. The second example, which is less mature, describes our initial approach to build spatial extremes into a VAE. Not only is this method flexible, but it is computationally more efficient than traditional Bayesian implementations of dependent spatial extremes models. The main use of our model is to emulate complex model output, such as output from climate models that exhibit extremes in space. Traditional emulation approaches for such models do not typically handle extremes in a realistic fashion.

## References


[1] Bopp, G.P., Shaby, B.A., Huser, R.: A hierarchical max-infinitely divisible spatial model for extreme precipitation. Journal of the American Statistical Association, **116**, 93–106 (2021)

[2] Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. John Wiley & Sons, Boca Raton (2011)

[3] De Haan, L. and Ferreira, A.: Extreme value theory: an introduction, vol. 21, Springer. (2006)

[4] Ferreira, A., de Haan, L.: The generalized Pareto process; with a view towards application and simulation. Bernoulli. **20**, 1717 – 1737 (2014)

[5] Gramacy, R.B.: Surrogates: Gaussian process modeling, design and optimization for the applied sciences. Chapman Hall/CRC, Boca Raton, Florida. (2020)

[6] Huser, R., Wadsworth, J.L.: Advances in statistical modeling of spatial extremes. Wiley Interdisciplinary Reviews: Computational Statistics. **14** (2022)

[7] Kennedy, M. C., O'Hagan, A.: Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). **63**,425–464 (2001)

[8] Kiriliouk, A., Rootzén, H., Segers, J. Wadsworth, J.L.: Peaks over thresholds modeling with multivariate generalized Pareto distributions. Technometrics. **61**, 123–135 (2019)

[9] Ma, X., Zhang, L., Wikle, C.K., Huser, R.: Emulating complex model output via a hybrid spatial extremes variational autoencoder. In progress.

[10] McDermott, P.L., Wikle, C.K.: An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. Stat. **6**, 315–330 (2017)

[11] McDermott, P.L., Wikle, C.K.: Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. Environmetrics. **30** (2019)

[12] Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics. **79** (12–49)

[13] Wikle, C.K.: Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. Journal of Agricultural, Biological and Environmental Statistics. **24**, 175–203 (2019)

[14] Wikle, C.K., Zammit-Mangion, A.: Statistical deep learning for spatial and spatio-temporal data. Annual Review of Statistics and Its Application. **10**, 247–270 (2023)

[15] Yoo, M., Wikle, C.K.: Using echo state networks to inform physical models for fire front propagation. Spatial Statistics. **54** (2023)


DRAFT

# 2 Invited Sessions

DRAFT

# Causal Discovery for complex survey data

Paola Vicard[a]

[a]Università Roma Tre; paola.vicard@uniroma3.it

**Abstract**

The association structure of a Bayesian network can be drawn based on subject matter or experts knowledge, or have to be learned from a database. In case of data driven learning, one of the most known procedures is the PC algorithm that is based on the assumption of independent and identically distributed observations. In practice, sample selection in surveys involves more complex sampling designs then the standard test procedure is not valid even asymptotically. In order to avoid misleading results about the true causal structure the sample selection process must be taken into account in the structural learning process. In this paper, a modified version of the PC algorithm is proposed for inferring casual structure from complex survey data. Finally, a simulation experiment is performed

***Keywords:*** Complex survey design, Faithfulness, PC algorithm

DRAFT

## 1. Introduction

Bayesian networks (BN) are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see (9). The advantage of using BNs to represent probability distributions is that they graphically show causal and conditional independence relations between random variables making easier for decision makers to evaluate and use the model. The nodes of the graph correspond to random variables and edges between nodes represent dependencies between the associated variables. Augmented with a table of parameters representing marginal probabilities, BNs are capable of representing the probabilities over any discrete sample space: the probability of any sample point in that space can be computed from the probabilities in the BN.

In recent years BNs have been successfully applied to official statistic problems. In particular, BNs appear to be very useful in missing item imputation ((5), (6)) and contingency table estimation for complex survey sampling (1). The use of BNs to model measurement errors has also been investigated in (7). However, there are still some limitations that may complicate their large scale application in the official statistics context. The main one is how to take into account the complexity of sampling design in the structural learning process when BN methodology is applied to sample surveys.

Building BNs by hand can be a difficult and time-consuming task. The association structure can be known in advance by subject matter knowledge or have to be learned from a database. In case of data driven learning, one of the most known procedures is the PC algorithm by which the structure is inferred carrying out several independence tests and building a Bayesian network in agreement with tests results. The algorithm is presented in detail in (11).

The PC algorithm is based on the assumption of independent and identically distributed observations (*i.i.d.*, for short). However, most of the commonly used survey designs employ stratification and/or cluster sampling and/or unequal selection probabilities. The impact of complex designs on *i.i.d.* based methods can be severe, as shown in (10). In this paper a modified version of the PC algorithm for complex survey data is proposed; it is named PC complex.

The paper is organized as follows. First of all, in Section 2, the PC algorithm for *i.i.d.* data and the basic assumptions on which it relies are briefly recalled. Secondly, in order to take into account the complexity of the sampling design in the learning process, a procedure for testing the association in a two-way table in complex sample surveys is described in subsection 2.1. To evaluate the accuracy of the proposed algorithm, a simulation study is performed in Section 3.

## 2. Model selection in graphical modeling: the PC algorithm

Let $V = \{X_1, X_2, \ldots, X_d\}$ be a set of variables with joint probability distribution $P$ whose casual structure can be represented by a DAG. The PC algorithm starts from a complete undirected graph and recursively deletes edges based on conditional independence decisions. More specifically, the algorithm consists of two phases. In phase 1 the skeleton of the graph is determined, in phase 2 arrows are oriented. The assumptions under which the PC algorithm correctly estimates the models are the following:

1. *Causal Sufficiency Condition* (CSC, for short). Variables that are the direct cause of at least two variables should be known.
2. *Causal Markov Condition* (CMC, for short) describes the independencies that follow from a causal structure: each variable is probabilistically independent from all its non-descendants conditional on its direct causes.
3. *Causal Faithfulness Condition* (CFC, for short) states that if a conditional independence relation is not entailed by the CMC, then it does not hold in the population. (12) decomposes the CFC in (i) *Adjacency-Faithfulness Condition*; (ii) *Orientation-Faithfulness Condition*. The former is necessary to recover the true skeleton. The latter is necessary for finding the correct orientations. The two conditions are strictly related to phase 1 and phase 2 of the PC algorithm, respectively.

Under the above assumptions and if the sample size is large enough for reliable inference on conditional independence, it is possible to infer the true causal graph from data. In practice, we need to do statistical inference based on a finite sample and the original graph could not be recovered due to errors in the statistical tests.

For instance, with regard to the CFC, it becomes very relevant to causal inference whether the probability distribution $P$, though faithful to the true casual structure, is far from or close to being unfaithful. Intuitively, a population distribution may be faithful but close to unfaithfulness, that is, the situation when independence seems to hold due to insufficient sample size. As a consequence, when applied to sample data, the PC algorithm can have problems of reliability.

The situation worsens for complex survey data. Roughly speaking, the CMC and the CFC may fail if the population is selected by a procedure that is biased toward two or more of the observed variables. In order to avoid misleading results about the true causal structure, the complexity of sampling design must be taken into account in the learning process. Here, a procedure (described in subsection 2.1) for testing the association in a two-way table for data coming from complex sample surveys is introduced in phase 1 of the PC algorithm giving rise to the PC complex algorithm.

### 2.1 Independence test for complex sample surveys via resampling

Denote by $A$ and $B$ two characters of interest with $H$ and $K$ categories, respectively. Furthermore, let the superpopulation parameters $p^{hk}$, $p^{h\cdot}$, $p^{\cdot k}$ be defined as

$$p^{hk} = Prob(A = A^h, B = B^k), \quad p^{h\cdot} = Prob(A = A^h), \quad p^{\cdot k} = Prob(B = B^k) \tag{1}$$

for $h = 1, \ldots, H$ and $k = 1, \ldots, K$. Let $\mathcal{U}_N$ be a finite population of size $N$, labeled by integers $1, \ldots, N$, generated from the superpopulation model (1). For each unit $i$ let $Y_i^{h\cdot}$ ($Y_i^{\cdot k}$) be the variable taking the value 1 if the unit $i$ assumes the modality $A^h$ ($B^k$) and 0 otherwise, for $h = 1, \ldots, H$ ($k = 1, \ldots, K$) and let $Y_i^{hk}$ be the variable given by the product $Y_i^{h\cdot} Y_i^{\cdot k}$.

DRAFT

For each unit $i$ of the population $\mathcal{U}_N$, let $D_i$ be a Bernoulli random variable (r.v.), such that $i$ is in the sample whenever $D_i = 1$, whilst $i$ is not in the sample whenever $D_i = 0$. Let $D = (D_1 \ldots D_N)$. An unordered, without replacement sampling design $P$ is the probability distribution of $D$, with $\pi_i = E_P[D_i]$ being the first order inclusion probability of unit $i$. Here we will consider fixed size sampling designs. Assumptions on the sampling design according to which the sample is drawn, are similar to those used in (3).

Let $p_N^{hk}$,$p_N^{h\cdot}$, $p_N^{\cdot k}$ be the finite population parameters defined as

$$p_N^{hk} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{hk}, \quad p_N^{h\cdot} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{h\cdot} = \sum_{k=1}^{K} p_N^{hk}, \quad p_N^{\cdot k} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\cdot k} = \sum_{h=1}^{H} p_N^{hk} \tag{2}$$

$h = 1, \ldots, H$, $k = 1, \ldots, K$. The parameters (2) can be estimated using the classical Hájek estimators

$$\widehat{p}^{hk} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{hk}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}}, \quad \widehat{p}^{h\cdot} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{h\cdot}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}} = \sum_{k=1}^{K} \widehat{p}^{hk}, \quad \widehat{p}^{\cdot k} = \frac{\sum_{i=1}^{N} \frac{D_i Y_i^{\cdot k}}{\pi_i}}{\sum_{i=1}^{N} \frac{D_i}{\pi_i}} = \sum_{h=1}^{H} \widehat{p}^{hk} \tag{3}$$

$h = 1, \ldots, H$, $k = 1, \ldots, K$. Here, similarly to (4), in order to prove the existence of the limiting distribution of the Hájek estimators (3) as the sample size and the population size increase, we consider the stochastic processes:

$$W^{HK} = \sqrt{n}(\widehat{p}^{hk} - p^{hk}), \quad W^{H} = \sqrt{n}(\widehat{p}^{h\cdot} - p^{h\cdot}), \quad W^{K} = \sqrt{n}(\widehat{p}^{\cdot k} - p^{\cdot k}) \tag{4}$$

for $h = 1, \ldots, H$, $k = 1, \ldots, K$.

Each of these processes can be partitioned into the sum of two stochastic processes

$$\begin{aligned}
W^{HK} &= \{W_n^{HK} + W_N^{HK}, h = 1, \ldots, H, k = 1, \ldots, K\} \\
&= \{\sqrt{n}(\widehat{p}^{hk} - p_N^{hk}) + \sqrt{f}\sqrt{N}(p_N^{hk} - p^{hk}), h = 1, \ldots, H, k = 1, \ldots, K\}, \qquad (5) \\
W^{H} &= \{W_n^{H\cdot} + W_N^{H\cdot} h = 1, \ldots, H\} \\
&= \{\sqrt{n}(\widehat{p}^{h\cdot} - p_N^{h\cdot}) + \sqrt{f}\sqrt{N}(p_N^{h\cdot} - p^{h\cdot}), h = 1, \ldots, H\}, \qquad (6) \\
W^{K} &= \{W_n^{\cdot K} + W_N^{\cdot K}, k = 1, \ldots, K\} \\
&= \{\sqrt{n}(\widehat{p}^{\cdot k} - p_N^{\cdot k}) + \sqrt{f}\sqrt{N}(p_N^{\cdot k} - p^{\cdot k}), k = 1, \ldots, K\} \qquad (7)
\end{aligned}$$

where $f = \frac{n}{N}$. The components $W_n^{HK}, W_n^{H\cdot}, W_n^{\cdot K}$ depend on the sample selection randomness while the components $W_N^{HK}, W_N^{H\cdot}, W_N^{\cdot K}$ depend on the superpopulation randomness.

In Proposition 1 in (8) it is shown that processes (4) converge to a Gaussian distributions. Specifically, under the assumptions A1-A6 of Proposition 1 in (4), as $n$ and $N$ increase, the sequences:

1. $W_n^{HK}$ and $W_N^{HK}$ converge in distribution to degenerate multivariate normal distributions with mean zero and singular covariance matrices $\Sigma_1^{HK}$ and $\Sigma_2^{HK}$ of order $HK$, respectively. As a consequence, the whole process $W^{HK}$ converges to a degenerate multivariate normal distribution with mean zero and singular covariance matrix $\Sigma^{HK} = \Sigma_1^{HK} + f\Sigma_2^{HK}$.
2. $W_n^{H}$ and $W_N^{H}$ converge in distribution to degenerate multivariate normal distributions with mean zero and singular covariance matrices $\Sigma_1^{H}$ and $\Sigma_2^{H}$ of order $H$, respectively. As a consequence, the whole process $W^{H}$ converges to a degenerate multivariate normal distribution with mean zero and singular covariance matrix $\Sigma^{H} = \Sigma_1^{H} + f\Sigma_2^{H}$.
3. $W_n^{K}$ and $W_N^{K}$ converge in distribution to degenerate multivariate normal distributions with mean zero and singular covariance matrices $\Sigma_1^{K}$ and $\Sigma_2^{K}$ of order $K$, respectively. As a consequence, the whole process $W^{K}$ converges to a degenerate multivariate normal distribution with mean zero and singular covariance matrix $\Sigma^{K} = \Sigma_1^{K} + f\Sigma_2^{K}$.

DRAFT

Next, we consider the hypothesis testing problem with null hypothesis of $A$ and $B$ independent, and alternative hypothesis of of $A$ and $B$ associated. Formally

$$\mathcal{H}_0 : p^{hk} = p^{h\cdot}p^{\cdot k} \quad \text{against} \quad \mathcal{H}_1 : p^{hk} \neq p^{h\cdot}p^{\cdot k}. \tag{8}$$

The following test statistic is used

$$\chi^2_H = n \sum_{h=1}^{H} \sum_{k=1}^{K} \frac{(\widehat{p}^{hk} - \widehat{p}^{h\cdot}\widehat{p}^{\cdot k})^2}{\widehat{p}^{h\cdot}\widehat{p}^{\cdot k}} \tag{9}$$

where $\widehat{p}^{hk}$, $\widehat{p}^{h\cdot}$ and $\widehat{p}^{\cdot k}$ are the Hájek estimators given in (3). In (8) it is shown that for complex survey data: (i) the test statistic (9) does have a limiting distribution; (ii) the limiting distribution does not necessarily approach to a Chi-square distribution because of the singularity of the matrices $\Sigma^{HK}$, $\Sigma^{H}$ and $\Sigma^{K}$. More specifically, $\chi^2_H$ tends in distribution to a quadratic form of a degenerate multinormal distribution.

The limiting sampling distribution of test statistic (9) under the independence null hypothesis is estimated resorting to resampling methods for finite population. Different versions of the bootstrap have been proposed in the literature, see (3), (4) and references therein. In (4) a class of resampling techniques based on a two-step procedure is proposed. Such a methodology consists in cr eating an artificial population (called pseudo-population) from the initial sample, and drawing bootstrap samples from it using the original sampling scheme.

DRAFT

## 3. Accuracy evaluation

A Monte Carlo simulation has been performed to assess the accuracy of the proposed algorithm. A finite population of size $N = 10000$ has been generated according to the network in Fig. 1a. The probability distributions of the nodes $X_1$, $X_2$ and $X_3|(X_1, X_2)$ are reported in Tables 1-3. The finite population causal structure has been estimated using the function *pc()* in Package pcalg. The finite population CPDAG in Fig. 1a has been obtained.

The effect of the sampling design on the structural learning process has been studied selecting 500 samples of size $n = 3000$ from the finite population according to (i) a simple random sampling design; (ii) a conditional Poisson sampling design with inclusion probabilities proportional to the $Z$-values, defined as follows

$$Z = \begin{cases} N(200, 2) + 10 & X_2 = 0 \\ N(10, 2) + 5 & X_2 = 1 \end{cases} \tag{10}$$

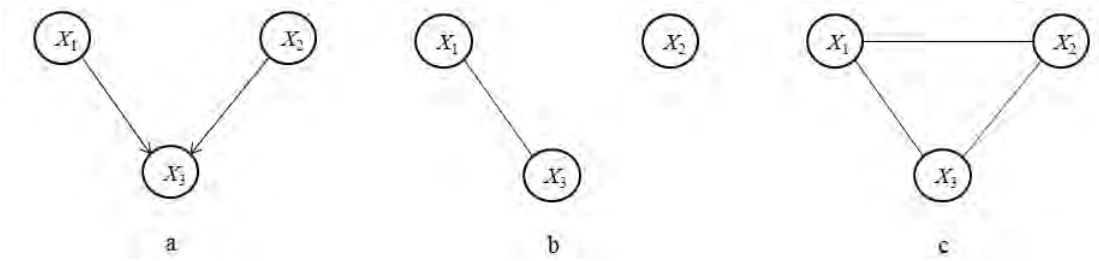



Figure 1: (a) True graph and finite population CPDAG, (b)-(c) Sampling design effects.

The significance level is fixed to 0.05. When the sample is selected according to a simple random sampling, the PC algorithm is not able to recover the true association structure in 3% of the selected samples.

Table 1: Probability distribution of $X_1$

| $X_1$ | $P(X_1)$ |
|---|---|
| 0 | 0.15 |
| 1 | 0.45 |
| 2 | 0.40 |

Table 2: Probability distribution of $X_2$

| $X_2$ | $P(X_2)$ |
|---|---|
| 0 | 0.7 |
| 1 | 0.3 |

Table 3: Probability distribution of $X_3|(X_1, X_2)$

| $X_3$ | $X_1$ | $X_2$ | $P(X_3\|(X_1, X_2))$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.10 |
| 1 | 0 | 0 | 0.50 |
| 2 | 0 | 0 | 0.40 |
| 0 | 1 | 0 | 0.40 |
| 1 | 1 | 0 | 0.20 |
| 2 | 1 | 0 | 0.40 |
| 0 | 2 | 0 | 0.20 |
| 1 | 2 | 0 | 0.20 |
| 2 | 2 | 0 | 0.60 |
| 0 | 0 | 1 | 0.70 |
| 1 | 0 | 1 | 0.20 |
| 2 | 0 | 1 | 0.10 |
| 0 | 1 | 1 | 0.30 |
| 1 | 1 | 1 | 0.50 |
| 2 | 1 | 1 | 0.20 |
| 0 | 2 | 1 | 0.35 |
| 1 | 2 | 1 | 0.25 |
| 2 | 2 | 1 | 0.40 |

The percentage of wrong graphs rises to $10.7\%$ when the sample is selected according to a conditional Poisson sampling. Fig. 1 shows the effects of the survey design on the association structure. The edge between the nodes $X_2$ and $X_3$ is missing in $34\%$ of the wrong graphs (Fig. 1b). An additional edge is placed between $X_1$ and $X_2$ in the remaining $66\%$ (Fig. 1c).

Next, the PC complex has been applied. For each sample, a pseudo population has been constructed and $M = 1000$ bootstrap samples have been drawn. It results that the percentage of wrong graphs decreases to $3.2\%$.

DRAFT

## References


[1] Ballin, M., Scanu, M., Vicard, P.: Estimation of contingency tables in complex survey sampling using probabilistic expert systems. J. Statist. Plann. Infer. **140**, 1501–1512, (2010)

[2] Conti, P.L, Marella, D.: Inference for quantiles of a fnite population: asymptotic vs. resampling results. Scand. Journal of Statistics, **42**, 545–561, (2015)

[3] Conti, P.L, Marella, D., Mecatti, F., Andreis, F.: A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. Bernoulli **26**, 1044–1069 (2019)

[4] Conti, P.L., Di Iorio, A. (2018). Analytic inference in finit populations via resampling, with applications to confidence intervals and testing for independence. *arXiv:1809.08035*.

[5] Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, A. : Bayesian networks for imputation. J. Roy. Statist. Soc. Ser. A. **167**, 309–322 (2004)

[6] Di Zio, M., Sacco, G., Scanu, M., Vicard, P.: Multivariate techniques for imputation based on Bayesian networks. Neu. Net. World. **4**, 303–309 (2005).

[7] Marella, D., Vicard, P. : Object-Oriented Bayesian Networks for Modeling the Respondent Measurement Error. Comm. Statist. Theory Methods, **42**, 19, 3463–3477, (2013)

[8] Marella, D., Vicard, P. : Bayesian Networks structural learning from complex survey data: a resampling based approach. Stat. Methods and Appl., **31**, 981–1013, (2022)

[9] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, (1988)

[10] Skinner, C.J., Holt, D., Smith, M.F.: Analysis of complex surveys. Wiley, (1989)
[11] Spirtes, P., Glymour, G., Scheines, R. : Causation, Prediction, and Search. Springer-Verlag, New York, (1993)
[12] Zhang, J., Spirtes, P.: Detection of Unfaithfulness and Robust Causal Inference. Minds and Machines, **18**, 239-271, (2008)

DRAFT

# Data Integration without conditional independence: a Bayesian Networks approach

Pier Luigi Conti[a], Paola Vicard[b], and Vincenzina Vitale[a]

[a]Sapienza Università di Roma; pierluigi.conti@uniroma1.it, vincenzina.vitale@uniroma1.it
[b]Università Roma Tre; paola.vicard@uniroma3.it

**Abstract**

Statistical Matching, at a macro level, consists in estimating the joint distribution of variables separately observed in independent samples. As a consequence of the lack of joint information on the variables of interest, uncertainty about the data generating model is the most relevant feature of matching. In the present paper the use of graphical models to deal with the statistical matching uncertainty for multivariate categorical variables is considered, under both a model-based and a model-assisted perspective.

***Keywords:*** Data integration, statistical matching, Bayesian networks, sampling design.

DRAFT

## 1. Introduction and basics

As a consequence of the *data deluge* phenomenon, in the last years there has been a considerable increase of available statistical data, coming from different sources: Official Statistics, Institutional entities, as well as private subjects. Data sometimes come from well-designed sample surveys, sometimes come from institutional archives (with some form of design), sometimes from surveys where sample units are essentially self-selected. The final effect of the data deluge is an increasing flow of data, including also data generated from online transactions, emails, videos, audios, images, click streams, logs, search queries, health records, social networking interactions, sensors and mobile phones, etc.. In different words, this is the big data era.

Using the above data has advantages and disadvantages.

- The main advantage is in terms of *cost*. An *ad hoc*, well-designed, sample survey where all variables of interest are observed is expensive, and frequently unaffordable. The availability of free data clearly offers an attracting alternative.
- The main disadvantages are two. First of all, freely available data hardly ever contain all variables of interest. In the second place, if data are collected through some form of self-selection of sample units, there is a potential bias that could make statistical estimates affected by serious errors.

Consider a finite population of $N$ units, and let $(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, N$ be a superpopulation model composed by independent and identically distributed (*i.i.d.*) random variables (r.v.s) with joint (either discrete or absolutely continuous) distribution P. In other terms, $(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i)$ are independent copies of a r.v. $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. With no essential loss of generality, it is assumed that $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iH})$, $\boldsymbol{Y} = (Y_{i1}, \ldots, Y_{iK})$, $\boldsymbol{Z} = (Z_{i1}, \ldots, Z_{iL})$ are vectors of r.v.s of dimension $H$, $K$, $T$, respectively. Furthermore, we will denote by $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ the joint density (w.r.t. either the counting measure - discrete case - or the Lebesgue measure - absolutely continuous case) of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Statistical matching

at a macro level consists in estimating the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$s (or possibly of some of its parameters) when:

1. $(\boldsymbol{X}, \boldsymbol{Y})$ are observed in a sample $\boldsymbol{S}_A$ of size $n_A$, possibly drawn according to a complex sampling design;
2. $(\boldsymbol{X}, \boldsymbol{Z})$ are observed in a sample $\boldsymbol{S}_B$ of size $n_B$, possibly drawn according to a complex sampling design;
3. $\boldsymbol{S}_A$ and $\boldsymbol{S}_B$ are independent, and the sets of observed units in the two samples do not overlap.

The main problem in estimating the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is essentially a *missing values* problem, where missingness is inherent, because the units in $\boldsymbol{S}_A$ have $\boldsymbol{Z}$ as missing values and the units in $\boldsymbol{S}_B$ have $\boldsymbol{Y}$ as missing values. As a consequence, the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is generally *not identifiable* from collected data because of the absence of joint observations of $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$.

The oldest approach to overcome this problem is to assume that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent conditionally on $\boldsymbol{X}$. This is the well-known Conditional Independence Assumption (CIA, for short). Under CIA, the model is actually identifiable (12) and consistent estimates of the involved distributions can be constructed (1).

Another approach that makes the model identifiable consists in using techniques based on the external auxiliary information; the most relevant case occurs when an additional, possibly small-scale, sample $\boldsymbol{S}_C$ where $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ are jointly observed is available; cfr. (14).

Unfortunately, in many cases CIA cannot be reasonably assumed, and an additional, complete sample $\boldsymbol{S}_C$ is hardly ever available. As a consequence of unidentifiability of the statistical model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, even for a large sample sizes $n_A$, $n_B$, one can only obtain a "blurred" estimate of the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. The blurring effect of unidentifiability can be numerically quantified through a *measure of uncertainty* in statistical matching; cfr. (1; 4).

DRAFT

## 2. Bayesian Networks for Statistical Matching: model-based approach

The goal of the present section is to illustrate the use of Bayesian networks (BNs) to deal with the statistical matching uncertainty in multivariate categorical data. Throughout the present section, the following assumptions are made.

A1. $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is a discrete, multivariate r.v., with joint probability function (p.f.) $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ and marginal p.f.s $f_X(\boldsymbol{x})$, $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$, $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$.

A2. The samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ are independent, and selected according to ignorable sampling designs. In this way, the r.v.s $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, $i \in \boldsymbol{S}_A$ are *i.i.d.*, and the r.v.s $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, $i \in \boldsymbol{S}_B$ are *i.i.d.*, as well.

In the sequel, we will essentially focus on *model-based* inference, where the sampling designs used to select $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ do not play any role. This allows to use standard methods both for parameters estimation and model construction.

The first attempt to use BNs in Statistical Matching of multivariate discrete data is in (7), where the CIA is assumed, and its connection with $d$-separation criterion exploited. Of course, under CIA both the dependence structure and the BN parameters are estimable from the sample data and there is no uncertainty. When CIA is unadequate, the final dataset may be significantly different from the one it would have been available if complete observations of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ were collected, and the application of standard inferential procedures may result in highly biased estimates.

A considerable step forward is (4), where Statistical Matching based on BNs without CIA is thoroughly studied. Differently from (7), BNs can help tackling the multivariate statistical matching problem both computationally and by targeting the effort for uncertainty assessment only to those components of the joint p.f. relative to variables separately observed in the two available samples.

A BN is a probabilistic graphical model encoding a joint probability distribution satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG) (5). A graph is a pair $G = (V, E)$, where $V$ is the set of vertices, and $E$ is the set of directed edges between pairs of nodes. In particular, directed acyclic graphs are considered: it is not possible to start from a node and go back

to the same node following arrows directions. Each node corresponds to a random variable and missing arrows between nodes imply (conditional) independence between the corresponding variables. In a BN each node $x_h$, say, is associated with the distribution of the corresponding variable given its parents, $pa(x_h)$, namely all nodes linked to $x_h$ by an arrow pointing to $x_h$. As a matter of fact, a BN consists of two components: the DAG and the set of the distributions parameters.

The use of BNs in the statistical matching problem is motivated by several advantages, listed below.

1. BNs can consider prior knowledge to be included in Statistical Matching.
2. BNs are widely used to describe dependencies among variables in multivariate distributions.
3. BNs admit convenient recursive factorizations of their joint probability useful for data integration and uncertainty evaluation in a multivariate context.

In Statistical Matching, the non-identifiability of the statistical model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, due to the absence of joint observations on the variables of interest, implies that both the components of a BN (namely, the underlying DAG and its parameters) cannot be estimated from samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$. As a matter of fact fact, when BNs are used for Statistical Matching, two operations are necessary.

$(i)$ Specification of the dependence structure of the variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$.
$(ii)$ Estimation of the parameters, *i.e.*, the local probability distributions associated to the edges between the components of $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

As a consequence, the two kinds of uncertainty reported below have to be accounted for.

1. Uncertainty on the DAG, namely on the dependence structure among the variables of interest.
2. Uncertainty on the parameters of the statistical relationships between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ (namely on the corresponding conditional probabilities) given the DAG, *i.e.* given the factorization of the joint p.f. of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$.

The above two forms of uncertainty will be discussed in the sequel. The steps on which the proposed Statistical Matching rests are listed below.

Step 1 Estimate the DAGs of $\boldsymbol{X}$, $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Z})$ from $\boldsymbol{S}_A \cup \boldsymbol{S}_B$, $\boldsymbol{S}_A$, $\boldsymbol{S}_B$, respectively.
Step 2 Extra sample information on the association structure between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ must be inserted in the DAGs. For instance, if a variable $Y_k$ is associated to $Z_l$ then a link between the vertices $Y_k$ and $Z_l$ must be added.
Step 3 Consider the class of plausible DAGs for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, and compute the uncertainty due to the corresponding dependence structure.
Step 4 Select a model from the class of plausible DAGs defined in Step 3.
Step 5 Estimate the parameters of the local p.f.s associated to the edges between the components of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, and compute the parameters estimation uncertainty.
Step 6 Compute the total uncertainty by adding the dependence structure uncertainty to the uncertainty in parameters estimation.

Step 3 requires the definition of a class $\mathcal{G}_{\boldsymbol{XYZ}} = \{G_{\boldsymbol{XYZ}}\}$ of plausible DAGs for the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Generally speaking, the widest class of DAGs one may consider is the class of all joint distributions for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ that are *collapsible* over $\boldsymbol{Y}$, $\boldsymbol{Z}$, respectively. In formal terms, $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ is collapsible over $Z_t$ if the estimate obtained for the probability function of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \setminus Z_t)$ does coincide with the estimate obtained by marginalizing w.r.t. $Z_t$ the estimate of the probability function of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Collapsibility w.r.t. subsets $\boldsymbol{Z}' \subseteq \boldsymbol{Z}$ or $\boldsymbol{Y}' \subseteq \boldsymbol{Y}$ is similarly defined. As a matter of fact, the plausible class of DAGs can be also defined in terms of $c$-removability; cfr. (4).

The most favorable case occurs under CIA, because the class $\mathcal{G}_{\boldsymbol{XYZ}}$ of all probability functions for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ collapses on a single graph where $Y$ and $Z$ are $d$-separated by $X$, with joint p.f. defined as $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{z}|\boldsymbol{x})$. When CIA does not hold, there is *uncertainty* on the dependence structure of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. As a matter of fact, there are *two* sources of uncertainty.

U1. Uncertainty of the specific graph $G_{\boldsymbol{XYZ}} \in \mathcal{G}_{\boldsymbol{XYZ}}$, *i.e.* on the independence structure of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Each graph in $\mathcal{G}_{\boldsymbol{XYZ}}$ represents an *equivalence class* of probability functions $f(\boldsymbol{x}\ \boldsymbol{y}, \boldsymbol{z})$ all sharing the same structure of local independence.

DRAFT

U2. Uncertainty on the parameter estimates of the non-identifiable probability function(s) corresponding to the selected graph $G_{\boldsymbol{XYZ}}$.

Both sources of uncertainty are intrinsic in the statistical matching problem.

As an *uncertainty measure* for the class $\mathcal{G}_{\boldsymbol{XYZ}}$, namely a measure to compare the plausible independence equivalence classes in terms of graph density/sparsity, the Structural Hamming Distance (SHD, for short) can be used. Such a distance, applied to independence equivalence classes, can be defined as the number of edge insertions or deletions necessary to transform a given equivalence class into another equivalence class.

As far as uncertainty in parameter estimation is concerned, once a graph $G_{\boldsymbol{XYZ}} \in \mathcal{G}_{\boldsymbol{XYZ}}$ has been chosen, let $\theta_{\boldsymbol{XYZ}}$ be parameter vector of the corresponding joint probability function $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, and let $\theta_{\boldsymbol{X}\cdot\cdot}$ be the parameters vector of the marginal distribution of $X$, let $\theta_{\boldsymbol{XY}\cdot}$ be the parameters vector of the marginal distribution of $(\boldsymbol{X}, \boldsymbol{Y})$, and let $\theta_{\boldsymbol{X}\cdot\boldsymbol{Z}}$ be the parameters vector od the marginal distribution of $(\boldsymbol{X}, \boldsymbol{Z})$. Note that $\theta_{\boldsymbol{X}\cdot\cdot}$, $\theta_{\boldsymbol{XY}\cdot}$, $\theta_{\boldsymbol{X}\cdot\boldsymbol{Z}}$ are identifiable, and can be consistently estimated on the basis of sample data.

Due to the unidentifiability issue, one can only say that it belongs to a set $\Theta_R$. At an estimation level, under the assumption that the sampling designs used to select $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ is *ignorable*, this identifies the *likelihood ridge*, *i.e.* the set of estimates

$$\widehat{\Theta}_R = \left\{ \widehat{\theta}_{\boldsymbol{XYZ}} : \sum_{\boldsymbol{z}} \widehat{\theta}_{\boldsymbol{XYZ}} = \widehat{\theta}_{\boldsymbol{XY}\cdot}, \sum_{\boldsymbol{y}} \widehat{\theta}_{\boldsymbol{XYZ}} = \widehat{\theta}_{\boldsymbol{X}\cdot\boldsymbol{Z}} \right\}$$

where $\widehat{\theta}_{\boldsymbol{XY}\cdot}$, $\widehat{\theta}_{\boldsymbol{X}\cdot\boldsymbol{Z}}$ are the Maximum Likelihood Estimates (MLEs) of the joint probabilities of $f(\boldsymbol{x}, \boldsymbol{y})$ and $f(\boldsymbol{x}, \boldsymbol{z})$, respectively.

As a measure of uncertainty for $\widehat{\Theta}_R$, the most natural choice consists is using supremum of the Lebesgue measure of all open subsets contained in $\widehat{\Theta}_R$. The main issue related to this measure is its computation. A much simpler alternative is proposed in (4). Furthermore, as a global measure of uncertainty, the sum of measures corresponding to U1, U2 are taken.

Before closing the present section, a few remarks.

1. The MLEs $\widehat{\theta}_{\boldsymbol{XY}\cdot}$, $\widehat{\theta}_{\boldsymbol{X}\cdot\boldsymbol{Z}}$ can be legitimately used as a consequence of the ignorability assumption. More importantly, ignorability has to be intended *conditionally on design variables for all population units*; cfr. (13).
2. The assumption of ignorability also allows one to use well-consolidated methodologies for testing for the presence of arcs in the marginal graphical models for $(\boldsymbol{X}, \boldsymbol{Y})$, $(\boldsymbol{X}, \boldsymbol{Z})$.
3. Statistical Matching under non-ignorable sampling designs is investigated in (9). The authors apply a parametric approach, basing the inference on the sample distribution, namely the distribution holding for the observed sample data. However, the maximization of sample likelihood can be complicated numerically and result in unstable estimates, depending on the population model and the model assumed for the sample selection probabilities, given the observed data. For this reason, in (10) the use of the empirical likelihood, which enables estimating the parameters governing sample distribution without specifying the corresponding population model, is proposed.

## 3. Statistical Matching under model-assisted approach

The goal of the present section is to extend the approach of Section 2. to the case when samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ are collected through complex, probabilistic sample designs. This case is of extreme interest in case of secondary analyses, where only design weights for sampled units are actually available. In these cases, that are actually very frequent in practice, the use of a model-based approach could be prone to a relevant source of bias, because the r.v.s $((\boldsymbol{X}_i, \boldsymbol{Y}_i) \;\; i \in \boldsymbol{S}_A)$ (or $((\boldsymbol{X}_i, \boldsymbol{Z}_i) \;\; i \in \boldsymbol{S}_B)$), given $\boldsymbol{S}_A$, $\boldsymbol{S}_B$, could be neither independent, nor identically distributed. In other terms, the conditional distributions of $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i);\, i \in \boldsymbol{S}_A\}|\boldsymbol{S}_A$, $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i);\, i \in \boldsymbol{S}_B\}|\boldsymbol{S}_B$ do not necessarily coincide with the corresponding

DRAFT

marginal distributions (which is exactly the assumption on which model-based inference rests); cfr. (13) and references therein for further considerations on this crucial point.

In the sequel, we will denote by $\pi_{iA}$, $\pi_{iB}$ the first order inclusion probabilities of units of samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$, respectively. Moreover, the two samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ will be still assumed independent.

The procedure described in Section 2. can be still used in a context of independent samples collected through complex sample designs, provided that two points remarked above (*i.e.* MLEs and model selection) are accounted for.

As far as estimates of $\widehat{\theta}_{\boldsymbol{XY}\cdot}$, $\widehat{\theta}_{\boldsymbol{X}\cdot\boldsymbol{Z}}$ are concerned, the simplest idea is to use an appropriate pseudo-likelihood. In addition to the notation already introduced, let us denote by $\theta_{\boldsymbol{Y}|\boldsymbol{X}}$ the parameters of the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X}$, and by $\theta_{\boldsymbol{Z}|\boldsymbol{X}}$ the parameters of the conditional distribution of $\boldsymbol{Z}|\boldsymbol{X}$. Clearly, there is a one-to-one map between $(\theta_{\boldsymbol{X}\cdot\cdot}, \theta_{\boldsymbol{Y}|\boldsymbol{X}})$ and $\theta_{\boldsymbol{XY}\cdot}$, and a one-to-one map between $(\theta_{\boldsymbol{X}\cdot\cdot}, \theta_{\boldsymbol{Z}|\boldsymbol{X}})$ and $\theta_{\boldsymbol{X}\cdot\boldsymbol{Z}}$.

The parameters $\theta_{\boldsymbol{X}\cdot\cdot}$ are then estimated by maximizing the *pseudo-log-likelihood*

$$\widetilde{l}_{\boldsymbol{X}}(\theta_{\boldsymbol{X}\cdot\cdot}) = \sum_{i\in\boldsymbol{S}_A} \frac{1}{\pi_{i,A}} \log f_{\boldsymbol{X}}(\boldsymbol{x}_i) + \sum_{i\in\boldsymbol{S}_B} \frac{1}{\pi_{i,B}} \log f_{\boldsymbol{X}}(\boldsymbol{x}_i). \tag{1}$$

In the context of file concatenation, pseudo-likelihood has been considered in (6).

Similarly, the parameters $\theta_{\boldsymbol{Y}|\boldsymbol{X}}$, $\theta_{\boldsymbol{Z}|\boldsymbol{X}}$ can be estimated by maximizing the conditional pseudo-log-likekihoods

$$\begin{aligned}
\widetilde{l}_{\boldsymbol{Y}|\boldsymbol{X}}(\theta_{\boldsymbol{Y}|\boldsymbol{X}}) &= \sum_{i\in\boldsymbol{S}_A} \frac{1}{\pi_{i,A}} \log f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_i|\boldsymbol{x}_i); \\
\widetilde{l}_{\boldsymbol{Z}|\boldsymbol{X}}(\theta_{\boldsymbol{Z}|\boldsymbol{X}}) &= \sum_{i\in\boldsymbol{S}_B} \frac{1}{\pi_{i,B}} \log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}_i|\boldsymbol{x}_i).
\end{aligned}$$

In is not difficult to see that the estimated obtained in this way are essentially equivalent to Hájek estimates, and hence they possess the asymptotic, large sample properties established in (3). Of course, considerable simplifications can be obtained by exploiting local independency relationships in the graphical models for $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Z})$. Exactly as before, in testing for the presence/absence of arcs (a crucial step in model construction) a major role is played by the sampling designs used to select $\boldsymbol{S}_A$, $\boldsymbol{S}_B$. As shown in (11), ignoring the effect of sampling designs by using a purely model-based approach is a source of relevant bias. In the same paper, a new approach to model construction, based on resampling under complex sampling design, is proposed.

Further extensions of the model-assisted approach to non-probability samples $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ may be considered, provided that for $\boldsymbol{S}_A$, $\boldsymbol{S}_B$ there are reference surveys $\boldsymbol{S}^*_{A'}$, $\boldsymbol{S}^*_{B'}$ allowing consistent estimates of the corresponding sample weights; cfr. (8) and references therein.

DRAFT

## References


[1] Conti, P L., Marella, D., Scanu, M.: On the matching noise of some nonparametric imputation procedures. Statistics and Probability Letters, **78**, 1593–1600 (2008)

[2] Conti, P L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. Journal of the American Statistical Association, **111**, 1715–1725 (2016)

[3] Conti, P L., Marella, D., Mecatti, F., Adreis, F.: A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. Bernoulli, **26**, 1044–1069 (2020)

[4] Conti, P L., Marella, D., Vicard, P., Vitale, V.: Multivariate statistical matching using graphical modeling. Journal of Approximate Reasoning, **130**, 150–169 (2021)

[5] Dawid, A P., Lauritzen, S L., Cowell, R G., Spiegelhalter, D J.: Probabilistic networks and expert system. Springer, New York (1999)

[6] D'Orazio, M., Di Zio, M., Scanu, M.: Old, new approaches in statistical matching when samples are drawn with complex survey designs. Atti della XLV Riunione Scientifica della Società Italiana di Statistica (Padova, June 16-18, 2010). Cleup, Padova (2008)

[7] Endres, E., Augustin, T.:Statistical matching of discrete data by Bayesian networks. In: JMLR: Workshop and Conference Proceedings, **52**, 159–170 (2016)
[8] Kim, J K., Park, S., Chen, Y., Wu, C.: Combining Non- Probability and Probability Survey Samples Through Mass Imputation. Journal of the Royal Statistical SOciety A, **184**, 941–963 (2021)
[9] Marella, D., Pfeffermann, D.: Matching information from two independent informative sampling. Journal of Statistical Planning and Inference, **203**, 70–81. (2019)
[10] Marella, D., Pfeffermann, D.: Accounting for Non-ignorable Sampling and Non-response in Statistical Matching. International Statistical Review. (2022) https://doi.org/10.1111/insr.12524
[11] Marella, D., Vicard, P.: Bayesian network structural learning from complex survey data: a resampling based approach. Statistical Methods and Applications, **31**, 981–1013 (2022)
[12] Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement, **1**, 325–342 (1972)
[13] Pfeffermann, D.: The Role of Sampling Weights When Modeling Survey Data. International Statistical Review, **61**, 317–337 (1993)
[14] Singh, A C., Mantel, H., Kinack, M., Rowe, G.: Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, **19**, 59–79, (1993)

DRAFT

# Mass imputation through Machine Learning techniques in presence of multi-source data

De Fausti F.[a], Di Zio M.[a], Filippini R.[a], and Toti S.[a]

[a]Istat, Via Cesare Balbo, 16, Roma;
defausti@istat.it, dizio@istat.it, filippini@istat.it, toti@istat.it

**Abstract**

The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of administrative information. Nevertheless, it is needed to resort to sample survey data to cope with delay of information and coverage problems. Due to the complexity and heterogeneity of the available information, the solution of the problem with standard statistical methods needs the construction of different imputation models with a strong effort in terms of human intervention. We study the use of a multilayer perceptron model to make the process more automatic, i.e., less costly in terms of human resources, and possibly more accurate in terms of estimates. Since a relevant quality aspect is the ability of the imputation process to provide a good estimate of the ALE frequency distribution, sampling weights are used in the estimation techniques. A comparison between different approaches in the use of weights for Machine Learning is carried out: results obtained from a weighted loss function are compared with those obtained from a data set expanded according to sampling weights.

***Keywords:*** Register-based statistics, imputation, sampling weights, weighted machine learning, expanded data

DRAFT

## 1. Introduction

The Attained Level of Education (ALE) of the Permanent Italian Census exploits a high amount of administrative information. Nevertheless, it is needed to resort to sample survey data to cope with the delay of information and coverage problems. The official procedure for the estimation of the ALE for Italian resident population in 2018 relies on a mass imputation approach using log-linear models, see (5). Mass-imputation is adopted because administrative data are not available for all the units in the population, and because they have a temporal lag with respect to the reference time (generally one-year later). Machine learning techniques could provide a valuable and more automated alternative for the imputation task (2). In (3), a comparison between the official imputation approach based on log-linear models, and multilayer perceptron models (MLP) is discussed. In (4), the study is repeated by considering sampling weights that are introduced in the loss function of the algorithm used for estimating MLP. In this work, we extend the study in (4) by adopting a similar approach to the one used for log-linear imputation, based on the pseudo-likelihood approach, in fact MLP are applied to a data set obtained expanding units of the sample according to their weights.

The paper is structured as follows. In Section 2, MLP and the use of sampling weights are presented. The experimental application for the empirical evaluation of the proposal is carried out on a subset of real data and is described in Section 3.

## 2. Multilayer perceptron model and sampling weights

National Statistics Institutes (NSIs) routinely use complex sampling designs to carry out sample surveys. Statistical analysis on complex survey data make use of sampling weights that typically enter estimator expressions in the form of weights attached to the survey units.

The use of survey sampling weights in machine learning applications is receiving increasing attention in the research community, see for instance (1), (2), (7).

To take into account sampling weights, in (4) we adopt a loss function that is weighted with sampling weights (MLP1, hereafter). A similar idea in a different context is in (6) where weights are used to correct for the difference between train and target distribution.

More in detail, the procedure (MLP1) studied in (4) is the following:

- Input variables of the perceptron multilayer are encoded as one-hot encoding[1];
- MLP is estimated by minimizing the cross-entropy loss function considering sampling weights as follows:

$$loss_w = -\sum_{ic} w_i T_{ic} \cdot log(P_{ic})$$

  where $w_i$ is the survey weight of the $i$-th training observation, $c$ is the modality index in a one-hot representation, $T_{ic}$ is the ground-truth value of the target variable for the i-th observation, and $P_{ic}$ is the corresponding softmax function[2] output probability distribution of the MLP.
- The MLP output is a probability distribution on the 8 ALE items: for each record we impute the ALE item randomly extracted from the probability distribution of the corresponding pattern of covariates.

In a classical survey statistical framework, statistics are produced by weighting data with sampling weights. In practice, roughly speaking, a unit $i$ with a sampling weight $w_i$ represents other $w_i$ units of the target finite population. Following this idea, we apply MLP on the data set obtained by repeating each unit of the sample according to the attached sampling weight (MLP2, hereafter). To limit the error that is made by rounding the weight associated with each individual, we duplicate the profile or cell that corresponds to individuals who have the same values of the input variables. The weight associated with a profile is given by the sum of the weights of all individuals with the same profile. We remark that in this case MLP is estimated by means of the not weighted cross-entropy loss function. Finally, the step of imputation through a random draw from the estimated ALE probability distribution is performed, as previously described.

The two approaches, MLP1 and MLP2, are applied to the estimation of ALE in the 2018 Italian resident population.

DRAFT

## 3. Experimental results

ALE for the Italian resident population in 2018 is estimated by using administrative data (covering the period from 2011 to 2017), traditional Census data (2011 Census) and 2018 sample survey data. In this work, we use data relatively to the 312,813 people residents in Lombardia region with no missing data on ALE 2018. The target variable is the self-declared ALE in the 2018 sample survey, which corresponds approximately to $5\%$ of the total population of interest. Core information like age, gender, citizenship, marital status, place of birth and place of residence are available for all individuals. The classification adopted for ALE is composed by 8 items: 1 - Illiterate, 2 - Literate but no formal educational attainment, 3 - Primary education, 4 - Lower secondary education, 5 - Upper secondary education,

[1] equivalent to dummy variables in statistics but where the vector of all zero is not admitted

[2] softmax is a function which applies the standard exponential function to each element of the input vector and normalizes these values by dividing by the sum of all these exponentials

6 - Bachelor's degree or equivalent level, 7 - Master's degree or equivalent level, 8 - PhD level.

We apply MLP1 to sample survey data to obtain an estimate of the ALE variable. Then we apply MLP2 to the sample survey data expanded according to sampling weights as described above. Sample survey data provide the observed ALE for all the individuals allowing for supervised estimation and assessment of accuracy.

The results by MLP1 are produced with a neural network having two hidden layers, each of 128 neurons, and an output layer with 8 neurons (one per modality of the target variable). Instead the architecture of the neural network for estimating ALE by expanded dataset has three hidden layers, two of 256 neurons and one of 128 neurons, and an output layer with 8 neurons. For both the best configuration of some hyper-parameters (number of hidden neurons, dropout probability, learning-rate) was explored through a suitable grid-search, but the configuration found for the latter is more complex than the first one.

The comparison between the two approaches is evaluated by computing accuracy indicators and by using a $k$-fold approach with $k = 5$. In particular, the dataset is partitioned into 5 folds and

1. the model is estimated on the training set, consisting of 4 of the 5 folds;
2. the results are applied on the test set, composed of the remaining fold;
3. accuracy indicators are calculated only on the test set as the difference between estimated and observed ALE.

Tasks 1-3 are repeated 5 times so to reconstruct the entire data set. In order to take into account the variability introduced by the imputation obtained through a random draw from the estimated probability distribution, the procedure is repeated 100 times and the results are obtained as the average of indicators over these runs.

In Table 1, we report the percentage of correct classification obtained by comparing the imputation and the value observed in the sample (considered as the target value). This is an indicator measuring the error prediction, it is a micro-level accuracy indicator. The indicator is listed for each k-fold.

Table 1: Micro-level accuracy in the 5 test sets averaged over 100 runs: MLP1 *vs* MLP2.

| K-fold | MLP1 | MLP2 |
|---|---|---|
| 1 | 71.521 | 71.622 |
| 2 | 71.648 | 71.598 |
| 3 | 71.350 | 71.616 |
| 4 | 71.405 | 71.545 |
| 5 | 71.385 | 71.494 |
| Mean | 71.462 | 71.575 |
| St. Dev. | 0.110 | 0.049 |

The means of the micro accuracy computed in the two approaches are very similar (71.5 vs 71.6) while MLP2 shows a lower variance between folds due to the fact that each fold is a simple random sample of the target population.

In order to evaluate the two MLP procedures at a macro level, the estimated ALE is compared with the observed ALE, appropriately weighted. In particular, we focus on the differences between the frequency distributions of estimated ALE ($\widehat{ALE}$) and the target distribution computed on weighted sample data ($\widehat{ALE}_S$). A synthetic measure of the difference between distributions is given by the average of the absolute values of the differences between percentage of each item, in absolute (AD) and relative

DRAFT

(RD) terms. Specifically:

$$AD = \frac{\sum_i^8 D_i}{8} = \frac{1}{8}\sum_i^8 |fr(\widehat{ALE})_i - fr(\widehat{ALE}_S)_i| \quad (1)$$

$$RD = \frac{\sum_i^8 Dr_i}{8} = \frac{1}{8}\frac{\sum_i^8 |fr(\widehat{ALE})_i - fr(\widehat{ALE}_S)_i|}{fr(\widehat{ALE}_S)_i} \cdot 100$$

where $fr(\widehat{ALE})_i$ is the percentage frequency of estimated ALE item $i$ by MLP1 or MLP2 and $fr(\widehat{ALE}_S)_i$ is the target percentage frequency of ALE item i estimated with the weighted sample.

Table 2: Macro-level accuracy in the 5 test sets averaged over 100 runs: AD and RD indicators for MLP1 and MLP2.

| | AD | | RD | |
|---|---|---|---|---|
| K-fold | MLP1 | MLP2 | MLP1 | MLP2 |
| 1 | 0.131 | 0.045 | 3.069 | 2.231 |
| 2 | 0.086 | 0.072 | 3.327 | 3.324 |
| 3 | 0.129 | 0.104 | 5.476 | 5.013 |
| 4 | 0.118 | 0.088 | 3.098 | 2.496 |
| 5 | 0.119 | 0.106 | 4.101 | 3.909 |
| Mean | 0.117 | 0.083 | 3.814 | 3.395 |
| St. Dev. | 0.016 | 0.023 | 0.911 | 1.005 |

DRAFT

As shown in Table 2 the MLP2 gives origin to a slightly better macro-level accuracy: the mean AD decreases from 0.117 to 0.083 and the mean RD from 3.814 to 3.395. We remark that, the improvement is only due to the expanded format of the sample data and not to the use of additional information. The increase of variability for MLP2 with respect to MLP1 reflects the imputation phase of the procedures: for MLP1 the ALE item is randomly extracted from the probability distribution of the correspondent modality, whereas for MLP2 the extraction is made for each unit from the expanded dataset.

Table 3: Macro-level accuracy by ALE modality in test set 2 averaged over 100 runs: AD and RD indicators for MLP1 and MLP2 and percentage frequency of the ALE modality in the weighted sample (n%)

| | AD | | RD | | |
|---|---|---|---|---|---|
| | MLP1 | MLP2 | MLP1 | MLP2 | n% |
| Illiterate | 0.02 | 0.05 | 6.53 | 12.09 | 0.38 |
| Literate but no ed. Att. | 0.05 | 0.03 | 3.51 | 2.33 | 1.47 |
| Primary education | 0.09 | 0.13 | 0.62 | 0.83 | 15.21 |
| Lower secondary ed. | 0.12 | 0.09 | 0.41 | 0.32 | 28.14 |
| Upper secondary ed. | 0.15 | 0.14 | 0.38 | 0.35 | 36.09 |
| Bachelor's degree | 0.10 | 0.04 | 2.75 | 1.05 | 3.70 |
| Master's degree | 0.11 | 0.07 | 0.96 | 0.58 | 11.59 |
| PhD | 0.04 | 0.03 | 11.45 | 9.04 | 0.42 |

In Table 3, the difference between estimated and target ALE for each modality is reported for the two approaches. A relevant advantage in the use of MLP2 is obtained for the Bachelor's degree: this

modality is difficult to impute due to the heterogeneity of the individual course of studies.

To conclude, the standard MLP applied to the expanded dataset allows a gain in micro and macro level accuracy without any additional information, on the other hand, increasing the neural network complexity entails an increase of the computational time.

## References


[1] Byrd, J., Lipton, Z. What is the Effect of Importance Weighting in Deep Learning? In: Kamalika C, Ruslan S, editors. Proceedings of the 36th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2019. p. 872–81.

[2] Dagdoug, M., Goga, C., Haziza, D.: Imputation procedures in surveys using nonparametric and machine learing methods: an empirical comparison. Journal of Survey Statistics and Methodology **11**, 143–188 (2023).

[3] De Fausti, F., Di Zio, M., Filippini, R., Toti, S., Zardetto, D.: Multilayer perceptron models for the estimation of the Attained level of Education in the Italian Permanent Census. Statistical Journal of the IAOS, **38**, 637–646 (2022)

[4] De Fausti, F., Di Zio, M., Filippini, R., Toti, S., Zardetto, D.: The imputation of the 'Attained Level of Education' in the base register of individuals through Neural Networks using sampling weights. UN/ECE Work Session on Statistical Data Editing. Vienna, Austria, 3-7 October 2022

[5] Di Zio, M., Filippini, R., Rocchetti, G.: An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. Rivista di Statistica Ufficiale, **2-3**, 143–174 (2019)

[6] Hashemi M., Karimi H. Weighted machine learning. Statistics, Optimization and Information Computing 6 (4), 497-525

[7] MacNell, N., Feinstein, L., Wilkerson, J., Salo, P.M., Molsberry, S.A., Fessler, M.B., Thorne, P.S., Motsinger-Reif, A.A., Zeldin, D.C. Implementing machine learning methods with complex survey data: Lessons learned on the impacts of accounting sampling weights in gradient boosting. PLoS One. 2023 Jan 13;18(1).

[8] Pfeffermann, D.: The role of sampling weights when modeling survey data. International Statistical Review/Revue Internationale de Statistique, 143–174 (1993).


DRAFT

# Evaluation of pollution containment policies in the US and the role of machine learning algorithms

Marco Di Cataldo[a], Margherita Gerolimetto[a], Stefano Magrini[a], and Alessandro Spiganti[a]

[a]Ca' Foscari Univesity Venice; marco.dicataldo@unive.it, margherita.gerolimetto@unive.it, stefano.magrini@unive.it, alessandro.spiganti@unive.it

**Abstract**

The aim of this study is to analyse policy actions and institutional changes in local governance structures as determinants of air pollutant reductions in US urban areas. First, we construct a dataset on traffic-related air pollution and socio-economic characteristics across urbanized areas of the US. Some of these data are available through Google Earth engine, others are instead provided by institutional sources. In general, raw data come from application of machine learning techniques to either satellite images or monitoring station records and are available at different temporal and spatial resolutions. Then we adopt regression discontinuity design techniques for the evaluations of pollution reduction policies, exploiting the designation of US Transport Management Areas as a quasi-experimental framework.

***Keywords:*** air pollution, policy evaluation, regression discontinuity design, machine learning

DRAFT

## 1. Introduction

Road traffic is one of the main contributors to air pollution and greenhouse gasses around the world (among others McDuffie et al., 2021). The mixture of vehicle exhausts from fuel combustion and non-exhaust from engine, brake, tire, and road surface wear and re-suspended street dust materials significantly contributes to particulate matter (PM), nitrogen oxides (NOx), and carbon dioxide (CO2) emissions (European Environment Agency, 2016). These emissions disperse into the ambient air as traffic-related air pollution (TRAP), which degrades ambient air quality.

Humans exposed to TRAP are at a higher risk of developing a wide range of adverse health effects, from premature mortality and cardiovascular illness to cognitive and metabolic effects (Fu et al., 2021; Khreis, 2020). On top of these direct effects, there are additional societal burdens: medical costs, missed school days and workdays (among others Nurmagambetov et al. 2018), reduced workers' productivity (Chang et al. 2016), and brain drain (Xue et al. 2021).

Most human exposures to TRAP happen in urban areas (Kura et al., 2013). Even if air quality in western countries has improved enormously in the past decades, most cities around the world struggle to meet air quality standards and guidelines (World Health Organization). Since the number of urban residents is expected to grow rapidly around the world (United Nations and Department of Economic and Social Affairs Population Division, 2022), a greater quantity of people will soon be at risk of TRAP exposure.

With this in mind, the aim of this work is to analyse policy actions and institutional changes in local governance structures as determinants of air pollutant reductions in US urban areas over the last decade. We exploit the designation of Transport Management Areas (TMAs) as a quasi-experimental framework. TMAs are designated by the US Secretary of Transportation for urbanized areas that overcome the population threshold of 200,000 as defined by the Bureau of Census, in recognition of the complexity of transportation issues. They are subject to several transportation planning requirements among which a Congestion Management Process and an Air Quality Plan.

From the methodological point of view, we rely on Regression Discontinuity Design (RDD) techniques, a long-standing way to obtain credible causal estimates that is gaining increasing popularity in recent times (among others, Cattaneo and Titiunik, 2022). Like other causal inference approaches, the RDD can benefit from the combination with machine learning methods, both to carry out supplementary analyses enhancing the credibility of the results and to handle specific computational issues, such as bandwidth determination (Athey and Imbens, 2017).

As for the data, we construct a dataset on TRAP and socio-economic characteristics across urbanized areas of the US. Some of these data are available through Google Earth engine, others are instead provided by institutional sources like NASA or Environment Protection Agency (EPA). In general, raw data come from application of machine learning techniques to either satellite images or monitoring station records and are available at different temporal and spatial resolutions. The preliminary aim of this work is to gather and merge data on specific variables from different sources and harmonize them to produce a novel dataset to be used for the main objective of the paper that is evaluating via RDD the effectiveness of policy actions implemented in US urban areas with the objective of containing TRAP.

The structure of the paper is as follows. In the second section we present our methodological framework. In the third section we describe the model and the data set. In the last section we illustrate some preliminary results.

DRAFT

## 2. Methods

In this section we will present our methodological framework.

### 2.1 Regression Discontinuity Design

A large literature on causal inference for policy evaluations has focused on methods for statistical estimation to answer a question about the counterfactual impact of change in a policy (or treatment). The policy change has not necessarily being observed before or may have been observed for a subset of the population. The goal is then to estimate the impact of small set of treatments using data from randomized experiments or, more commonly, observational studies (that is non-experimental data). The literature identifies a variety of assumptions that, when satisfied, allow the researcher to draw the same type of conclusions that would be available from a randomized experiment.

Drawing inference about the causal effect of a policy from observational data is rather challenging. The main issue is that there are factors, which may be unobserved that are said confounders in the sense that they induce correlation that is not indicative of what would happened if the policy had been changed. In this sense, identification strategies are central to causal inference and in economics the RDD is among the most credible, because it relies on weak and easy to implement non parametric identifying assumptions which permit flexible and robust identification and inference for local treatment. The key feature of RDD is the existence of a score, or running variable, for each unit of the sample, which determines treatment assignment via hard-thresholding: all units whose score is above a known cutoff are treated, while all units below the cutoff are not treated. Identification, estimation, and inference proceed by comparing the outcomes of units near the cutoff taking those below (control group) as counterfactuals to those above (treatment group). For extensive literature reviews, see, among others, Lee and Lemieux (2010) and Cattaneo and Titiunik (2022).

In this work we adopt the potential outcome (or continuity) approach[1], introduced by Hahn et al.

[1] As a complement to the potential outcome approach, Cattaneo et al. 2015, introduced the local randomization

(2001), where potential outcomes are taken as random variables, with the $n$ units of analysis forming a random sample from an underlying population and the running variable, or score, $X$ is assumed to be continuously distributed. Let $Y_i(0), Y_i(1)$ denote the pair of potential outcomes for unit $i$ and let $T_i \in \{0, 1\}$ denote the treatment[2]. The realized outcome is $Y_i = Y_i(T_i)$. The treatment received is a function of the running (pretreatment) variable $X_i$, more specifically $T_i = I_{X_i \geq c}$, where $c$ denotes the threshold or cutoff, that must be exogenous. Formally, the treatment effect is defined from the following identity

$$\tau = E\left(Y_i(1) - Y_i(0)|X = c\right) \tag{1}$$

The two key assumptions for identifications are: a) the regression functions $E\left(Y_i(0)|X_i = x\right)$ and $E\left(Y_i(1)|X_i = x\right)$ are continuous in $x$ at $c$ and b) the density of the running variable near the cutoff is positive. These assumptions capture the idea that units that are barely above and below the cutoff $c$ would exhibit the same average response if their treatment status did not change. Then by implication, any difference between the average response of treated and control units at the cutoff can be attributed to the treatment and can be interpreted as the causal average effect, estimated as the discontinuity in the conditional expectation of $Y_i$ as a function of the running variable at the cutoff:

$$\tau = lim_{X \to c^-} E\left(Y_i|X_i\right) - lim_{X \to c^+} E\left(Y_i|X_i\right) \tag{2}$$

In practice, we have

$$\begin{array}{c|c} -h \leq X_i < c & c \leq X_i < h \\ Y_i = \beta_{0-} + (X_i - c)\beta_{1-} + \epsilon_{i-} & Y_i = \beta_{0+} + (X_i - c)\beta_{1+} + \epsilon_{i+} \end{array}$$

where $\hat{\tau}_{RD} = \hat{\beta}_{0+} - \hat{\beta}_{0-}$ is the vertical distance between the two estimated expectations, $h$ is bandwidth that guarantees that only units that are close to the cutoff $c$ are involved, $\epsilon_{i-}$ and $\epsilon_{i+}$ are error terms.

The idea is to estimate regression functions for control and treatment group locally and this, in its most basic fashion, can be done by estimating

$$Y_i = \alpha + \tau_{RD} T_i + (X_i - c)\beta_1 + \epsilon_i, \quad -h \leq X_i \leq h \tag{3}$$

where $\hat{\tau}_{RD}$ is the desired estimated discontinuity and $\epsilon_i$ is the error term. As mentioned above, the regression is estimated on a subsample of the data that is statistically optimally close to the cutoff so that units are most comparable each other and this should reduce the influence of the confounding factors. However, this reduces also the number of available observations and thus makes estimates and causal inference increasingly imprecise, limiting the ability to access policies. Thus one needs to identify an optimal bandwidth around the cut-off that optimally balances variance and bias. There are several methods to find the optimal bandwidth (e.g. Imbens and Kalyanaraman, 2012). A very interesting recent proposal is a machine learning based method developed by Long and Rooklyn (2020)

Typically researches tend to estimate model (3) adopting local polynomial methods tailored to flexibly approximate, above and below the cutoff, the unknown conditional mean function of the outcome variable given the running variable. In practise, researchers often choose a local linear polynomial and perform the estimation using weighted linear least squares, giving higher weights to observations close to the cutoff. If present, this discontinuity is interpreted as some average response to the treatment at the cutoff, depending on the assumptions and the setting under examination.

## 2.2 Machine Learning Algorithms

The machine learning literature has traditionally focused on discovering pattern and on prediction, using data-driven approaches to build rich models and relying on cross-validation as powerful tool for model selection.

framework that is built on the idea that near the cutoff the RD design can be interpreted as a randomized experiment or more precisely as a natural experiment.

[2]See for example Imbens and Rubin (2015) for more details on this set-up.

DRAFT

Supervised machine learning tools could be useful for causal inference given that, similarly to regressions, can summarise linear and non-linear relationships in the data and make predictions (Varian, 2014). However, given the lack of interpretable coefficients for some of the algorithms and the lack of standard errors of the obtained coefficients, predictions tools from these literature cannot be readily used for causal inference. Nevertheless, one particularly mature strand of literature includes approaches that incorporate supervised machine learning techniques in the so-called supplementary analyses to improve the credibility of the policy evaluations, such as placebo analysis, internal validity, external validity for RDD methods (Athey and Imbens, 2017). Moreover, in causal inference many estimators involve the specification of parameters, such as the optimal bandwidth in RDD, which are not interest per se, but are necessary to estimate the target parameter, and their determination can be done via machine learning (Long and Rooklyn, 2020). Overall, while most machine learning methods cannot be used to infer causal effects, they can surely help the process.

From a different perspective, machine learning methods can have a fundamental role in the data processing that might be preliminary to subsequent causal inference analyses. This comes from the literature on machine learning methods to estimate variables investigated in geo- and environmental sciences (Wuepper and Finger, 2023). For example, in case of greenhouse gas emissions, data might not be available and researchers must rely on proxies. Most natural phenomena are non-linear, multivariate, highly variable and correlated at many spatio-temporal scales. The analysis and treatment of such complex data and their integration/assimilation with science-based models is a difficult problem that is addressed by contemporary machine learning approaches, e.g. random forest algorithms to generate new variables even directly in Google Earth Engine (2022). Given the growing abundance and improving resolution (spatial, temporal, and spectral) of satellite imagery, in a recent review of the field, Burke et al. (2021) discuss the rapidly increasing literature regarding satellite imagery and measurements of different human outcomes, with specific attention to approaches that combine imagery with machine learning. Researchers in economics also increasingly use satellite imagery, and particularly nightlights imagery, for a variety of applications (among others, Henderson et al., 2012).

DRAFT

## 3. Model and Data

In order to evaluate the impact of policy actions and institutional changes in local governance structures on air pollutant reductions in US urban areas we will consider the TMAs designation as quasi-natural framework. TMAs are designated by the US Secretary of Transportation for urbanized areas that overcome the population threshold of 200,000 as defined by the Bureau of Census, in recognition of the complexity of transportation issues. When an urban area is designated as a TMA, the Metropolitan Planning Organization (MPO) responsible for that urban area (an MPO is mandatory for urban areas with population over 50,000) is subject to several transportation planning requirements among which a Congestion Management Process

Here, with reference to TMA designation after 2010 Census, we estimate a (very preliminary) RDD model for year 2015. The statistical units are the urbanized areas and the running variable is the population. Treated units are those that, overtaking the threshold $c = 200,000$, have been designed as TMAs. The RDD model we have in mind is:

$$Y_i = \alpha + T_i \tau_{RD} + (X_i - c)\beta + \mathbf{Z}_i \gamma + \epsilon_i, \quad -h \leq X_i \leq h \tag{4}$$

where for each statistical unit $i$, $Y$ is the level of traffic related pollutant C02, $X$ is the population (running variables), $\mathbf{Z}$ is a vector of covariates, specifically income (proxied by the nightlights) and meteorological variables, such as wind and precipitations, $\epsilon_i$ is the error term. The regression is run considering urbanized areas whose value of the running variable is close to the cutoff $c = 200,000$, according to the bandwidth $h$.

To conduct this analysis, we build a novel dataset, from the following data sources:

**Population, TMA designation** (source US Census Bureau and US Federal Register): these sources provide information also on the shapefiles for the administrative boundaries of the urbanized areas.

**Traffic-related CO2** (source NASA): DARTE (Database of Road Transportation Emissions) data set provides a 38-year, 1-km resolution inventory of annual on-road CO2 emissions for the conterminous United States based on roadway-level vehicle traffic data and state-specific emissions factors for multiple vehicle types on urban and rural roads.

**Nightlights** (source Google Earth Catalog): VIIRS Nighttime Day/Night Band Composites Version 1 provides monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB).

**Wind and precipitations** (source Google Earth Catalog): TerraClimate is a dataset of monthly climate and climatic water balance for global terrestrial surfaces.

## 4. Some results

In the presentation of some preliminary results, we show the estimation of the basic version of model (4), that is the RDD regression without covariates:

$$Y_i = \alpha + T_i\tau_{RD} + (X_i - c)\beta + \epsilon_i, \quad -h \leq X_i \leq h \tag{5}$$

Figure 1 shows the local estimation of model (5). It is a local polynomial estimation of degree 2, with triangular kernel where the bandwidth has been optimally selected by MSE minimization, as suggested in Calonico et al. (2014), leaving for a more advanced version of the paper the estimation with a bandwidth selected with machine learning criteria (see section Sect. 2.). In the graph it is evident the discontinuity at the cutoff of the conditional expectation of $Y$ as a function of the running variable: this represents the local effect of the treatment.

Moving now to Table 1, we can see that the treatment effect is significant at 10% level and it has a negative sign, as expected. This confirms the idea of a possible effect of the TMAs designation on the levels of traffic related C02, that will be further explored in a more advanced version of the paper through the estimation of model (4), i.e. in the version including covariates. However, given that the effect is not so strongly significant, we shall also consider other case studies of pollution containment policies.

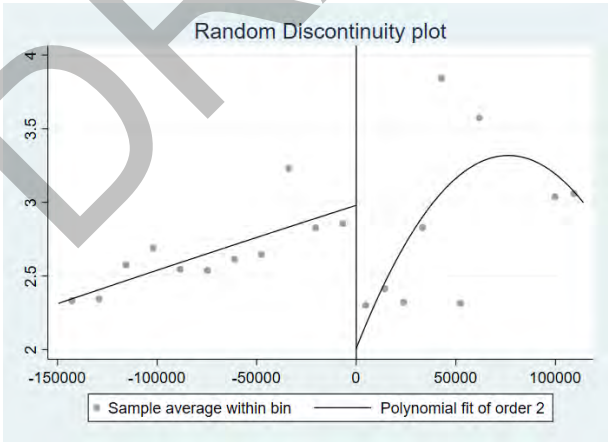



Figure 1: Random discontinuity plot for basic RDD with local polynomial regression

Table 1: Basic RDD estimation with local polynomial regression

| Method | Coeff | Std. Error | $z$ | $P > \|z\|$ |
|---|---|---|---|---|
| Conventional | -1.2167 | .68582 | -1.7741 | 0.076 |
| Robust | - | - | -1.7200 | 0.085 |

DRAFT

## References


1. Athey, S., Imbens, G.W.: The State of Applied Econometrics: Causality and Policy Evaluations. Journal of Economics Perspectives **31**, 3−32 (2017)
2. Burke, M., Driscoll, A., Lobell, D.B., and Ermon, S. . Using satellite imagery to understand and promote sustainable development. Science **371**, 1−12 (2021)
3. Calonico, S., Cattaneo, M., Titiunik, R.: Robust non parametric confindence intervals for regression discontinuity designs. Econometrica. **82**, 2295−2326 (2014)
4. Cattaneo, M. D.,Frandsen, B., Titiunik,R.: Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. Journal of Causal Inference **3**, 1−24 (2015)
5. Cattaneo, M. D., Titiunik, R.: Regression Discontinuity Designs. Annual Review of Economics, **14**, 821−851 (2022)
6. Chang, T., Graff Zivin J., Gross, T., Neidell, M.: Particulate Pollution and the Productivity of Pear Packers. American Economic Journal: Economic Policy, **8** 141−69. (2016)
7. European Environment Agency: European Environment Agency: Explaining Road Emissions. 10.2800/71804 (2016)
8. Fu, S., Viard, V. B., Zhang, P.: Air Pollution and Manufacturing Firm Productivity: Nationwide Estimates for China. The Economic Journal **131**, 241−3273 (2021)
9. Google Earth Engine (2022) `hiips://earthengine.google.com/`
10. Hahn, J., Todd, P., Van Der Klaauw,W.: Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica, **69**, 201−209 (2001)
11. Henderson J., Soreygard, A. Weil D.: Measuring Economic Growth from Outer Space, American Economic Review **102**, 994−1028 (2012).
12. Imbens, G., Kalyanaraman, K.: Optimal Bandwidth Choice for the Regression Discontinuity Estimator. The Review of Economic Studies, **79**, 933−959 (2012)
13. Imbens, G., Rubin, D.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press (2015)
14. Khreis, H.: Traffic, air pollution, and health. In: Nieuwenhuijsen, M.J., Khreis, H. (eds.), Advances in Transportation and Health. Elsevier (2020)
15. Kura, B., Verma, S., Ajdari, E., Iyer, A.: Growing Public Health Concerns from Poor Urban Air Quality: Strategies for Sustainable Urban Living. Comput. Water, Energy. Environ. Eng. **02**, 1−9 (2013)
16. Lee, D. S., Lemieux, T.: Regression Discontinuity Designs in Economics. Journal of Economic Literature, **48**, 281−355 (2010)
17. Long, M.C., Rooklyn, J.: NEXT: Stata module to perform regression discontinuity. Statistical Software Components S458238, Boston College Department of Economics, revised 13 Oct 2020 (2020)
18. McDuffie, E.E., Martin, R.V., Spadaro, J.V., Burnett, R., Smith, S.J., O Rourke, P., Hammer, M.S., van Donkelaar, A., Bindle, L., Shah, V., Jaegle, L.: Source sector and fuel contributions to ambient PM2. 5 and attributable mortality across multiple spatial scales. Nat. Commun. **12**, 1−12 (2021)
19. Nurmagambetov, T., Kuwahara, R., Garbe, P.: The Economic Burden of Asthma in the United States, 2008−2013. Ann. Am. Thorac. Soc. **15**, 348−356. (2018)
20. United Nations, Department of Economic and Social Affairs Population Division. World Population Prospects 2022: Summary of Results. UN DESA/POP/2022/TR/ NO. 3. (2022)
21. Varian, H.R.: Big data: new tricks for econometrics. Journal of Economics Perspectives **23**, 317−320 (2014)
22. Wuepper, D., Finger, R.: Regression discontinuity designs in agricultural and environmental economics. European Review of Agricultural Economics, **50**, 1−28 (2023)
23. Xue, S., Zhang, B., Zhao, X.: Brain drain: The impact of air pollution on firm performance. Journal of Environmental Economics and Management. **110**, 102546 (2021)


DRAFT

# Gaussian Processes and Deep Neural Networks for Spatial Prediction

Alex Cucco[a], Luigi Ippoliti[b], Nicola Pronello[b], Pasquale Valentini[b], and Carlo Zaccardi[b]

[a]National Heart and Lung Institute, Imperial College London, UK; a.cucco20@imperial.ac.uk
[b]University G. d'Annunzio, Chieti-Pescara; luigi.ippoliti@unich.it, nicola.pronello@unich.it, pasquale.valentini@unich.it, carlo.zaccardi@unich.it

DRAFT

**Abstract**

Spatial prediction, accounting for spatial dependence, has been a topic of significant interest in spatial statistics. While the Gaussian process (GP) has been a popular tool for spatial prediction, it can be limited by the need to estimate a stationary spatial covariance function and its computational demands for large datasets. As a result, deep neural networks (DNNs) have emerged as an alternative to standard approaches, such as kriging, for spatial prediction. This paper briefly discusses the use of DNNs with basis functions as input variables for modelling spatial dependence and obtaining spatial predictions.

***Keywords:*** Spatial Prediction, Gaussian process, Deep learning, Deep Gaussian processes, Low-rank models

## 1. Introduction

Spatial statistics is concerned with the analysis of data that have spatial locations associated with them. These spatial locations are used to model statistical dependence between the data, as they can provide important information about the underlying spatial process.

Statistical methods for modelling spatial data are well summarized in literature and relevant references are (2) and (4). The spatial data are treated as a single realization from a probability model that encodes the dependence through both fixed effects and random effects, where randomness is manifest in the underlying spatial process and in the noisy and usually incomplete measurement process. If $\mathbf{z} = \big(z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n)\big)'$ are measurements observed at $n$ spatial locations, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, from a real-valued spatial process $\big\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in D\big\}$, with $D \subset \mathcal{R}^d$, a widely used representation of the random process is:

$$Z(\boldsymbol{s}) = Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \tag{1}$$

where $\epsilon(\boldsymbol{s})$ is a sequence of uncorrelated Gaussian random variables, each with mean 0 and variance $\sigma^2_\epsilon$. The process $Y(\boldsymbol{s})$ corresponds to the true (latent) spatial process vector of interest for which we further assume the following decomposition:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta(\boldsymbol{s}) \tag{2}$$

where, $\mu(\boldsymbol{s})$, represents a spatial mean component and, $\eta(\boldsymbol{s})$, is a zero mean spatially-dependent random process with covariance matrix $\boldsymbol{\Sigma}_\eta$. Usually, the mean, $\mu(\boldsymbol{s})$, is a parametrized function of spatial covariates, $X_j(\boldsymbol{s}),\ j = 1, \ldots, p$.

Predicting or interpolating the variable of interest, $Y$, at some unobserved location $\boldsymbol{s}_0$, it is of particular interest in spatial analysis. Some data sets, in fact, may have missing values at some sites and these missing values may be separated or clumped. While separated missing values may occur in case of instrument malfunctions, an example of clamped missing values, may be represented by passive satellite images due to cloud presence.

In practice, many different methods exist for spatial prediction, and the choice of a given method will depend on the specific characteristics of the data being analysed. Gaussian processes, neural networks, and deep learning models are all powerful tools for spatial interpolation, and can be used to capture complex spatial structures and dependencies that may be difficult to model with traditional methods, such as kriging. This paper provide a brief introduction on the use of low-rank model representations for using these methods in the context of spatial interpolation.

## 2. Gaussian Processes

A popular assumption for the noise-free process $Y(\boldsymbol{s})$ of Eq. (2) is that it can be interpreted as a Gaussian Process. From a practical perspective, finite dimensional realizations of a GP are simply multivariate Gaussian and we can thus assume that $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\eta)$, where $\mathbf{Y} = \big(Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\big)'$, $\boldsymbol{\mu} = \big(\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n)\big)'$ and $\boldsymbol{\Sigma}_\eta$ depends on a parameter vector $\boldsymbol{\theta}$ which specifies the spatial covariance function. If the mean, $\mu(\boldsymbol{s})$, is taken to be a linear combination of covariates such that $\mu(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}$, where $\boldsymbol{x}(\boldsymbol{s}) = \big(1, x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s})\big)'$ and $\boldsymbol{\beta}$ are associated fixed effects, simple estimation of the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, can be obtained by maximum likelihood and predictions at new locations simply reduces to computing the conditional expectation $\hat{Y}(\boldsymbol{s}_0) = E\big[Y(\boldsymbol{s}_0)|\mathbf{z}\big]$.

## 3. Multi-Layer neural networks

Neural networks (NN) can be used for a variety of tasks, including, for example, nonparametric regression. One advantage of NN in this framework, is that they can model non-linear relationships between the input and output variables, which can be difficult to capture with traditional regression methods. The regression function is of the form, $E\big[Z(\boldsymbol{s}_i)\big] = f\big(\widetilde{\mathbf{x}}(\boldsymbol{s}_i)\big)$, where $\widetilde{\mathbf{x}}(\boldsymbol{s}_i)$, is a suitable vector of covariates and $f(\cdot)$ is a non-linear function. The unknown non-linear regression function $f(\cdot)$, which describes how the inputs $\widetilde{\mathbf{x}}(\boldsymbol{s}_i)$ are translated into outputs $Z(\boldsymbol{s}_i)$, is learnt through the specification of multiple layers (i.e., multi-layer perceptrons, MLP) of interconnected processing nodes, or neurons, which include: an input layer, one or more hidden layers, and an output layer. Each neuron in the hidden layer receives inputs from the neurons in the previous layer and computes a weighted sum of the inputs, which is then passed through an activation function to produce the output. Formally, if $L$ is the number of hidden layers in the model, and $P_l$ is the number of hidden neurons for layer $l$, the steps of the network compute the following quantities:

$$\begin{aligned} \widetilde{\boldsymbol{x}}_l(\boldsymbol{s}_i) &= f_l\Big(\mathbf{b}_l + \mathbf{W}_l\, \widetilde{\boldsymbol{x}}_{l-1}(\boldsymbol{s}_i)\Big), \;\; l = 1, \ldots, L \\ E\big[Z(\boldsymbol{s}_i)\big] &= f_{L+1}\Big(b_{L+1} + \sum_{j=1}^{P_L} w_{L+1,j}\, \widetilde{x}_{L,j}(\boldsymbol{s}_i)\Big) \end{aligned} \tag{3}$$

where $\widetilde{\boldsymbol{x}}_0(\boldsymbol{s}_i) = \widetilde{\boldsymbol{x}}(\boldsymbol{s}_i)$, $\widetilde{\boldsymbol{x}}_l(\boldsymbol{s}_i)$ represents a $P_l$-dimensional vector of neuron values at layer $l$, $\mathbf{W}_l$ is a $(P_l \times P_{l-1})$ matrix of weights, $\mathbf{b}_l$ is a $(P_l \times 1)$ vector of intercepts, and $f_l(\cdot)$ is the activation function that accounts for non-linear relationships between the inputs and the outputs. The choice of the activation functions is made before the training, and the weight matrices are estimated during the training process. In any case, the final activation function $f_{L+1}$ is chosen so as to match the support of $Z(\boldsymbol{s}_i)$ at the output layer. If $Z(\boldsymbol{s}_i)$ is real-valued, then $f_{L+1}$ is typically the identity function. Parameter estimation in NN is done using back-propagation and, during the training, the weights and biases of the neurons are adjusted to minimize the error between the predicted output and the actual output. For classical

DRAFT

regression problems, the fit accuracy is usually measured by the mean squared error (MSE) between observed and predicted values.

## 4. Deep Gaussian Processes and low-rank models

The two paradigms reviewed in Sect. 2. and 3. are complimentary in their scope. The popularity of GP is owed to their simplicity and model interpretability. However, it is also known that they suffer from the following three limitations:

1. applying GPs and Kriging requires estimating the spatial covariance function, which is commonly assumed to be stationary;
2. the mean of the process, $\mu(\boldsymbol{s})$, very often relies on the strong assumption of a linear covariate effect;
3. GPs and Kriging can be computationally prohibitive for large spatial datasets, since they involve computing the inversion of $\boldsymbol{\Sigma}_\eta$, which may be large in many applications.

Given that spatial processes can exhibit non-Gaussian distributions, non-stationarity, and complex spatial patterns, many practitioners are turning to neural networks to address these limitations. However, while neural networks are capable of estimating arbitrary non-linear covariate effects, they assume that the data units are independent. To address these challenges, deep neural networks (DNNs) have received considerable attention in the context of geospatial data and interpolation (see (12) for a comprehensive review).

In particular, recent approaches based on neural networks to model spatial data often use spatial coordinates or some form of transformation, such as distances or basis functions, as additional covariates (see, for example, (3)). These methods thus incorporate all spatial information (covariates and space) directly into the mean.

To provide a link between this approach and Eq. (1) and (2), we note that the process $Z(\boldsymbol{s})$ has the following basis-function representations:

$$\begin{aligned} Z(\boldsymbol{s}) &= \boldsymbol{x}(\boldsymbol{s})'\,\boldsymbol{\beta} + \eta(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \\ &= \boldsymbol{x}(\boldsymbol{s})'\,\boldsymbol{\beta} + \sum_{j=1}^{q} \alpha_j\,\psi_j(\boldsymbol{s}) + \xi(\boldsymbol{s}) \\ &= \widetilde{\mathbf{x}}(\boldsymbol{s})'\,\boldsymbol{\gamma} + \xi(\boldsymbol{s}), \end{aligned} \tag{4}$$

where $\tilde{\mathbf{x}}(\boldsymbol{s}) = \big(\boldsymbol{x}(\boldsymbol{s})',\ \boldsymbol{\psi}(\boldsymbol{s})'\big)$ is a $P$-dimensional dimensional vector of covariates, $\boldsymbol{\psi}(\boldsymbol{s}_i)$ is a $q$-dimensional expansion vector that maps the low-dimensional latent process, $\boldsymbol{\alpha}$, to $\eta(\boldsymbol{s})$, $\boldsymbol{\gamma} = \big(\boldsymbol{\beta}', \boldsymbol{\alpha}'\big)'$, and the error term $\xi(\boldsymbol{s})$ is a stochastic process independent of the basis-function coefficients. The inclusion of $\xi(\boldsymbol{s})$ in Eq. (4) accounts for the error incurred by using only a finite number of basis functions. It preserves variability of $Z(\boldsymbol{s})$, and in that sense, Eq. (4) is a statistical model, not an approximation, of spatial variability.

The representation given in Eq. (4) can be obtained as a result of the Karhunen-Loéve (KL) theorem (1). Furthermore, the basis functions in Eq. (4) may be defined by truncations of a countable basis, or they may simply be functions thought to be important for representing the spatial variability. For example, basis functions can come in the form of splines (11), wavelets (10), bisquare functions (5), Wendland functions (9), and finite elements (7). Physical basis functions or the output of a numerical model, such as an atmospheric-transport model could also be used as basis functions.

The relationship between the GP and a NN can be seen by first noting that, in a single layer NN, assuming $f_l$ is an identity function and no covariates $\boldsymbol{x}(\boldsymbol{s})$ are involved, $E\big[Z(\boldsymbol{s}_i)\big]$ results in a linear function of the basis functions $\psi_j(\boldsymbol{s})$ similarly to Eq. (4). Furthermore, upon the orthogonalisation of the set of basis functions, $\big(\widetilde{\mathbf{x}}_{L,j}(\boldsymbol{s}_1), \ldots, \widetilde{\mathbf{x}}_{L,j}(\boldsymbol{s}_n)\big)'$, $j = 1, \ldots, P_L$, Eq. (3) demonstrates that, for $P_L \to \infty$, an infinitely wide NN provides a representation of a GP. To this purpose, (8) and (6) establish that it is possible to fit a NN of infinite widths via Bayesian training using GP priors. Clearly, rather

DRAFT

than using infinitely wide NNs, an NN with large hidden layer widths and depths can be specified to approximate the process.

After selecting a family of basis functions, the training process of the neural network involves iteratively switching between feed-forward (obtaining the fits based on the current estimates of the weights) and back-propagation (updating the weights based on the current fits). In general, the NN can be trained over all the available sample though estimation can be expedited by splitting the data into smaller and disjoint mini-batches. After estimating the full set of parameters, say $\widehat{\boldsymbol{\Omega}}$, of the neural network, spatial prediction of the process at an unobserved location can be obtained by evaluating Eq. (3) at the desired site, denoted by $\boldsymbol{s}_0$.

## 5. Discussion

By drawing from recent literature on DNNs and GPs, we have briefly discussed a framework that uses low-rank model representation of a spatial process and basis functions as input variables in DNNs for modeling spatial dependence. This approach opens up avenues for exploring the potential of deep learning in spatial prediction. For future work, we plan to evaluate the effectiveness of neural networks for spatial prediction under different scenarios, including non-Gaussian and non-stationary processes, and to compare their performance with traditional prediction methods such as kriging. We will also investigate various choices of basis functions and show that, for certain options, specific forms of DeepKriging (3) can be obtained in both one- and two-dimensional settings. Additionally, we will explore methods for quantifying the uncertainty associated with predictions. Results from simulations as well as real environmental data, will be presented to demonstrate the effectiveness of the approach for spatial prediction.

DRAFT

## References


[1] Adler, R.J.: The Geometry of Random Fields. SIAM, Chichester (2010)

[2] Banerjee S., Carlin, B.P., Gelfand, A.E.: Hierarchical modeling and analysis for spatial data. CRC Press (2014)

[3] Chen, W., Li, Y., Reich, B.J., Sun, Y.: Deepkriging: Spatially dependent deep neural networks for spatial prediction. Stat. Sin. (2022) doi:10.5705/ss.202021.0277

[4] Cressie, N.: Statistics for Spatial Data. Wiley. Rev. ed., Hoboken (1993)

[5] Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B. **70** 209–226 (2008)

[6] Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., Bahri, Y.: Deep neural networks as Gaussian processes. In: International Conference on Learning Representations (2018)

[7] Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). J.R. Stat. Soc. Ser. B **73**, 423–98 (2011)

[8] Neal, R.M.: Priors for infinite networks. (Tech. Rep. no. crg-tr-94-1). University of Toronto. (1994)

[9] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S.: A multiresolution Gaussian process model for the analysis of large spatial datasets. J. Comput. Graph. Stat. **24**, 579–599 (2015)

[10] Vidakovic, B., Müller, P.: An introduction to wavelets. In: B. Vidakovic, P. Müller (eds.) Bayesian Inference in Wavelet Based Models, pp. 1–18. Springer, New York (1999)

[11] Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)

[12] Wikle, C.K., Zammit-Mangion, A.: Statistical deep learning for spatial and spatiotemporal data. Annu. Rev. Stat. Appl. **10**, 247–270 (2023)

# How can we explain Random Forests in a spatial framework?

Natalia Golini[a], Luca Patelli[b], and Xavier Barber[c]

[a]University of Torino, Department of Economics and Statistics Cognetti de Martiis, Lungo Dora Siena, 100A, Torino; natalia.golini@unito.it
[b]University of Pavia, Department of Economics and Management, Via San Felice al Monastero, 5, Pavia; luca.patelli01@universitadipavia.it
[c]Universidad Miguel Hernández de Elche, Centro de Investigación Operativa, Avenida de la Universidad, Elche; xbarber@umh.es

DRAFT

**Abstract**

Random Forest (RF) is a Machine Learning algorithm, very popular in environmental applications thanks to its flexibility and predictive performances. Even if its working mechanism is simple and intelligible, RF is considered a *black box* model since it prevents grasping how predictors are combined to generate the response variable predictions. This lack of interpretability represents a limitation of RF, especially when some knowledge is required on the response-predictors relationship from the decision-making perspective. In this work, we aim to explain RF using a Post-Hoc approach, i.e. by extracting a compact and simple list of rules from an estimated RF focusing on a spatial regression context. By means of a spatial dataset, we compare the final sets of rules and discuss the predictive accuracies of the standard RF and its *gold standard* for the case of spatially correlated data.

***Keywords:*** Explainable Machine Learning, inTrees, Post-hoc methods, Rule extraction, RF-GLS

## 1. Introduction

The Machine Learning (ML) era has given rise to complex and powerful methods that can process vast amounts of data and make predictions with remarkable accuracy. However, the inherent *black-box* nature of some of these techniques has raised concerns about their lack of interpretability. Often the term *interpretability* is used as a synonym for *explainability*, but actually they refer to two different concepts. According to Rudin *et al.* (11), interpretability is referred to models that are built to be interpretable, while explainability is obtained by applying further techniques to non-interpretable models in order to extract information. On the topic of explainable ML methods, the recent paper by Wikle *et al.* (14) is worth to be mentioned. In particular, the authors discuss the use of explainability techniques in spatial ML to understand the role of specific inputs in predicting environmental variables. Even if from a statistical point of view the gold standard would be to use interpretable ML methods, when this is not possible it is a good practice to try to extract information from non-interpretable ML methods that have proven good performance.

In this work, among ML techniques, we consider Random Forest which is well known for its high prediction accuracy. It is a non-parametric supervised algorithm that, thanks to its flexibility, can model complex non-linear relationships between the response variable (categorical or continuous) and the predictors (3). RF is defined as an ensemble model as the result of aggregating a set of decision trees. Each tree is the result of a recursive binary splitting process obtained using re-sampled data and a random

set of predictors evaluated at each node as splitting candidates. Given its adaptability, RF has also been widely applied in the spatial framework with different strategies to deal with the spatial autocorrelation of the data. Patelli *et al.* (10) have recently proposed a literature review and a novel taxonomy of the existing strategies adopted to adjust RF for spatially correlated data. In particular, the authors highlight that the most interesting strategy is the RF-GLS method proposed by Saha *et al.* (12), who extend the RF by estimating trees using generalized least squares (GLS). It was proven that RF-GLS outperforms the classical RF in the presence of spatial correlation, thus representing the gold standard to be used in the spatial framework.

In any case, spatially aware or not, RF remains a non-interpretable algorithm. However, it is possible to use specific methods to explain the RF resulting model, as described in the review by Haddouchi and Berrado (7). In particular, "Internal Processing" (IP) methods try to get "insights that are inherent to internal processing" providing a global overview of the model. "Post-Hoc" (PH) methods instead are based on RF post-processing, such as for example the "Rule Extraction" (RE) approaches (see e.g. inTrees (5), SIRUS (2), Node harvest (9) and RuleFit (6) among others). These methods aim to find a limited set of rules (each defined as the combination of predictors and split values) that is common to many trees in the RF and that allow representing the prediction mechanism of RF.

The main aim of this contribution is to verify if, for a spatial regression problem, there exist differences in the rules obtained by using - so far - the inTrees approach applied to two different cases: trees grown by RF-GLS and by a classical RF. We expect that taking or not into account the spatial correlation when implementing RF will have an impact also in its extracted rules. The analysis is carried out by using a dataset regarding daily meteorological records measured by 159 monitoring stations in Croatia. We present here preliminary results followed by a discussion on further steps.

DRAFT

## 2. Data and methods

The explainability of RF in the spatial framework is illustrated using meteorological daily data from the national network of 159 stations in Croatia for the year 2008, provided by the Croatian National Meteorological Service (available at `https://github.com/AleksandarSekulic/RFSI`). At this stage of the work, we have not considered the temporal dimension of the data confining the analysis to a single day: $14^{th}$ June 2008. The locations of the 151 stations working at this date are shown in Fig. 1. In particular, dots and crosses represent training and test data considered to implement the RF-GLS and RF algorithms. For this dataset, we randomly selected $90\%$ of the data (i.e., 135 observations) for training the algorithms and used the remaining $10\%$ of the data (i.e., 16 observations) for testing the algorithms. Croatia is a country located in southeastern Europe, bordering the Adriatic Sea. It has a diverse topography with flat plains in the east, a hilly central region, and mountainous terrain in the west. The response variable is the mean daily temperature[1] [`TEMP`], measured in degrees Celsius (°C). The minimum and maximum observed mean daily temperature values are 1.8°C and 21.5°C, respectively. The highest temperatures are recorded along the coast and at low altitudes. The variables used as predictors are latitude [`lat` (in meters)], longitude [`lon` (in meters)], distance-to-coastline [`HRdsea` (in km)], elevation [`HRDdem` (in meters)], wetness index [`HRtwi`], seasonal fluctuation [`ctd` (in days)], insolation (total incoming solar radiation) [`INSOL` (in Joules)], and Moderate Resolution Imaging Spectroradiometer land surface temperature [`MODIS.LST`] images. The dataset and predictors are detailed in (8) and references therein. In particular, this dataset was used by Sekulić *et al.* (13) to evaluate and compare the performance of a spatial interpolation method they proposed, i.e. the Random Forest Spatial Interpolation.

With the aim of obtaining simple, stable and accurate rules, we implemented the inTrees approach proposed by Deng (5) and implemented in the homonym `R` package `inTrees`[2]. The set of algorithms proposed in the work of Deng (5) can be applied to all tree ensemble methods to perform different tasks: extract, prune, select and summarize the rules. Each step is not mandatory, and the procedure can be tailored based on the specific explanatory necessity.

[1] On most meteorological stations `TEMP` is measured three times a day: at 7 am, 1 pm and 9 pm.
[2] `https://cran.r-project.org/web/packages/inTrees/index.html`

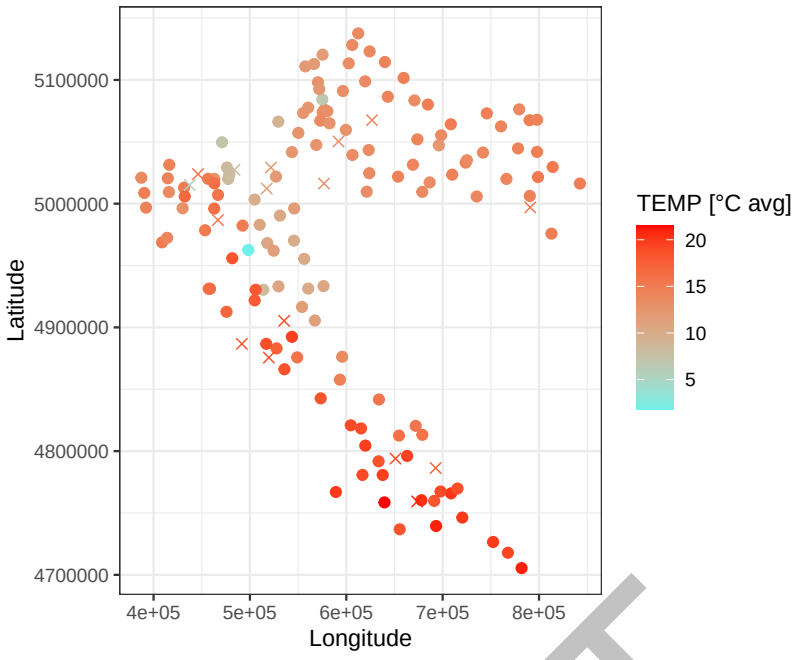



Figure 1: Mean daily temperature recorded on 2008-06-14 in 151 Croatian meteorological stations. Dots represent the mean daily temperature registered in the 135 training sites; crosses represent the mean daily temperature measured in the 16 test sites.

DRAFT

In order to extract and analyze rules by means of `inTrees`, the first step consists in running the chosen RF algorithm to have a collection of trees grown over a set of training data. Each tree results in the combination of all its splits, i.e. the conditions that permit splitting of the predictor space and getting predictions in the final regions. Then the obtained rules can be evaluated by using the relative "frequency" of occurrence, the prediction "error" and their "length" representing the rule complexity.

Using these metrics and considering opportune (decay) functions, the rules can be further simplified by pruning irrelevant predictor-split values. In order to have a compact rule set containing relevant and non-redundant rules, a complexity-guided condition selection method can be used, e.g. guided regularized Random Forest (GRRF) (4). In the end, the extracted rules can also be summarized by a rule-based learner that should be comparable in terms of prediction accuracy to the standard RF but more interpretable, named Simplified Tree Ensemble Learner (STEL). Note that in `inTrees` it is possible to build a STEL only for classification problems.

## 3. Preliminary empirical results

This section shows our preliminary results by applying the inTrees approach to extract insights from the RF-GLS and RF algorithms applied to the temperature spatial dataset.

We started by training the regression RF-GLS and RF on the same training set, by means of the R packages `randomForestGLS`[3] and `randomForest`[4], respectively. We used the same setting for the hyperparameters. In particular, we have set to 1000 the number of trees (`ntree` in `R`) and to 3 (one-third of the total number of predictors) the number of the variables randomly sampled as candidates at each split (`mtry` in `R`). For the RF-GLS, the covariance function used in modelling the spatial dependence structure among the observations was the default value, i.e. the exponential covariance function (`cov.mat` in `R`). Note that the coordinates [`lat`, `long`], measured in meters, have also been considered as predictors in both algorithms. In order to stabilize the forest structure, we followed the strategy pro-

[3] https://cran.r-project.org/web/packages/RandomForestsGLS/index.html
[4] https://cran.r-project.org/web/packages/randomForest/index.html

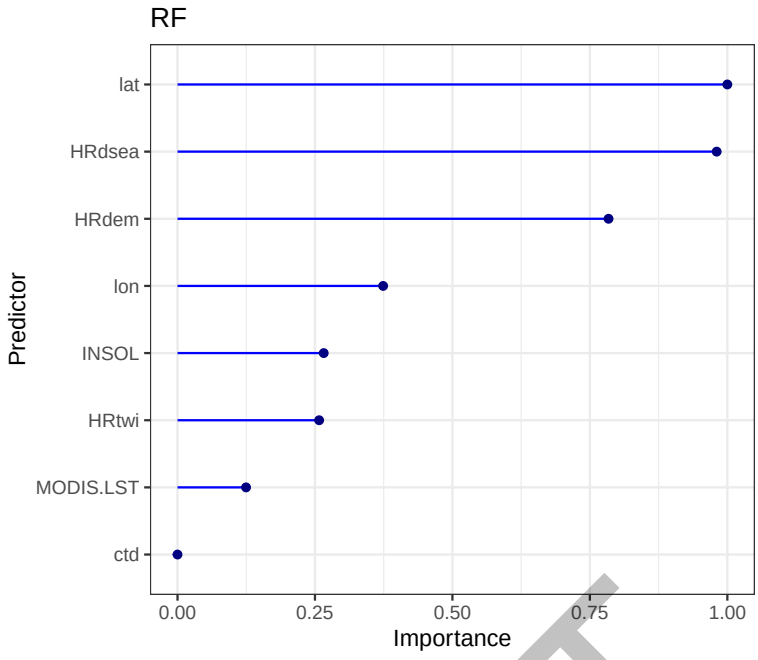



Figure 2: Variable importance plot for RF. The importance index is scaled to a maximum of 1.

DRAFT

posed in Bénard *et al.* (2) for rule generation consisting in restricting the node splits to the $q$-empirical quantiles of the predictors. This modification to Breiman's original regression tree algorithm is expected to have a small impact on predictive accuracy but is essential for stability.

Table 1 shows the test accuracy in terms of root mean square error and percentage of explained variance of the two algorithms when the node splits are restricted to the 10-empirical quantiles of the predictors. Different values of $q$ will be considered in the next steps of the work. As expected, RF-GLS shows a better predictive performance than RF because it is able to capture the spatial autocorrelation of the response variable.

Table 1: Root Mean Square Error (RMSE) and percentage of explained variance (Var explained) values evaluated for the test dataset.

| Algorithm | RMSE [°C] | Var explained [%] |
|---|---|---|
| RF-GLS | 1.057 | 93.52 |
| RF | 1.357 | 89.32 |

Latitude, distance-to-coastline and DEM are the most important predictors for RF (see Fig. 2). This information is not reported for RF-GLS since the `R` package `randomForestGLS` does not provide the variable importance as output yet.

Given the two forests, we applied the inTrees approach described in Section 2. For both algorithms, RF-GLS and RF, we used the same setting for the tuning parameters of the `inTrees` functions. We extracted the rule conditions from the set of trees with a maximum length of 3 (`maxdepth` in `R`) from each tree. The distinct rule conditions extracted from the 1000 trees of RF-GLS and RF were 2,836 and 3,007, respectively. Then, we assigned the outcome values (mean of the response variable values of the training observations that satisfy the condition) [pred] to the conditions and measured the quality of the rules by "frequency" [freq], "error" [err], and "length" [len]. We pruned the extracted rules' irrelevant or redundant variable-value pairs considering the metric "error" and the "relative" decay function. With the irrelevant variable-value pairs being removed, the pruned rules have shorter conditions and a frequency that increases without an increase in error. Finally, we applied the complexity-guided regularized random forest (GRRF) to the set of distinct pruned rules in order to have two compact lists of stable rules ($\leq 30$)

able to explain the results of both algorithms. We grew 1000 trees, setting the importance threshold to 0.1 and using the default values for the other tuning parameters of the function `selectRuleRFF` in `R`. From a run of this function we obtained a list of 19 and 25 rules starting from the forests grown by the RF-GLS and RF algorithms, respectively. By applying both these lists of rules to test data we obtained a very good predictive performance: the percentage of variance explained was 92.01 and 90.17, respectively.

Table 2 and Table 3 show the two lists of the first ten rules output for the meteorological dataset. The scores [impRF] of the selected conditions are calculated by building an RF on the selected rules. In general, the two lists of selected rules have 17 rules in common. An example is represented by the first rule in Table 2 and Table 3. More specifically, the first rule in both lists shows that the interaction of a low latitude with a low elevation and a low distance to the coastline induces a higher mean daily temperature. The third rule in Table 2 (and then the fifth rule in Table 3) displays that the interaction of low longitude and a high elevation induces a mean daily temperature of about 9°C. This is composed of two conditions (lon<= 589199.5 & HRdem>317.40), and satisfied by the 14.8% of the observations in the training dataset and has an RMSE (the square root of "err") of about 2.2°C. One can notice that rule scores (importance values) and the rules metrics are not related. For instance, the fourth rule in Table 2 (and then the second rule in Table 3) has a larger frequency than the three most important ones.

Table 2: First ten rules extracted, measured, pruned and selected via GRRF, generated by RF-GLS. The rules are ordered by scores (importance value - ImpRF)

| rule | len | freq | err | condition | pred | impRF |
|---|---|---|---|---|---|---|
| 1 | 3 | 0.252 | 3.082 | lat <= 4931735.37 & HRdem <= 609.20 & HRdsea <= 26.14 | 18.534 | 1 |
| 2 | 3 | 0.289 | 3.998 | lon > 457787.2 & HRdem<=609.20 & HRdsea <= 26.14 | 18.160 | 0.893 |
| 3 | 2 | 0.148 | 4.882 | lon<= 589199.5 & HRdem>317.40 | 9.325 | 0.673 |
| 4 | 2 | 0.681 | 5.564 | lat>4780743.3 & HRdsea>1.34 | 12.885 | 0.602 |
| 5 | 2 | 0.230 | 2.472 | lon>457787.2 & HRdsea<=1.34 | 18.714 | 0.586 |
| 6 | 3 | 0.148 | 4.882 | lon<=620344.9 & lat>4873835.0 & HRdem>317.40 | 9.325 | 0.577 |
| 7 | 3 | 0.230 | 2.690 | lat<=4931735.37 & HRdem<=317.40 & HRdsea<=26.14 | 18.743 | 0.505 |
| 8 | 3 | 0.148 | 4.882 | lat>4873835.0 & HRdem>317.40 & HRdsea<=195.44 | 9.325 | 0.489 |
| 9 | 2 | 0.259 | 5.886 | lat<=4931735.37 & HRdsea<=26.14 | 18.242 | 0.365 |
| 10 | 2 | 0.237 | 2.905 | lat<=4931735.37 & HRdem<=317.40 | 18.645 | 0.347 |

Table 3: First ten rules extracted, measured, pruned and selected via GRRF, generated by RF. The rules are ordered by scores (importance value - ImpRF)

| n | len | freq | err | condition | pred | impRF |
|---|---|---|---|---|---|---|
| 1 | 3 | 0.252 | 3.082 | lat<=4931735.37 & HRdem<=609.20 & HRdsea<=26.14 | 18.534 | 1 |
| 2 | 2 | 0.681 | 5.564 | lat>4780743 & HRdsea>1.34 | 12.885 | 0.560 |
| 3 | 3 | 0.148 | 4.882 | lon<=620344.9 & lat>4873835 & HRdem>317.40 | 9.325 | 0.487 |
| 4 | 3 | 0.148 | 4.882 | lat>4873835 & HRdem>317.4 & HRdsea<=195.44 | 9.325 | 0.486 |
| 5 | 2 | 0.148 | 4.882 | lon<=589199.5 & HRdem>317.40 | 9.325 | 0.475 |
| 6 | 3 | 0.267 | 5.879 | lon>457787.2 & lat<=4982676 & HRdsea<=26.14 | 18.212 | 0.439 |
| 7 | 2 | 0.230 | 2.472 | lon>457787.2 & HRdsea<=1.34 | 18.714 | 0.433 |
| 8 | 3 | 0.230 | 2.690 | lat<=4931735 & HRdem<=317.4 & HRdsea<=26.14 | 18.743 | 0.366 |
| 9 | 3 | 0.207 | 1.914 | lon>457787.2 & HRdsea<=1.34 & INSOL>8.082524 | 18.994 | 0.359 |
| 10 | 3 | 0.23 | 2.963 | lon>503554.1 & HRdem<=609.2 & HRdsea<=26.14 | 18.723 | 0.294 |

DRAFT

## 4. Discussion and next steps

This work represents a first attempt to "open" an RF that is specifically designed for spatially dependent data, i.e. RF-GLS. This algorithm should be the gold standard in a spatial framework. We compared

the predictive performance and explainability of RF-GLS and RF applied to a Croatian meteorological dataset. Both algorithms have shown high and similar predictive performance in our application. A cross-validation procedure will be implemented to confirm this result. Among the different approaches existing in the literature to obtain explainability from RF, we focused on the rule extraction methods. In particular, we considered the approach proposed by Deng (5) applying the same constraints to the node splits proposed in Bénard *et al.* (2). We found two compact lists of rules with high predictive performance sharing a large number of rules in common. However, the shared rules have different scores (importance values) within their respective membership lists. As next step, we aim to tune the GRRF hyperparameters to reduce the number of rules in the two lists while maintaining their predictive performance. Moreover, we aim to set up a comparison study considering the main competitors of inTrees, i.e. SIRUS (2), Node harvest (9) and RuleFit (6). Unfortunately, the `R` functions implementing RF-GLS (`RFGLS_estimate_spatial` and `RFGLS_predict_spatial`) return objects that are not valid inputs for the `R` functions implementing the competitor rule extraction methods. This will require further investigation.

## References


[1] Aria, M., Cuccurullo, C., Gnasso, A.: A comparison among interpretative proposals for Random Forests. MLWA **6**, 100094 (2021)

[2] Bénard, C., Biau, G., Da Veiga, S., Scornet, E.: Interpretable random forests via rule extraction. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, pp. 937–945 (2021)

[3] Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)

[4] Deng, H., Runger, G.: Gene selection with guided regularized random forest. Pattern Recognit. **46**, 3483–3489 (2013)

[5] Deng, H.: Interpreting tree ensembles with intrees. Int J Data Sci Anal. **7**, 277–287 (2019)

[6] Friedman, J. H., Popescu, B. E.: Predictive learning via rule ensembles. Ann Appl Stat. **2**, 916-954 (2008)

[7] Haddouchi, M., Berrado, A.: A survey of methods and tools used for interpreting random forest. In: Proceedings of the 2019 1st International Conference On Smart Systems And Data Science (2019) doi:10.1109/ICSSD47982.2019.9002770

[8] Hengl, T., Heuvelink, G.BM., Perčec Tadić, M., Pebesma, E.J.: Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. Theor Appl Climatol. **107**, 265–277 (2012)

[9] Meinshausen, N.: Node harvest. Ann Appl Stat. **4**, 2049–2072 (2010)

[10] Patelli, L., Cameletti, C., Golini, N., Ignaccolo, R.: A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review. arXiv **2303.04693** (2023)

[11] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Stat Surv. **16**, 1–85 (2022)

[12] Saha, A., Basu, S. Datta, A.: Random forests for spatially dependent data. JASA **118**, 665–683 (2023)

[13] Sekulić, A., Kilibarda, M., Heuvelink, G. BM, Nikolić, M., Bajat, B.: Random forest spatial interpolation. Remote Sens. **12**, 1687 (2020)

[14] Wikle, C., Datta, A., Hari, B., Boone, E., Sahoo, I., Kavila, I., Castruccio, S., Simmons, S., Burr, W., Chang, W.: An illustration of model agnostic explainability methods applied to environmental data. Environmetrics **34**, e2772 (2023)


DRAFT

# Recent approaches in coupling deep learning methods with the statistical analysis of spatial point patterns

Jorge Mateu[a] and Abdollah Jalilian[b]

[a]Department of Mathematics, University Jaume I of Castellon, Spain; mateu@uji.es
[b]Department of Statistics, Razi University, Kermanshah, Iran; stat4aj@gmail.com

**Abstract**

Although literature on spatial point process models and associated techniques is now widely available, in problems related to multivariate settings and classification the methodology is scarce and far from being flexible enough. We here provide a mathematical framework for coupling neural network methods with the statistical analysis of point patterns with a focus on two problems. We first use deep convolutional neural networks and employ a Siamese framework to build a discriminant model for distinguishing structural differences between spatial point patterns. Then, we discuss an example of deep neural networks in the statistical analysis of highly multivariate spatial point patterns and provide a new strategy for building spatio-temporal point processes using variational autoencoder generative neural networks.

***Keywords:*** Classification, Deep neural networks, Log-Gaussian Cox processes, Multi-layer perceptron, Siamese architecture, Spatial point processes

DRAFT

## 1. Introduction

Locations of trees in a forest stand, epicenters of earthquakes in a geographical region, or locations of crime incidents in a city are typical examples of spatial point pattern data. Many point process models, statistical inference approaches, and data analysis methods have been developed for spatial and spatio-temporal point patterns on different domains (12; 5; 13).

However, there is still room for substantial improvements and further developments in some areas of point pattern analysis. In particular, point process models and statistical methods for analyzing a set of spatial point patterns of $m \geq 2$ groups (populations, species, types, or other categorical characteristics), each observed at $T \geq 2$ different time instances on the same observation window are scarce. Although several multivariate point process models have already been developed (9; 15; 14; 2; 6), due to the curse of dimensionality, modeling and statistical analysis of highly multivariate point patterns observed at several time instances are both theoretically and computationally challenging.

In a related context within the analysis of spatial point patterns, it is important to distinguish between random and structural differences among observed point patterns from several populations. For example, analyzing the ecological distance or dissimilarity between species can reflect the effect of underlying processes on the spatial structure of ecological communities. Point pattern matching in pattern recognition and computer vision tasks also tries to quantify how one set of point patterns differs from another set of point patterns by considering isometric transformations and random noise or jittering of the position of points in each pattern (16; 1). Thus, detecting and quantifying relative similarities and

differences at large and small scales among partially observed point patterns provide an understanding of hidden differences in their corresponding populations. Moreover, such similarities and differences can be used to classify an unlabeled spatial point pattern into one of the populations.

Recent rapid developments in machine learning algorithms and methodologies have provided data analysis tools for various types of data (7). Because of the flexibility, generalization ability, and scalability of machine learning algorithms, they can be employed for statistical analysis of complex and highly multivariate data with spatial and temporal attributes (8). This makes spatial point processes and neural networks a convenient couple for statistical analysis of highly multivariate point patterns observed on several time instances. Consequently, more attention has recently been directed to merge machine learning algorithms in general, and deep neural networks in particular, into the statistical analysis of spatial point patterns.

Focusing on the modeling of the intensity functions of highly multivariate point processes, (17) used kernel density estimation and a variational autoencoder approach for analyzing the inhomogeneity in point patterns of location-based social networks. (4) provided a general framework for supervised statistical learning for point processes, and discussed cross-validation through independent thinning and the performance evaluation metrics in terms of bivariate innovations for spatial point patterns. (10) used deep convolutional neural networks and employ a Siamese framework to build a discriminant model for distinguishing structural differences between spatial point patterns. Recently, (11) provided a mathematical framework for coupling neural network methods with the statistical analysis of point patterns.

This short contribution summarises the ideas developed in these two previous papers (10; 11). On the one side, we use deep convolutional neural networks to extract features in a given spatial point pattern. Deep convolutional neural networks provide a flexible yet tractable family of transformations, are suitable for data with spatial correlations, and have been widely used in pattern recognition and image classification (8). We then use a Siamese framework of deep convolutional neural networks to construct a parametric discriminant model to distinguish structural differences in a set of spatial point patterns from several groups, each with a number of replicates. On the other side, we discuss an example of deep neural networks in the statistical analysis of highly multivariate spatial point patterns with the aim of providing a new strategy for building spatio-temporal point processes using variational autoencoder generative neural networks.

DRAFT

## 2. Statistical and neural network methodology

### 2.1 Dissimilarity and classification for spatial point patterns

Given two spatial point patterns, random similarities or differences between them provide no information about the underlying differences between their corresponding generative point processes, and only structural similarities or differences are of interest. To this end, major determinants of given point patterns, that include the most relevant information about the underlying point processes that have generated the observed point patterns, must be extracted by a suitable transformation. This transformation has to reflect appropriate large and small scale features and/or incorporate necessary first, second, and higher-order characteristics of a spatial point process and ignore less pronounced and redundant aspects. Such transformation is called feature extraction in machine learning and pattern recognition literature.

For a given pair of observed point patterns $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, with $\mathcal{X}$ denoting the space of all spatial point patterns on the observation window $W$, it is of interest to know whether $\mathbf{x}$ and $\mathbf{x}'$ are *structurally* different or if their differences are due to chance (random or unstructured differences). In fact, any difference between $\mathbf{x}$ and $\mathbf{x}'$ is due to pure chance if their corresponding point processes, $X$ and $X'$, are statistically indistinguishable; i.e., $\mathbb{P}_X = \mathbb{P}_{X'}$ or equivalently $f_X = f_{X'}$, with $\mathbb{P}$ being the probability distribution (or measure), and $f$ the corresponding density function. Even in the case that $\mathbb{P}_X \neq \mathbb{P}_{X'}$ ($f_X \neq f_{X'}$), some of differences between $\mathbf{x}$ and $\mathbf{x}'$ are still due to chance, but it is expected to observe some further differences that reflect the structural differences between the probability distributions of $X$ and $X'$. Thus, random similarities or differences between $\mathbf{x}$ and $\mathbf{x}'$ provide no information about the underlying differences between $X$ and $X'$, and we are only interested in structural similarities or

differences.

We can have dissimilarities by pattern or feature matching. A dissimilarity (distance) function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ takes any pairs of point patterns $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and quantifies the differences between them by a real number $D(\mathbf{x}, \mathbf{x}')$. For example, dissimilarity functions can be based on distances between quadrat counts or pairwise point-to-point distances between observed point patterns. A suitable dissimilarity function takes small values when $\mathbb{P}_X \approx \mathbb{P}_{X'}$, or equivalently $f_X \approx f_{X'}$, and large values when $\mathbb{P}_X$ and $\mathbb{P}_{X'}$, or $f_X$ and $f_{X'}$ are very different. Thus, we want to extract only major determinants (called features) of $\mathbf{x}$ and $\mathbf{x}'$ that include most relevant information about $\mathbb{P}_X$ and $\mathbb{P}_{X'}$ and remove less pronounced and redundant aspects.

Let $G : \mathcal{X} \to \mathcal{F}$ be a transformation that maps any observed point pattern $\mathbf{x} \in \mathcal{X}$ into a point in the feature space $\mathcal{F}$, which is assumed to be a metric space with metric $D_{\mathcal{F}}$. Then, the dissimilarity function

$$D(\mathbf{x}, \mathbf{x}') = D_{\mathcal{F}}\left(G(\mathbf{x}), G(\mathbf{x}')\right) \tag{1}$$

captures and quantifies differences between any pair of point patterns $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ in terms of their extracted features by $G$ (3). If $G$ is selected such that it incorporates the most relevant information about $\mathbb{P}_X$, then Eq. (1) compares structural differences between $\mathbf{x}$ and $\mathbf{x}'$. The transformation $G(\mathbf{x})$ might be chosen to be any summary statistic of the observed point pattern $\mathbf{x}$ such as an estimate of the intensity function, pair correlation function, nearest neighbor distance distribution function, empty space function or $J$ function.

To use convolutional neural networks (CNNs), the data are required to have a known grid-like topology. For the ease of representation, we consider the planar case $W \subset \mathbb{R}^2$ and partition the observation window $W$ into a $d_1 \times d_2$ regular grid of cells $B_{ij}$, $i = 1, \ldots, d_1, j = 1, \ldots, d_2$. Then, any point pattern $\mathbf{x} \in \mathcal{X}$ can be discretized over partition cells by cell counts $x_{ij} = n(\mathbf{x} \cap B_{ij})$. The discretized point pattern $\tilde{\mathbf{x}} = [x_{ij}]$ is the input layer of the network and it can be thought of as an image, which consists of a two-dimensional grid of cell counts as pixel values. Feature maps are obtained by repeating the convolution and pooling layers a suitable number of times to obtain a final feature vector $\mathbf{G} = [g_{k'}]$, which contains all biases and weights of the layers. Note that the elements of $\mathbf{G}$ encompass more exhaustive information about $\mathbb{P}_X$ than elements of all previous layers.

Let $\boldsymbol{\vartheta}$ denote a parameter vector that contains all biases, kernels of convolution layers and weights and biases of the last fully connected layer. Then, given $\boldsymbol{\vartheta}$, the convolutional neural network is a very flexible transformation

$$G_{\boldsymbol{\vartheta}} : \mathbf{x} \mapsto \mathbf{G} \tag{2}$$

from the observation space $\mathcal{X}$ to the feature space $\mathcal{F} = [0, 1]^{\ell_L}$ that maps a discretized point pattern $\tilde{\mathbf{x}}$ to its corresponding final feature vector $\mathbf{G}$. Note that $\mathbf{x}$ is arbitrary and no assumption, such as homogeneity or isotropy, is made on the underlying generative point process of $\mathbf{x}$.

In our context, let $\mathcal{D} = \{\mathbf{x}^{(s,t)}, s = 1, \ldots, m, t = 1, \ldots, T\}$ be a set of spatial point patterns of $m \geq 2$ groups, each observed with $T \geq 4$ replicates on the same observation window $W$. A relevant research question here is how we can use the observed point patterns in $\mathcal{D}$ to discriminate differences between groups. It falls under the topic of discriminant analysis, which is concerned with the relationship between the grouping variable $s$ of a point pattern $\mathbf{x} \in \mathcal{D}$ and its extracted features. And we use a Siamese framework (3) to construct a parametric discriminant model to distinguish differences between any pair of observed point patterns in $\mathcal{D}$ based on their extracted features.

## 2.2 Variational autoencoder generative neural networks for multivariate spatial point patterns

In this second part, we use generative neural networks within the context of point process methodology to provide a predictive statistical model in the context of log-Gaussian Cox processes. In particular, we discuss a framework to construct predictive models for multivariate (multi-type) spatial point patterns using multilayer perceptrons (also known as feedforward) neural networks to relate the spatial point patterns to a small set of latent random fields and obtain a generative variational autoencoder model. (15)

DRAFT

considered common and species-specific hidden Gaussian random fields to account for unobserved influential factors, and analyzed the environmental inhomogeneity and intra and inter-specific interactions of $m = 9$ species in only one temporal instant. We provide an extension of such multivariate framework to the spatio-temporal case by considering spatio-temporal independent zero mean and unit variance Gaussian random fields, accounting for species specific random effects and intra-species spatio-temporal correlations, together with inter-species spatio-temporal correlations with different scales. Parameter estimation is done through variational inference that approximates the corresponding posterior density. A variational distribution under the neutral hypothesis of no inter- and intra-specific interactions among species is used.

## 3. Data analysis of the Barro Colorado Island (BCI)

The paper is wrapped up with reporting several analysis carried out over BCI data. For the classification problem, we restrict our attention to the $m = 130$ most abundant species with 1808725 alive trees in $T = 8$ censuses, while we consider the $m = 100$ most abundant species in the BCI data with 1767880 trees in the 8 censuses to fit the multivariate spatio-temporal LGCP model. See complete details in (10; 11).

DRAFT

## References


[1] Caetano, T.S., Caelli, T., Schuurmans, D., Barone, D.A.C.: Graphical models and point pattern matching. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1646–1663 (2006)

[2] Choiruddin, A., Cuevas-Pacheco, F., Coeurjolly, J.F., Waagepetersen, R.: Regularized estimation for highly multivariate log gaussian cox processes. Stat. Comput. **30**, 649–662 (2020)

[3] Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 539–546 (2005)

[4] Cronie, O., Moradi, M., Biscio, C.A.: Statistical learning and cross-validation for point processes. arXiv preprint arXiv:2103.01356 (2021)

[5] Diggle, P.J.: Statistical Analysis of Spatial and Spatio-temporal Point Patterns. Chapman and Hall/CRC (2013)

[6] Eckardt, M., Gonzalez, J.A., Mateu, J.: Graphical modelling and partial characteristics for multitype and multivariate-marked spatio-temporal point processes. Comput. Stat. Data Anal. **156**, 107–139 (2021)

[7] Golden, R.M.: Statistical Machine Learning: A Unified Framework. Chapman and Hall/CRC (2020)

[8] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press. http://www.deeplearningbook.org (2016)

[9] Jalilian, A., Guan, Y., Mateu, J., Waagepetersen, R.: Multivariate product-shot-noise cox point process models. Biometrics **71**, 1022–1033 (2015)

[10] Jalilian, A., Mateu, J.: Assessing similarities between spatial point patterns with a Siamese neural network discriminant model. Adv. Data Anal. Classif. **17**, 21–42 (2023)

[11] Jalilian, A., Mateu, J.: Spatial point processes and neural networks: a convenient couple. Spat. Stat. (2023). doi: 10.1016/j.spasta.2022.100644.

[12] Møller, J., Waagepetersen, R.: Modern statistics for spatial point processes. Scand. J. Stat. **34**, 643–684 (2007)

[13] Møller, J., Waagepetersen, R.: Some recent developments in statistics for spatial point patterns. Annu. Rev. Stat. Appl. **4**, 317–342 (2017)

[14] Rajala, T., Murrell, D.J., Olhede, S.C.: Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection. J. R. Stat. Soc., C: Appl. Stat. **67**, 1237–1273 (2018)

[15] Waagepetersen, R., Guan, Y., Jalilian, A., Mateu, J.: Analysis of multispecies point patterns by

using multivariate log-gaussian cox processes. J. R. Stat. Soc., C: Appl. Stat. **65**, 77–96 (2016)
[16] Wang, H., Hancock, E.R.: A kernel view of spectral point pattern matching. In: Joint IAPR International Workshops on Statistical Techniques in PatternRecognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp. 361–369 (2004)
[17] Yuan, B., Wang, X., Ma, J., Zhou, C., Bertozzi, A.L., Yang, H.: Variational autoencoders for highly multivariate spatial point processes intensities. In: International Conference on Learning Representations (2019)

DRAFT

# A Kernel-based Nonparametric Multivariate CUSUM for Location Shifts

Konstantinos Bourazas[a,c], Konstantinos Fokianos[a], Christos Panayiotou[b,c], and Marios Polycarpou[b,c]

[a]Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus;
bourazas.konstantinos@ucy.ac.cy, fokianos.konstantinos@ucy.ac.cy

[b]Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus;
panayiotou.christos@ucy.ac.cy , mpolycar@ucy.ac.cy

[c]KIOS Research and Innovation Center of Excellence, University of Cyprus, Nicosia, Cyprus

DRAFT

**Abstract**

Online shift detection in multivariate data with unknown parameter settings is quite challenging. In statistical process control and monitoring, various self-starting methods that have been proposed aim to overcome such issues. In high dimensions, any distributional assumption, with the most typical being that of normality, can be strict or even unrealistic, calling into question the efficient performance of any method. In this work, we propose a Nonparametric multivariate CUSUM-type control chart to detect translocations for the mean vector. The proposed methodology is adaptive, free from tuning parameters that need to be set by the user.

*Il rilevamento in real-time del cambiamento della media di dati multivariate caratterizati da parametri sconosciuti è piuttosto complesso. In controllo statistico di processo e monitoraggio, vari metodi auto-avvianti sono stati proposti come possible soluzione. Quando i dati presentano alta dimensionalità, fare assunzioni sulla loro distribuzione, inclusa la tipica assunzione di distribuzione Gaussiana, può essere limitante, o persino irrealistico, mettendo in discussione l'efficienza dei risultati ottenuti da qualsiasi metodo. In questo lavoro, proponiamo una carta di controllo non parametrica multivariata di tipo CUSUM per rilevare il mutamento della media dei dati. La metodologia proposta è adattiva e non presenta parametri di configurazione che un utente deve inizializzare/impostare.*

***Keywords:*** Adaptive, CUSUM, Location shifts, Kernel Density Estimation, Self-Starting

## 1. Introduction

Multivariate detection schemes have been increasingly common in the last decades, as they allow monitoring a process by considering associations between the tested variables. The interest is placed on online monitoring procedures, as they can detect a disorder in real-time. A process is called In Control (IC) when it operates under such common causes of variation only. We refer to a process as Out Of Control (OOC) when special/assignable causes of variation are present. Statistical Process Control and Monitoring (SPC/M) is a set of powerful statistical methods that aim to detect early the presence of unusual observation/irregularities.

Several multivariate methods in SPC/M have been proposed for detecting a system disorder for a continuous monitoring process; an extensive literature review is provided in (1). In most of the methods, the standard setup requires a long training phase (phase I) under in control conditions for performing calibration before starting the testing (phase II). This requirement is restrictive when preliminary data are not available. Other limitations (especially for multivariate processes) are the requirement to define a set of tuning parameters or/and the parametric assumption, i.e., that the data follow a known distribution; for example see (5) or (2). Nonparametric methods have been introduced in the literature to relax the latter constraint. However, in many cases, they also have challenging issues when applied in practice due to, for example, not considering dependencies between variables (e.g., (4)) or requiring strict assumptions such as prior knowledge of both the IC and the OOC state; (7).

In this work, we focus on detecting location shifts in multivariate processes by proposing a nonparametric CUSUM-type chart. The proposed methodology is an online cumulative likelihood ratio test of two competing distributions, estimated through the Kernel Density Estimation (KDE) method. The use of KDE allows us to have the information for the entire IC distribution and not just one summary statistic, which is advantageous, especially for detection. Furthermore, the KDE-CUSUM is self-starting, as it can provide testing without the requirement of any lengthy historical dataset. The main contribution of the method is that it is adaptive, because it does not require any preliminary specification for the magnitude or the direction of a change.

The paper is organized as follows. In Section 2, we provide the theoretical background and the KDE-CUSUM derivation, while in Section 3, we discuss the stopping rule and the appropriate performance metrics. A short simulation study is presented in Section 4. Finally, Section 5. provides a short discussion, highlighting a few points that deserve further work.

DRAFT

## 2. Statistical modeling

An online change point model can be considered as sequential testing between two hypotheses (states); the IC state with distribution function $F_0$ against the OOC with distribution function $F_1$. Assuming that these distributions are known is strict and unrealistic, especially for multivariate processes. In practice, the pre-change density $f_0$ may be known under specific conditions, but it is highly restrictive to assume that the post-change density $f_1$ is known in advance. For this reason, we will adopt the nonparametric approach to estimate the unknown IC density via kernel smoothing.

In this setup, assume a random sample of multivariate data is obtained sequentially, and at time $t$ we have $\boldsymbol{x}_{1:t} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t)$. Each observation $\boldsymbol{x}_i$ is $p$-dimensional and specifically $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})$. The KDE at time $t$ will be based on the previous data $\boldsymbol{x}_{1:(t-1)}$. We study the KDE in the context of a mean process change. The key idea for suitably expressing the OOC state is to shift the parameters of the IC state, estimating the potential translocation based on the most recent data points. Thus, the estimated OOC density will be a translocated version of the IC density. We test whether the observation $\boldsymbol{x}_t$ supports the IC state or the OOC state by considering the following ratio 1 with estimated densities $\hat{f}_0$ and $\hat{f}_1$, i.e.:

$$L_t = \frac{\hat{f}_1(\boldsymbol{x}_t|\boldsymbol{x}_{1:(t-1)})}{\hat{f}_0(\boldsymbol{x}_t|\boldsymbol{x}_{1:(t-1)})} = \frac{\sum_{i=1}^{\hat{\tau}-1} K_{\boldsymbol{H}}\left(\boldsymbol{x}_t - (\boldsymbol{x}_i + \boldsymbol{\delta}_{\hat{\tau}:(t-1)})\right)}{\sum_{i=1}^{\hat{\tau}-1} K_{\boldsymbol{H}}\left(\boldsymbol{x}_t - \boldsymbol{x}_i\right)}, \tag{1}$$

where $K_{\boldsymbol{H}}$ is the Normal kernel, the matrix $\boldsymbol{H}$ is the bandwidth, a smoothing parameter, whose choice is discussed in more detail below, while $\hat{\tau}$ and $\boldsymbol{\delta}_{\hat{\tau}:(t-1)}$ represent the estimation for the change point and the shift respectively for which we will provide details next. As the estimated densities $\hat{f}_0$ and $\hat{f}_1$ play the role of predictive distributions, as they are sequentially updated based on the previous data. Using predictive distributions allows the estimates to be sequentially updated without wasting the IC information the test data can provide, rather than using distributions with estimated and fixed parameters

from a calibration (training) phase.

As the goal is to create a memory-based control chart based on $\log L_t$, the cumulative statistic at time $t$ is given by the following expression:

$$S_t = \max\{0, S_{t-1} + \log L_t\}, \tag{2}$$

where $S_0 = 0$. For more details on deriving the recursive formula, see (12).

As both the states and the change point are unknown, we do not provide any test for the first three observations obtained, but we "sacrifice" them to obtain rough representatives of them, initiating the chart. Precisely, we set $\hat{\tau} = 3$, $\boldsymbol{\delta}_{3:3} = \boldsymbol{x}_3$ and $S_{1:3} = 0$, and the control chart can provide online testing starting from the fourth data point, i.e., the next observable. After this, the estimate $\hat{\tau}$ follows from an inherent property of CUSUM. Specifically, the log-ratio provides evidence in favor of the denominator (IC state) when it is negative, while when it is positive supports the nominator (OOC state). At the same time, the zero truncation in 2 fades away the memory in favor of the IC state. The first observation for which the cumulative statistic is greater than 0 denotes the start of the persistent change. Taking this into account, we conclude that the first OOC observation is $\hat{\tau} = \max\{t : S_t = 0\}$. Regarding the shift $\boldsymbol{\delta}_{\hat{\tau}:(t-1)}$, it is the difference between the post-change and the pre-change mean vectors, i.e.:

$$\boldsymbol{\delta}_{\hat{\tau}:(t-1)} = \sum_{j=\hat{\tau}}^{t-1} \boldsymbol{x}_j/(t-\tau) - \sum_{j=1}^{\hat{\tau}-1} \boldsymbol{x}_j/(\tau-1). \tag{3}$$

In this way, we split the data before and after the estimated change point, avoiding the involvement of the contaminated data in the calculation of the IC estimates. The benefit of this formulation is twofold. On the one hand, the control chart is adaptive to successfully detect either small or large shifts, being free from tuning parameters. On the other hand, it is directional invariant as it can detect a location shift in any direction. Turning now to the bandwidth $\boldsymbol{H}$, which is thoroughly discussed in the literature (for example, see (3)), it should be carefully chosen to avoid a degeneracy or oversmoothing for the density estimation. We suggest the use of the unconstrained bandwidth introduced by (11), which minimizes the Asymptotic Mean Integrated Squared Error (AMISE) for the Normal kernel, and it is given by:

$$\hat{\boldsymbol{H}} = \hat{\boldsymbol{\Sigma}}_{1:(\hat{\tau}-1)} \left( \frac{4}{(p+2)\cdot(\hat{\tau}-1)} \right)^{2/(p+4)}, \tag{4}$$

where $\hat{\boldsymbol{\Sigma}}_{1:(\hat{\tau}-1)}$ is the sample covariance matrix of the $p$-dimensional data $\boldsymbol{x}_{1:(\hat{\tau}-1)}$.

It is worth noting that locally adaptive bandwidths have also been proposed and discussed in the literature. Apart from the fixed bandwidths, adaptive choices have also been proposed and discussed in the literature. For further information see (9) and (10). In any case, since our primary goal is detection and not estimation, the choice of $\boldsymbol{H}$ is not crucial as long as it is chosen in a reasonable way.

## 3. Stopping time and performance metrics

KDE-CUSUM is a sequential hypothesis testing procedure where two competing states are compared via the log-ratio of their kernel-estimated multivariate densities within a memory-based control scheme. The goal is to detect a transition from the IC to the OOC state as soon as possible while keeping the number of false alarms at a low predetermined level. In this spirit, we define the stopping time $T$ by:

$$T = \inf\{t : S_t \geq h\}, \tag{5}$$

i.e., it is the first time that the cumulative statistic $S_t$ exceeds a threshold $h$, which is determined by a pre-specified false alarms tolerance.

An important aspect is the evaluation of the proposed method via appropriate metrics. Regarding the IC performance, we suggest using the IC Average Run Length ($ARL_0$) for the decision limit $h$

DRAFT

derivation, as described in (8). $ARL_0$ corresponds to the average number of consecutive IC observations obtained sequentially until the first false alarm is raised. Regarding the OOC detection, an appropriate performance metric should consider the location of $\tau$. Thus, we adopt the framework of (6), estimating the Conditional Expected Delay (CED), which is given by:

$$CED(\tau) = E_\tau(T - \tau + 1 | T \geq \tau) = \frac{E_\tau((T - \tau + 1)^+)}{P(T \geq \tau)}, \tag{6}$$

where $(T - \tau + 1)^+ = \max\{0, T - \tau + 1\}$. $CED(\tau)$ is the expected delay of an alarm after the change point occurs.

## 4. Short Simulation study

In this section we provide a brief simulation study, comparing the performance of the proposed KDE-CUSUM against the distribution free EWMA (DFEWMA, (4)). As IC distribution we consider a two-dimensional ($p = 2$) Normal with mean zero vector and a covariance matrix $\boldsymbol{C}$ with $c_{k,l} = 0.6^{|k-l|}$. Simulating 1,000 IC sequences, we derive the decision limits (for KDE this is $h$) setting $ARL_0 = 100$. Regarding the OOC scenarios, we will examine the performance in detecting several persistent shifts, setting two change point locations $\tau \in \{51, 101\}$. Regarding the DFEWMA, we set the smoothing parameter $\lambda = 0.1$ and the weight parameter $w = 29$ based on the authors' recommendations. The simulated results are given in Table 1. As can be seen, KDE-CUSUM detects a change faster in almost all cases establishing its good diagnostic ability. The only setting where DFEWMA performs better corresponds to the case of shift $(1, 1)$. This is because DFEWMA does not consider correlations between variables, and its statistic is based only on ranks. So in this specific case, with a small positive and simultaneous shift in the variables, it is favored in terms of detection. On the other hand, KDE-CUSUM performs better in all other cases and maintains its good diagnostic ability for changes in all directions.

Table 1: The simulated results

| $\boldsymbol{\mu}_{OOC}$ | $\tau = 51$ | | $\tau = 101$ | |
|---|---|---|---|---|
| | **KDE-CUSUM** $CED(\tau)$ | **DFEWMA** $CED(\tau)$ | **KDE-CUSUM** $CED(\tau)$ | **DFEWMA** $CED(\tau)$ |
| $(1, 1)$ | 10.259 | **6.660** | 8.874 | **6.559** |
| $(2, 2)$ | **3.057** | 4.118 | **2.930** | 4.116 |
| $(1, 0)$ | **7.645** | 11.440 | **6.763** | 10.876 |
| $(2, 0)$ | **2.699** | 5.087 | **2.688** | 4.884 |
| $(0.5, -0.5)$ | **9.687** | 26.052 | **8.219** | 20.913 |
| $(1, -1)$ | **3.107** | 6.173 | **3.041** | 6.012 |

## 5. Discussion and Future Work

In this work, we develop a methodological framework to efficiently detect changes in multidimensional processes, while relaxing the strict assumption of a known distribution. Precisely, we developed a multivariate nonparametric CUSUM focusing on detecting shifts of the mean vector. The adaptability of the proposed methodology allows the detection of systematic changes (small or large) in any direction, as well as being free of tuning parameters, which is important in high dimensions where intuition may not

DRAFT

be available. At the same time, the density estimation via kernels allows us to consider the associations between the variables and, generally, the data structure, which is crucial for detection performance. Furthermore, the method does not require a lengthy historical dataset to estimate the IC state; it can provide effective testing for mean vector shifts after a few dozen IC data while updating the estimates using the current process data.

Despite the promising results obtained with the proposed method, this is still work in progress. In the following steps, we will extend the research on the most appropriate bandwidth choice while considering the performance and addressing the computational cost in higher dimensions. Furthermore, we will extend the detection scheme for the case of changes in the covariance matrix, while demonstrating its performance in practical applications with real data.

## 6. Acknowledgements

This work has been supported by the European Union Horizon 2020 program under Grant Agreement No. 739551 (TEAMING KIOS CoE) and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

DRAFT

## References

[1] Bersimis, S., Psarakis, S., Panaretos, J.: Multivariate statistical process control charts: an overview. Quality and Reliability Engineering International, **23**, 5, 517-543 (2007).

[2] Capizzi, G., Masarotto, G.: Self-starting CUSCORE control charts for individual multivariate observations. Journal of Quality Technology, **42**, 2, 136-151 (2010).

[3] Chacón, J. E., Duong, T.: Multivariate kernel smoothing and its applications. Chapman and Hall/CRC (2018).

[4] Chen, N., Zi, X., Zou, C. : A distribution-free multivariate control chart. Technometrics, **58**, 4, 448-459 (2016).

[5] Hawkins, D. M., Maboudou-Tchao, E. M. (2007): Self-starting multivariate exponentially weighted moving average control charting. Technometrics, **49**, 2, 199-209 (2007).

[6] Kenett, R. S., Pollak, M.: On assessing the performance of sequential procedures for detecting a change. Quality and Reliability Engineering International, **28**, 5, 500-507 (2012).

[7] Li, W., Zhang, C., Tsung, F., Mei, Y.: Nonparametric monitoring of multivariate data via KNN learning. International Journal of Production Research, **59**, 20, 6311-6326 (2021).

[8] Montgomery, D. C.: Introduction to statistical quality control. John Wiley & Sons (2020).

[9] Ruppert, D., Wand, M. P.: Multivariate locally weighted least squares regression. The Annals of Statistics, **22**, 3, 1346-1370 (1994).

[10] Sain, S. R.: Multivariate locally adaptive density estimation. Computational Statistics & Data Analysis, **39**, 2, 165-186 (2002).

[11] Wand, M. P.: Error analysis for general multtvariate kernel estimators. Journal of Nonparametric Statistics, **2**, 1, 1-15 (1992).

[12] West, M., Harrison, J.: Bayesian forecasting and dynamic models. Springer Science & Business Media (2006).

# Anomaly Detection in Circular Data

Houyem Demni[a] and Giovanni C. Porzio[a]

[a]Department of Economics and Law, University of Cassino and Southern Lazio;
houyem.demni@unicas.it, porzio@unicas.it

**Abstract**

Circular data arise as directions, rotations, axes, clock, or calendar measurements. Applications are found in industry, envirometrics, Earth sciences and many other fields. Detecting outliers is an important problem that has been studied in several research areas. In this study, an outlier identification procedure for circular data is suggested. The proposed method is based on robust estimates of distribution parameters on the circle and it is illustrated through two real data examples.

***Keywords:*** angles, directions, Ko estimator, outliers, robust statistics.

DRAFT

## 1. Introduction

A circular observation lies on the circumference of the unit circle and it can be described in polar coordinates by an angle $\phi \in [-\pi, \pi)$ or $[0, 2\pi)$ measured in a specified direction from a specified origin, as well as in Cartesian coordinates through the vector $x = (\cos\phi, \sin\phi)^T$ for which $||x|| = 1$. Circular data arise in many fields such as in Earth sciences (5), biology (20), bioinformatics (16) and also in industry (11). Books covering many aspects of circular data are available within the literature (15; 23).

When dealing with circular data, as with any kind of data analysis, outlying observations or anomalies may influence the main findings and conclusions. They can also reveal unexpected patterns in the data.

Outliers can be defined as observations that are different from the majority. These outlying observations may occur due to copying or recording errors, they could have been recorded under exceptional circumstances, or they simply come from another population.

Detecting these anomalous cases can be thus essential. However, numerous difficulties can arise while performing this task. In practice, as it will be discussed shortly, we found that the available techniques may be not as effective as they should be. Particularly, outliers may not be detected, a notorious effect called masking, or some good observations might be flagged as outliers (which is known as the swamping effect). To avoid these effects, a potentially useful approach is to rely on robust statistical procedures, and this work is aimed at investigating this perspective.

The paper is organized as follows. Section 2. provides a review on outlier detection techniques on the circle, while Section 3. describes the robust anomaly detection technique. Finally, in Section 4, two real data examples are used in order to illustrate the proposed methodology.

## 2. Outlier Detection on the Circle

Within the literature, several tools have been considered to detect outliers in circular data. One option is to detect outliers by deletion. That is, one or more points are deleted, the analysis is performed without

them, the deleted points are then somehow compared with the obtained results. Within this context, many techniques have been made available. For instance, a statistic that identifies an observation as an outlier if it appears as the most influential observation on the mean resultant length has been proposed (14). Four tests of discordancy for outlier detection have been described and compared in (6). These techniques can be only used for small sample sizes and to detect a single outlier. A discussion on outlier detection on the circle has been also provided in textbooks (7; 15) where the proposal of (6; 14) has been considered.

An outlier detection rule based on the locally most powerful invariant statistic and the likelihood ratio test has been introduced in (21) under the assumption that the data follow a von Mises distribution. They compared their proposed method with the ones in (6; 14). However, their method relies on assuming that the concentration parameter is known.

Unfortunately, outlier detection techniques by deletion suffer from the masking effect. That is, an outlier is undetected because of the presence of another adjacent anomalous observation.

More recently, a series of new statistical tests for anomaly detection in circular data, based on a circular distance (3; 12), the sums of these distances (2) and on the spacings theory (17) have been introduced and compared with existing techniques. Nevertheless, each of these techniques has some limits. The procedures in (2; 3) are able to detect only single outliers, while the cutoff value for the one in (12; 17) is obtained through simulations under a specific data model. Additionally, the detection rule in (17) imposes that multiple outliers are well separated from the rest of the data.

Other authors discussed how to identify outliers in multivariate directional settings (i.e., when data lie on a sphere or on a torus) (1; 8; 24). Although these methodologies can be adapted to the circular case, no specific study is available along this direction.

Alternatively, robust statistical techniques can be used. However, this concept have been only considered in (4) or within the context of circular regression (19). In (4), the weighted likelihood and minimum disparity methods are extended to the circular case under the von Mises distribution assumption. Their proposal is rather complex to be applied and it strongly depends on the choice of a bandwidth and of a certain $\alpha$ parameter.

DRAFT

## 3. Robust Anomaly Detection

Anomaly detection is a task strongly related to the idea of robust statistics. Outliers can be detected by fitting the majority of the data and flagging as potential outliers the observations that deviate from it (22).

Robust procedures assume that the majority of the data (that are supposed to be clean) follows a specific probability distribution (9). For instance, for data on a line, the data are assumed to follow the Normal probability density function with unknown mean and standard deviation. Under this assumption, the location and dispersion parameters of the distribution are estimated in a robust way, and a cutoff threshold is identified in order to recognize and discard potential outliers. As cutoff, the quantile of the assumed distribution is typically considered.

For Normal data, thus, an observation $x_i$ will be flagged as outlier if

$$\frac{x_i - \hat{\mu}}{\hat{\sigma}} < \Phi^{-1}(1 - \alpha/2),$$

where $\hat{\mu}$ and $\hat{\sigma}$ are some robust estimates of the corresponding parameters, and $\Phi$ is the standard Normal cumulative distribution function (cdf).

Within the circular domain, we apply this same procedure and we assume data come from the von Mises distribution. The von Mises distribution is the most used distribution to model circular data, and its circular density is given by:

$$h(\phi; \mu, \kappa) := \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\phi - \mu)), \tag{1}$$

with $I_0$ the modified Bessel function of the first kind and of order 0, and where $\mu$ is a location parameter and $\kappa \geq 0$ is the concentration parameter. For $\kappa = 0$, the distribution reduces to the uniform distribution

on the circle. When $\kappa > 0$, the distribution is symmetric around $\mu$, which is both the directional mean and median of the distribution.

In our setting, we also assume $\kappa > 0$. This is because (a) under uniformity on the circle it would be odd to find points that "deviate from the majority of the data", and (b) in such a case the parameter $\mu$ is undefined.

Under this model, robust anomaly detection will be performed by first robustly estimating $\mu$ and $\kappa$. Then, the (shortest arc) distance of each observed angle $\phi_i$ from the estimated $\mu$ will be computed, and then compared with a cutoff value $c_\alpha(\kappa), \alpha > 0$. The value of $\alpha$ will be set by the analyst, keeping in mind that it represents the expected proportion of the points that will be flagged as outliers while actually they are not.

The cutoff value $c_\alpha(\kappa)$ will be the $1 - \alpha/2$ quantile direction of a von Mises distribution centered in $\mu = 0$. It will be thus obtained by solving the equation: $\int_0^{c_\alpha(\kappa)} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\phi)) d\phi = (1 - \alpha/2)$.

Hence, in practice, a circular observation $\phi_i$ will be flagged as potential outlier/anomalous data if

$$d(\phi_i, \hat{\mu}) > c_\alpha(\hat{\kappa}), \tag{2}$$

where $d(\phi_i, \phi_j) := \pi - |\pi - |\phi_i - \phi_j||$ is the length of the shortest arc joining $\phi_i$ with $\phi_j$.

Robust estimators of $\mu$ and $\kappa$ must be adopted in Equation 2 in order to get an effective anomaly detection rule. This will guarantee protection against the masking effect, while the level of swapping will be controlled by the chosen value of $\alpha$.

Within this work, as robust estimators of $\mu$ and $\kappa$, we suggest to use the Fisher circular median and the simple concentration estimator discussed in (10), respectively. The first is defined as the point minimizing the shortest arc distances of the sampled observations from it. That is, let $C = \{\phi_1, .., \phi_i, .., \phi_n\}$ be a circular data set. Its Fisher median is given by:

$$\hat{\mu} := \arg\min_{\eta \in \mathcal{S}} \sum_{i=1}^{n} d(\phi_i, \eta)). \tag{3}$$

The robust estimator of the concentration parameter described in (10) is instead given by

$$\hat{\kappa} = (\Phi^{-1}(0.75)/CMAD(C))^2, \tag{4}$$

where $CMAD(C)$ is the circular median absolute deviation of the set $C$, this latter being the median of the shortest arc distances of the observed values $\phi_i$ from $\hat{\mu}$.

At the end, as a peculiar property of data on the circle, we note that the minimization problem in Equation 3 can result in a disconnected set of values (if the set is connected, the median will be given by its central point). Should this unlikely event occur, a different robust estimator of the location parameter $\mu$ must be adopted (e.g. the circular trimmed mean estimator).

DRAFT

## 4. Illustrative Examples

For illustrative purposes, the anomaly detection procedure proposed in Section 3. is here applied to two real data sets. The first is the well-known Sardinian sea stars while the second is related to an industrial application, and it considers some wind directions.

### 4.1 Sea stars

The Sea stars data was provided by (7) and it is available within the library *circular* in *R*. It refers to the resultant directions in degrees moved by 22 Sardinian sea stars over a period of 11 days after their displacement from their natural habitat. We transform the given angles from degrees to radians, and we consider that 0 radians is the North pole and the rotation is clockwise.

The data are plotted in Figure 1. According to (7), the 13th and 14th observations (2.565 and 5.201 radians) are outliers. In fact, these values emerge as far-out values with respect to the the majority of the

data. The sample circular median of the data is $\hat{\mu} = 0.040$ radians and it is shown by the black arrow in Figure 1 (left and right panels) while the corresponding estimator of the concentration parameter $\hat{\kappa} = 6.942$. Then, we compute the cutoff threshold at a level $\alpha = 0.01$ (which turns out to be equal to 0.936 radians), and the shortest arc distances between each data point and the circular median. By comparing the computed distances with the threshold, observations 13 and 14 are flagged as outliers (Figure 1, right panel, highlighted in red).

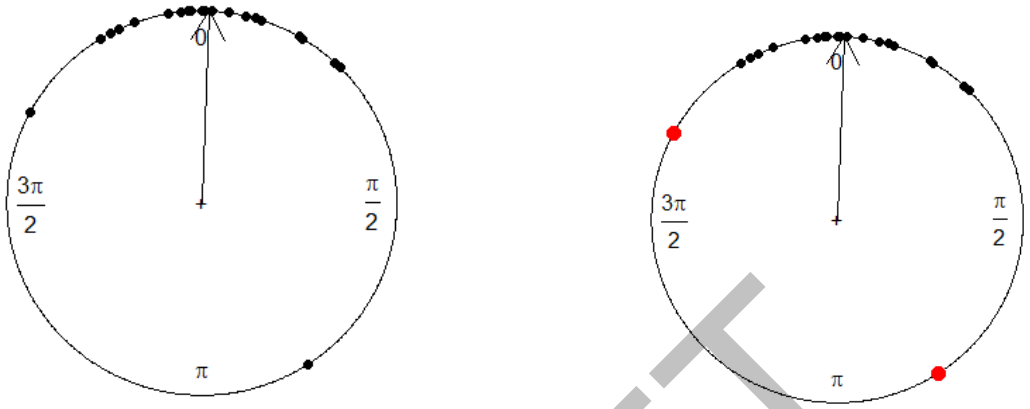



Figure 1: Resultant directions moved by Sardinian sea stars over a period of 11 days after displacement from their natural habitat. Raw data circular plots: the black arrow shows the sample circular median direction (left), outliers are flagged in red (right).

DRAFT

## 4.2 Wind directions

The modeling and the monitoring of wind directions play an important role in the industry of wind power generation (18). The data comes from the recording of wind directions from the meteorological station at "Col de la Roa" in the Italian Alps via data-logger every 15 minutes. Daily recorded wind directions between 3:00 am and 4:00 am inclusive from January 29, 2001 to March 31, 2001 are considered. Accordingly, there are five directions recorded every day leading to a total of 310 measurements (in radians). The data are also available within the *R* package *circular*.

These wind directions have a sample circular median $\hat{\mu} = 0.165$ radians and an estimate of the concentration $\hat{\kappa} = 3.848$. The associated cutoff value at $\alpha = 0.01$ is given by 1.351 radians. The wind directions are depicted in Figure 2 (left panel) and their circular median is drawn by the black arrow. By evaluating the shortest arc distances between each point and the median, and comparing them to the threshold, outliers are flagged (Figure 2, right panel). We found sub-populations of outliers located around the East-Southeast, Southeast, South and West directions.

The same data example was considered in (4), where the presence of sub-populations of outliers located around the East-Southeast, Southeast and South directions was visually inferred by means of a non parametric density estimator.

**Acknowledgments** This work has been partially funded by the BiBiNet project (grant H35F21000430002) within the POR-Lazio FESR 2014-2020.

## References

[1] Abuzaid, A. H.: Identifying density-based local outliers in medical multivariate circular data. Stat. Med. **39(21)**, 2793-2798 (2020).

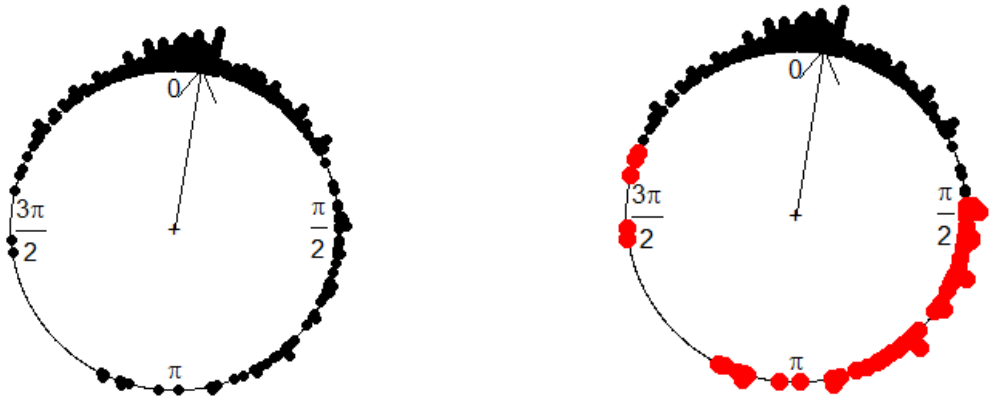



Figure 2: Daily recorded wind directions. Raw data circular plots: the black arrow shows the sample circular median direction (left), outliers are flagged in red (right).

DRAFT

[2] Abuzaid, A. H., Mohamed, I. B., and Hussin, A. G.: A new test of discordancy in circular data. Commun. Stat. Simul. Comput. **38(4)**, 682-691 (2009) doi: 10.1080/03610910802627048.

[3] Abuzaid, A. H., Hussin, A. G., Rambli, A., and Mohamed, I.: Statistics for a new test of discordance in circular data. Commun. Stat. Simul. Comput. **41(10)**, 1882-1890 (2012) doi:10.1080/03610918.2011.624239.

[4] Agostinelli, C.: Robust estimation for circular data. Comput. Stat. Data. Anal. **51(12)**, 5867-5875 (2007).

[5] Cabella, P. and Marinucci, D.: Statistical challenges in the analysis of cosmic microwave background radiation. Ann. Appl. Stat. **3(1)**, 61-95 (2009) doi:10.1214/08-aoas190.

[6] Collett, D.: Outliers in circular data. J. R. Stat. Soc. C-Appl. **29(1)**, 50-57 (1980).

[7] Fisher, N. I.: Statistical analysis of circular data. Cambridge, UK: Cambridge University Press. (1993) doi: 10.1017/cbo9780511564345.

[8] Greco, L., Saraceno, G., and Agostinelli, C.: Robust fitting of a wrapped normal model to multivariate circular data and outlier detection. Stats. **4(2)**, 454-471 (2021).

[9] Hubert, M., and Van der Veeken, S.: Outlier detection for skewed data. J. Chemom. **22(3-4)**, 235-246 (2008) doi: 10.1016/j.simpat.2018.05.010.

[10] Ko, D.: Robust estimation of the concentration parameter of the von Mises-Fisher distribution. Ann. Stat. **20(2)** 917-928 (1992).

[11] Lima–Filho, L. M., Bayer, F. M., and da Silva, A. M.: Control chart to monitor circular data. Qual. Reliab. Eng. **37(3)**, 966-983 (2021).

[12] Mahmood, E. A., Rana, S., Midi, H., and Hussin, A. G.: Detection of outliers in univariate circular data using robust circular distance. J. Mod. Appl. Stat. Methods. **16(2)** 22 (2017) doi:10.22237/jmasm/1509495720.

[13] Mardia, K.V.: Statistics of directional data. Academic Press, London (1972).

[14] Mardia, K. V.: Statistics of directional data. J. R. Stat. Soc. Series B Stat. Methodol. **37(3)**, 349-371 (1975).

[15] Mardia, K. V., and Jupp, P. E.: Directional statistics. Chichester, UK: John Wiley and Sons Ltd. (2000) doi: 10.1002/9780470316979.

[16] Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H.: A multivariate von Mises distribution with applications to bioinformatics. Can. J. Stat. **36(1)**, 99-109 (2008) doi:10.1002/cjs.5550360110

[17] Mohamed, I. B., Rambli, A., Khaliddin, N., and Ibrahim, A. I. N.: A new discordancy test in circular data using spacings theory. Commun. Stat. Simul. Comput. **45(8)**, 2904-2916 (2016).

[18] Koivisto M., Ekström J., Mellin I., Millar J., Lehtonen M.: Statistical wind direction modeling for

the analysis of large scale wind power generation. Wind. Energy **20(4)**, 677-694 (2017).
[19] Rana, S., Mahmood, E. A., Midi, H., and Hussin, A. G.: Robust detection of outliers in both response and explanatory variables of the simple circular regression model. Malays. J. Math. Sci. **10(3)**, 399-414 (2016).
[20] Ranalli, M., and Maruotti, A.: Model-based clustering for noisy longitudinal circular data, with application to animal movement. Environmetrics **31(2)**, e2572 (2020).
[21] Rao, J. S., and Sengupta, A.: Topics in circular statistics. World Scientific Press, Singapore, **10**, 4031 (2001) doi: 10.1142/4031.
[22] Rousseeuw, P. J., and Hubert, M.: Anomaly detection by robust statistics. Wiley. Interdiscip. Rev. Data. Min. Knowl Discov. **8(2)**, e1236 (2018) doi: 10.1002/widm.1236.
[23] Pewsey, A., Neuhäuser, M., and Ruxton, G. D.: Circular statistics in R. Oxford University Press, UK (2013).
[24] Sau, M. F., and Rodriguez, D.: Minimum distance method for directional data and outlier detection. Adv. Data. Anal. Classif. **12**, 587-603 (2018).

DRAFT

# Boosting Diversity in Regression Ensembles

Mathias Bourel[a], Jairo Cugliari[b], Yannig Goude[c], and Jean-Michel Poggi[d]

[a] Facultad de Ingenieria, Universidad de la Republica, Uruguay; mbourel@fing.edu.uy
[b] Université de Lyon 2, France; jairo.cugliari@univ-lyon2.fr
[c] EDF R&D, & LMO, Université Paris-Saclay, France; yannig.goude@edf.fr
[d] Université Paris Cité & LMO, Université Paris-Saclay, France; jean-michel.poggi@universite-paris-saclay.fr

**Abstract**

Ensemble methods, such as Bagging, Boosting or Random Forests, often enhance the prediction performance of single learners on both classification and regression tasks. In the context of regression, we propose a gradient boosting-based algorithm incorporating a diversity term with the aim of constructing different learners that enrich the ensemble while achieving a trade-off of some individual optimality for global enhancement. We present a simple convergence result ensuring that the associated optimization strategy reaches the global optimum. In the experiments, we consider a variety of different base learners with increasing complexity: stumps, CART trees, Purely Random Forests and Breiman's Random Forests. Finally, we consider simulated and benchmark datasets and a real-world electricity demand dataset to show, by means of numerical experiments, the suitability of our procedure by examining the behaviour not only of the final or the aggregated predictor but also of the whole generated sequence.

**Keywords:** Diversity, Boosting, Tree-based methods, Regression

DRAFT

## 1. Introduction

The practical interest of using ensemble methods has been highlighted in several works. Ensemble methods, such as Bagging, Boosting or Random Forests, often enhance the prediction performance of single learners on both classification and regression tasks.

Aggregation estimation as well as sequential prediction provide natural frameworks for studying ensemble methods and for adapting such strategies to time series data. Sequential prediction focuses on how to combine by weighting a given set of individual experts while aggregation is mainly interested in how to generate individual experts to improve prediction performance.

## 2. Boosting diversity

A key quantity for analysing ensemble strategies is to introduce the concept of diversity (see [2]). Considering a linear combination of individual predictors, the squared error of the mixture can be decomposed as the sum of the weighted average error of the predictors minus a positive diversity term.

We propose, in the regression context, a gradient boosting-based algorithm by incorporating in the classical Boosting with the l2 loss (see [3]) a diversity term to guide the gradient boosting iterations. The idea is to trade off some individual optimality for global enhancement. The improvement is obtained with progressively generated predictors by boosting diversity.

We establish a new algorithm, Boosting Diversity (BoDi), which takes diversity into account at each step. The goal is to focus on the intermediate learners and on the ensemble at the same time to enrich prediction. The BoDi algorithm is implemented in the homonymous R package (see [4]).

## 3. A convergence result

Verifying the hypotheses of a theorem from [2] a convergence result is given ensuring that the associated optimisation strategy reaches the global optimum.

## 4. Experimental results

In the experiments, we consider a variety of different base learners with increasing complexity: stumps, CART trees, Purely Random Forests and Breiman's Random Forests.

We consider simulated and benchmark datasets and a real-world electricity demand dataset to show, by means of numerical experiments, the suitability of our procedure by examining the behaviour not only of the final or the aggregated predictor but also of the whole generated sequence.

## 5. Conclusions

In this study, we propose a new boosting algorithm for regression problems based on the diversity formula. This method constructs a base learner at each step improving the diversity term of the diversity formula and then tries to reduce the mean square error.

Our initial experiments on simulated data and tree-based base learners confirm the potential of themethod when the base learner is rich enough to generate diversity (Random Forests and Purely Random Forests). This could be considered as a surprise, even if combining Random Forests and Boosting is a way to improve initial Random Forests.

**Acknowledgments** This work benefited from the support of the PGMO/IRSDI program (https://www.fondation-hadamard.fr/en)

### References

[1] Biau G, Cadre B (2021) Optimization by gradient boosting. In: Advances in Contemporary Statistics and Econometrics, Festschrift in Honour of C. Thomas-Agnan, Springer

[2] Brown G, Wyatt JL, Tino P (2005) Managing diversity in regression ensembles. J Mach Learn Res 6:1621-1650

[3] Buhlmann P, Yu B (2003) Boosting with the l2 loss. JASA 98(462):324-339

[4] Goude Y, Bourel M, Cugliari J, Poggi J.-M. Bodi: Boosting Diversity in Regression Ensembles, 2022. URL `https://CRAN.R-project.org/package=Bodi`. R package version 0.1.0

DRAFT

# How ENBIS has contributed to the UK Universities Research Excellence Framework

Shirley Coleman [a]

[a] NU Solve, School of Mathematics, Statistics and Physics, Newcastle University, UK;
shirley.coleman@newcastle.ac.uk

**Abstract**

Universities should be Centres of Excellence in Research, Learning and Fellowship in Society. As publicly funded institutions it is not surprising that periodically Universities are asked to demonstrate how they are having a positive impact on Society. Research Excellence Framework assessments were carried out in the UK in 2014 and 2021 and on each occasion, academic departments were required to submit a specified number of impact case studies. Impact case studies present an opportunity to showcase the wide range of benefits that statistical thinking can bring. They are coincident with the culture of ENBIS, the European Network of Business and Industrial Statistics, and the benefit of belonging to such a network is evident in all aspects of impact work. The talk will consider the concept of Excellence Frameworks and present an example from the 2021 Research Excellence Framework. The benefits of ENBIS involvement will be demonstrated in all aspects of the impact case study ensuring a successful outcome.

**Keywords:** societal benefit, continuous improvement, open data, data science, impact case studies, Industry 4.0, monetising data

DRAFT

## 1. Introduction

Universities in the UK are publicly funded institutions and it is not surprising that periodically they are asked to demonstrate their worth. Assessment is carried out with regards to teaching and learning, research and knowledge exchange. In the most recent Research Excellence Framework (REF) assessments in 2014 and 2021, academic departments were required to submit a specified number of impact case studies from each department.

Impact case studies are extremely well suited to applied statisticians and are highly relevant to the culture of ENBIS, the European Network of Business and Industrial Statistics. They represent an opportunity to showcase the wide range of benefits that statistical thinking can bring. One aspect of impact is in the monetising of data. This focuses on making the absolute best of the effort that goes into collecting data in all parts of a process or operation. With Industry 4.0 thinking prevalent in most companies and the wide availability of sensors plus the expectation of a digital profile the world is awash with data.[7] Making sense of this data and using it to understand and continuously improve the quality and efficiency of output makes sound sense both from a financial and a sustainable environmental point of view. [1].

The next section gives an overview of Excellence Frameworks and the REF in particular. In section 3 an impact case study from 2021 is described highlighting the value to the author of being part of ENBIS. Finally, section 4 gives plans for future contributions and concluding remarks.

## 2. Excellence Frameworks

Excellence Framework assessments are carried out with regards to teaching and learning, research and knowledge exchange. These all aim to help Universities improve but can understandably cause anxiety when staff feel they are being judged. They fear the implications could range from negative press and league table scores to censure and reduction in public funding, even though this is not (usually) the case. In theory, the concept is to help the Universities in their continuous improvement.

### 2.1 Knowledge Exchange Framework

The Knowledge Exchange Framework (KEF) shows the underlying supportive rationale. Universities are required to contribute data on a number of dimensions over a number of years. The aggregated scores are then compared with "like" Universities to give insight and to encourage a positive attitude to change. The KEF website https://kef.ac.uk/dashboard gives output from the KEF which is open access and appealing for users to explore.

### 2.2 Research Excellence Framework

The Research Excellence Framework (REF) is interesting because as well as an appraisal of research output in terms of academic papers in highly regarded journals, there is also a requirement to present impact case studies.

Impact case studies are based on the following outputs:

1. A narrative about research on a theme which can extend back up to 20 years
2. Supporting academic publications in relevant journals with at least a 2* ranking
3. Evidence demonstrating impact on society

Academic departments are allocated to an appropriate Unit of Assessment (UoA) depending on their specialty. There must be one impact case study for every 7-10 academic staff members. Before each REF there is considerable uncertainty as to what is expected and what will be judged well. After each REF the case studies are all made available on a Government website and can be freely scrutinised for future ideas. Here we consider UoA 10 which is for Mathematics and Statistics.

#### 2.2.1 Research theme

The long time scale for the research theme is helpful to show how the work progressed and to allow older academic papers to be included. The evidence for the impact, however, has to be in the last 5 years

#### 2.2.2 Academic publications

There must be at least 3 or 4 respectable academic publications on the theme in journals associated with the UoA. At least one author must have been a member of the department at the time of the publication. It can be difficult to publish in high ranking subject specific journals when the research is applied and often straddles several areas of interest.

#### 2.2.3 Societal impacts

Impacts can be in terms of creating new jobs or increasing sales, exports, profits or investments, developing intellectual property or implementing new products. In all cases the impacts must be accompanied by evidence and be corroborated by business or partners external to the academic department.

### 2.3 Continuation from 2014 to 2021 Research Excellence Framework

The dates of each REF are not pre-set and guidelines can change between REFs. However the concept was similar from 2014 to 2021 and it was permitted to present a continuation of a previous impact case study provided there was substantial progression.

DRAFT

## 3. Example of Impact Case Study from 2021 REF

The impact case study presented in this section is of particular interest as it drew significantly on the ENBIS network which facilitated the opportunities that were fundamental to the work.

### 3.1 Research theme

An impact case study entitled "Faster Fault Tracking for National Grid Gas", [2] was successfully reported in REF 2014. Research in statistical applications in gas and other utilities continued and formed part of the subsequent REF 2021 impact case study entitled "Statistics-based Data Analytics for Industry - A Focus on Small and Medium Enterprises"[5].

The focus on SMEs was very much influenced by extensive European funding achieved in conjunction with pro-ENBIS, the EC Framework 5 Thematic Network funding obtained by ENBIS President Dave Stewardson for the newly formed ENBIS.

### 3.2 Academic papers

The academic publications supporting the impact case study included early work with Stewardson who was a founder member of ENBIS and principal investigator of pro-ENBIS [9]. The concept of using open data integrated with company internal data was initiated by involvement with ENBES, the European Network for Business Establishment Statistics and ICOTS, the International Conference on Teaching Statistics which were ENBIS partners. [3, 4]. A pivotal paper was written by a team of ENBIS members.[6]. Publications on data science were written in collaboration with Italian ENBIS members [10] and as a result of involvement in a European Conference of Mathematics in Industry with a Hungarian ENBIS member [4].

The papers provided an opportunity to report on the results of Knowledge Transfer Partnerships (KTPs) funded by Innovate UK. Seven two-year KTPs were won by the author in open competition between 2014 and 2021 due in part to the impressive benefits of being an active ENBIS member with access to experts from all parts of Europe and the opportunity to participate in external examining, research workshops, conferences and publishing. [8].

### 3.3 Societal impact

Societal impact was shown by a combination of testimonials and evidence from reports. For example, the impact case study references: "The testimonial from the Technical Director of Advanced Engineering Solutions (AES). *Provides evidence of the financial impact and global reach of the research.*"

The KTPs were particularly valuable in providing evidence of the benefits of the partnership corroborated by the business partner. This evidence forms part of the obligatory completion of the final report pending payment of the final installment of the Innovate UK funding to the business partner. Impacts from one KTP were included in the impact case study as follows:
The research work has led to the following benefits since 2016:

- an increase in annual sales turnover of £4M;
- an increase in annual exports of £4M;
- an increase in annual profit of approximately £1M;
- the creation of 5 new jobs within the business;
- an investment of greater than £1M in new software and hardware implementation and support.

These are clearly valuable impacts and contributed to the impact case study being suitably highly rated in the REF.

## 4. Concluding Remarks

Excellence frameworks fit well with the ENBIS philosphy. In particular as regards the Research Excellence Framework but also whenever an international flavour showing reach is helpful in obtaining research funding. The ENBIS website states that it is "a platform connecting individuals and organisations, interested in theoretical developments and practical applications in the field of business and industrial

DRAFT

statistics".The interaction that such a platform facilitates is clearly invaluable for opening minds to research ideas and statistical possibilities. International collaboration leads to exchange of work and study and boosts the global recognition of members' Universities, companies and Institutions, as well as staff esteem.

It is interesting that most impact case studies in the Mathematics and Statistics Unit of Assessment arise from statistical applications rather than pure or applied mathematics. This is probably due to the explosion of interest in data science that is driving businesses to seek out academic partnerships to explore new capabilities. It also reinforces the importance of the work of ENBIS and shows the value of a network where statisticians can share ideas and develop their research.

**Acknowledgments** The author is indebted to the camaraderie of the ENBIS community and the valuable skills, knowledge and experience of the members met and worked with over the years.

## References

[1] Ahlemeyer-Stubbe, A., Coleman, S.Y.: Monetising data – how to uplift your business. Wiley, Chichester. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119125167 (2018)

[2] Coleman, S.Y.: Faster Fault Tracking for National Grid Gas. http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=21204 (2014) Accessed 23 March 2023

[3] Coleman, S.Y.: Data mining opportunities for small to medium enterprises from official statistics. *Journal of Official Statistics*, 32, 849-866. doi.org/10.1515/jos-2016-0044 (2016)

[4] Coleman, S.Y.: Data science in Industry 4.0. *Mathematics in Industry* 30, 559-566. doi.org/10.1007/978-3-030-27550-1_71 (2019)

[5] Coleman, S.Y.: Statistics-based Data Analytics for Industry - A Focus on Small and Medium Enterprises. https://results2021.ref.ac.uk/impact/208ade51-7f10-4379-b7a4-c478a70648a4?page=1 (2021) Accessed 23 March 2023

[6] Coleman, S.Y., Gob, R., Manco, G., Pievatolo, A., Tort-Martorell, X., Reis, M.: How can SMEs benefit from big data? Challenges and a path forward. *Journal of Quality and Reliability Engineering International* 32, 2151–2164. doi.org/10.1002/qre.2008 (2016)

[7] Kenett, R.S., Coleman, S.Y.: Data and the Fourth Industrial Revolution. RSS Significance. https://doi.org/10.1111/1740-9713.01523 (2021) Accessed 23 March 2023

[8] Kenett, R.S., Coleman, S., Zempléni, A., De Frenne, A.: European Network for Business and Industrial Statistics (ENBIS): a journey. https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat08435 (2023) Accessed 23 March 2023

[9] Stewardson, D.J., Coleman, S.Y.: *Journal of Applied Statistics* 28, 469-484. doi.org/10.1080/02664760120034180 (2001)

[10] Vicario, G., Coleman, S.Y.: A review of data science in business and industry and a future view. *Applied Stochastic Models in Business and Industry* 36, 6-18. doi.org/10.1002/asmb.2488 (2020)

DRAFT

**Climate change impacts on fertility in low- and middle-income countries: An analysis based on global sub-national data**

Côme Cheritel[1], Roman Hoffmann[2], Raya Muttarak[3]

[1] Paris School of Economics
[2] Wittgenstein Centre for Demography and Global Human Capital (IIASA, OeAW, University of Vienna)/ International Institute for Applied Systems Analysis
[3] Department of Statistical Sciences, University of Bologna

**Abstract**

There is scarce empirical evidence on the direction and the extent to which climate change affects fertility. This paper aims to close the gap in the literature by compiling a comprehensive database of fertility data derived from the Demographic and Health Surveys (DHS) for over 60 low- and middle-income countries at a regional level covering the period from 1980 to 2020. We explore: 1) whether and in what direction exposure to extreme climatic events measured by temperature and rainfall anomalies affect fertility patterns; and 2) demographic heterogeneity in the effect of climatic shocks on fertility. The high resolution climate data (e.g. temperature and rainfall) are matched with the fertility data at a subnational region, with a temporal resolution of up to one month. Our analysis focuses on the effects of global warming (measured as temperature and rainfall anomalies) on age-specific fertility rates. We will ultimately use this information to project the future impact of global warming on fertility, in order to evaluate its potential implications for future population dynamics in a next step of this project.

DRAFT

# Environmental Exposures and Under-5 Mortality in India: A Survival Analysis of DHS Data

**Joseph Vinod and Joseph Kannankeril**

## INTRODUCTION

Climate change is expected to increase the frequency and intensity of environmental extremes, leading to significant public health concerns in the coming years. Most low- and middle-income countries are located in relatively environmentally challenged regions compared to developed countries. However, child growth faltering and mortality in developing countries are invariably attributed to poor nutrition, infection and poverty. Despite the scale-up of nutrition, water, and sanitation interventions and widespread reductions in diarrhoea and all-cause mortality over the past decades, child growth faltering remains a major source of morbidity and mortality in Sub-Saharan Africa and Asia. Globally, nutrition-related factors contribute to approximately 45% of under-five deaths, with Sub-Saharan Africa and Central and Southern Asia accounting for over 80% of under-five deaths, despite only accounting for 52% of the global under-five population. Furthermore, five countries, including Nigeria, the Democratic Republic of the Congo, Ethiopia, India, and Pakistan, account for half of all under-five deaths in 2019.

There has been increasing interest in assessing the impact of rising environmental risk factors on children's health and survival as climate change continues. Despite the demonstrated implications of environmental exposures, early life growth's possible effects have been rarely studied. . In the Indian context, previous studies have largely focused on child, maternal, and household-level risk factors of infant and child survival while ignoring the potential effect of environment and geographic space. Additionally, little has been done to investigate how under-five mortality is associated with socio-demographic and environmental aspects that affect child survival.

Identifying prevalent environmental hazards in different settings where children live is crucial for developing setting-based interventions. Understanding the threats that environmental hazards pose to human health can help mitigate risks and increase preparedness. Thus, this study aims to investigate the associations between environmental exposures and under-five mortality in India. By filling this knowledge gap, this study will help elucidate the effects of environmental exposures on child survival in the country. This study aims to explore the linkages between environmental exposures and under-five mortality in India.

DRAFT

## METHODOLOGY

### Data source

The data for this study was obtained from the National Family Health Survey (NFHS)-4, which is the Indian version of the Demographic and Health Surveys, conducted in a representative sample of households across India. The study also utilized data on air pollution, obtained from The Air Quality Life Index (AQLI) initiative by the Energy Policy Institute at the University of Chicago (EPIC), which uses satellite-derived PM2.5 raw data from the Atmospheric Composition Analysis Group at Dalhousie University at a high resolution of 10km x 10km. In addition, data on drought-prone areas in the year 2015 was obtained from the Ministry of Agriculture and Farmers Welfare, Government of India, and the Mahalanobis National Crop Forecast Centre (MNCFC) in New Delhi.

### Statistical Analysis

To investigate the association between environmental and socio-demographic factors and under-five mortality in India, we employed Kaplan-Meier survival plots, log-rank test (Mantel test), and Cox proportional hazard model. The synthetic cohort component approach was used to calculate death rates. The Kaplan-Meier estimator is a non-parametric statistic that estimates the survival function from lifetime data. It is the simplest way of computing survival over time despite all these difficulties associated with subjects or situations. For each time interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk. Kaplan-Meier survival plots are useful for graphical comparison of potential differences. The log-rank test is used to test the null hypothesis that the survival distribution is the same in two or more groups. The test compares the number of observed events to the number of expected events (based on Kaplan-Meier curves) at all points where the events occur. The numerical value of the test statistic is compared to the chi-square distribution with degrees of freedom that are equal to the number of groups being compared minus one. To estimate the hazard ratios (HRs) with their 95% confidence intervals (CIs) for under-five mortality, we used Cox proportional hazard regression. Adjusted hazard ratios for under-five death were calculated for the total sample, adjusting for each independent variable.

## RESULTS

The results of the log-rank test, mortality rates, and hazard ratios indicate the influence of environmental factors on child survival (Table 1). The under-five mortality rate was 59.09 deaths per thousand live births in air polluted areas, compared to 38.27 deaths per thousand

DRAFT

live births in non-air polluted areas, with a gap of 20.82 deaths per thousand live births. In drought-prone areas, the under-five mortality rate was 53.87 deaths per thousand live births compared to 46.63 deaths per thousand live births in non-drought prone areas, with a gap of 7.2 deaths per thousand live births. The national under-five mortality rate was 49.75 deaths per thousand live births.

The log-rank test statistics for air pollution and drought were 344.27 and 83.54, respectively, with p-values less than .001 for all environmental factors, indicating a significant difference in survival experience between groups. The hazard ratios showed that children living in air polluted areas had a 1.09 times higher risk of under-five death compared to those in non-air polluted areas, while children living in drought-prone areas had a 1.055 times higher risk of under-five death compared to those in non-drought prone areas. Figure 1 presents the Kaplan-Meier survival plots comparing child survival across the different environmental exposures.

**CONCLUSIONS**

The results of this study demonstrate the significant influence of environmental and socio-demographic factors on under-five mortality and child survival in India. It is crucial to recognize these threats and work collaboratively to mitigate the risks and be prepared. To improve people's health, a multi-sectoral, holistic approach is necessary, integrating efforts inside and outside healthcare systems. The high levels of air pollution in India require effective policies and regulation for environmental resource management. Cross-border strategies are necessary to address the issue of pollution.

A quarter of the global disease burden can be prevented by reducing environmental and social risk factors. Understanding these threats can help us work together to mitigate the risks and be prepared, particularly in the context of anthropogenic climate change. Creating healthier environments will not only benefit health and the quality of life but also reduce the burden on the healthcare system. A sustainable approach in this regard may help countries achieve the goals set out in the 2030 Agenda for Sustainable Development. This study highlights the urgent need to improve our understanding of the relationship between environmental risk factors and human health, especially at this crucial time.

DRAFT

***Table-1:*** *Association of under-five deaths with environmental predicators and other background characteristics India, 2015*-16

| Background characteristics | Total (N) | Weighted Percentage | Percentage of under five deaths | Under-five mortality rate *5q0* (95% CI) | Log-Rank test statistic | P-value | Adjusted HR (95% CI) | P-value |
|---|---|---|---|---|---|---|---|---|
| **Air pollution** | | | | | | | | |
| Not polluted | 112368 | 44.95 | 3.41 | 38.27 [36.58-39.96] | 344.27 | <0.001 | 1.00 | |
| Polluted | 137599 | 55.05 | 5.22 | 59.09 [57.43-60.75] | | | 1.095 [1.0414 1.1523] | <0.001 |
| **Drought** | | | | | | | | |
| Not drought prone | 142305 | 24.87 | 4.18 | 46.63 [45.16-48.09] | 83.54 | <0.001 | 1.00 | |
| Drought prone | 107662 | 75.13 | 4.72 | 53.87 [52.26-55.49] | | | 1.055 [1.0084 1.1029] | 0.020 |
| **Sex of child** | | | | | | | | |
| Male | 130572 | 52.24 | 4.66 | 51.47 [49.78-53.16] | 34.35 | <0.001 | 1.121 [1.0813 1.1623] | <0.001 |
| Female | 119395 | 47.76 | 4.14 | 47.85 [46.47-49.23] | | | 1.00 | |
| **Residence** | | | | | | | | |
| Urban | 70118 | 28.05 | 3.08 | 34.40 [32.16-36.65] | 210.86 | <0.001 | 0.970 [0.9203 1.0229] | 0.263 |
| Rural | 179849 | 71.95 | 4.93 | 55.76 [54.27-57.25] | | | 1.00 | |
| **Mothers age** | | | | | | | | |
| 15-24 | 87491 | 35.00 | 4.63 | 52.59 [50.49-54.69] | 187.14 | <0.001 | 1.073 [0.9990 1.1516] | 0.053 |
| 25-34 | 141171 | 56.48 | 4.04 | 44.92 [43.46-46.37] | | | 0.871 [0.8194 0.9253] | <0.001 |
| 35-49 | 21305 | 8.52 | 5.99 | 63.46 [59.63-67.29] | | | 1.00 | |
| **Mothers education** | | | | | | | | |
| No education | 75140 | 30.06 | 5.99 | 67.44 [65.29-69.58] | 734.74 | <0.001 | 1.218 [1.1600 1.2783] | <0.001 |
| Primary | 35120 | 14.05 | 5.36 | 59.56 [56.29-62.83] | | | 1.204 [1.1399 1.2713] | <0.001 |
| Secondary or more | 139707 | 55.89 | 3.32 | 36.47 [35.09-37.84] | | | 1.00 | |
| **Birth order** | | | | | | | | |
| One | 96475 | 38.59 | 4.37 | 47.82 [45.81-49.83] | 446.81 | <0.001 | 0.872 [0.8202 0.9274] | <0.001 |
| Two-Three | 117983 | 47.20 | 3.77 | 42.87 [41.64-44.11] | | | 0.744 [0.7050 0.7845] | <0.001 |
| Four and above | 35510 | 14.21 | 6.63 | 75.62 [72.75-78.50] | | | 1.00 | |
| **Religion** | | | | | | | | |
| Hindu | 196629 | 78.66 | 4.49 | 50.52 [49.43-51.62] | 103.92 | <0.001 | 1.178 [1.0496 1.3213] | 0.005 |
| Muslim | 41379 | 16.55 | 4.39 | 49.94 [47.42-52.46] | | | 1.218 [1.0765 1.3781] | 0.002 |
| Christian | 5111 | 2.04 | 2.77 | 32.21 [26.29-38.13] | | | 1.085 [0.9439 1.2464] | 0.252 |
| Others | 6848 | 2.74 | 3.37 | 38.97 [33.78-44.16] | | | 1.00 | |
| **Caste/Tribe** | | | | | | | | |
| Scheduled caste | 53851 | 21.54 | 4.91 | 55.86 [53.72-57.99] | 130.55 | <0.001 | 1.120 [1.0516 1.1926] | <0.001 |
| Scheduled tribe | 26350 | 10.54 | 4.93 | 57.24 [53.71-60.77] | | | 1.056 [0.9848 1.1314] | 0.127 |
| Other backward class | 110399 | 44.17 | 4.54 | 50.78 [49.22-52.34] | | | 1.043 [0.9878 1.1013] | 0.129 |
| Others | 59366 | 23.75 | 3.48 | 38.97 [36.77-41.16] | | | 1.00 | |
| **Wealth quintile** | | | | | | | | |
| Poorest | 63394 | 25.36 | 6.27 | 71.70 [69.45-73.96] | 958.53 | <0.001 | 1.944 [1.7732 2.1310] | <0.001 |
| Poorer | 54939 | 21.98 | 5.07 | 57.33 [54.82-59.84] | | | 1.786 [1.6368 1.9489] | <0.001 |
| Middle | 49577 | 19.83 | 4.20 | 46.16 [42.99-49.32] | | | 1.603 [1.4717 1.7460] | <0.001 |
| Richer | 45305 | 18.12 | 3.11 | 34.90 [32.35-37.45] | | | 1.381 [1.2663 1.5057] | <0.001 |
| Richest | 36752 | 14.70 | 2.08 | 22.57 [20.89-24.26] | | | 1.00 | |
| **Region** | | | | | | | | |
| North | 32928 | 13.17 | 3.97 | 44.43 [42.43-46.42] | 903.55 | <0.001 | 1.354 [1.2339 1.4856] | <0.001 |
| Central | 67799 | 27.12 | 6.41 | 73.50 [71.66-75.35] | | | 1.857 [1.6864 2.0444] | <0.001 |
| East | 63638 | 25.46 | 4.46 | 49.55 [47.44-51.67] | | | 1.263 [1.1446 1.3942] | <0.001 |
| Northeast | 8839 | 3.54 | 4.45 | 50.09 [46.52-53.66] | | | 1.258 [1.1292 1.4014] | <0.001 |
| West | 31836 | 12.74 | 2.93 | 33.58 [29.95-37.21] | | | 1.107 [0.9889 1.2388] | 0.077 |
| South | 44927 | 17.97 | 2.69 | 29.42 [26.86-31.97] | | | 1.00 | |
| **India** | **249967** | **100.00** | **4.41** | **49.74 [48.60-50.88]** | | | - | - |

DRAFT

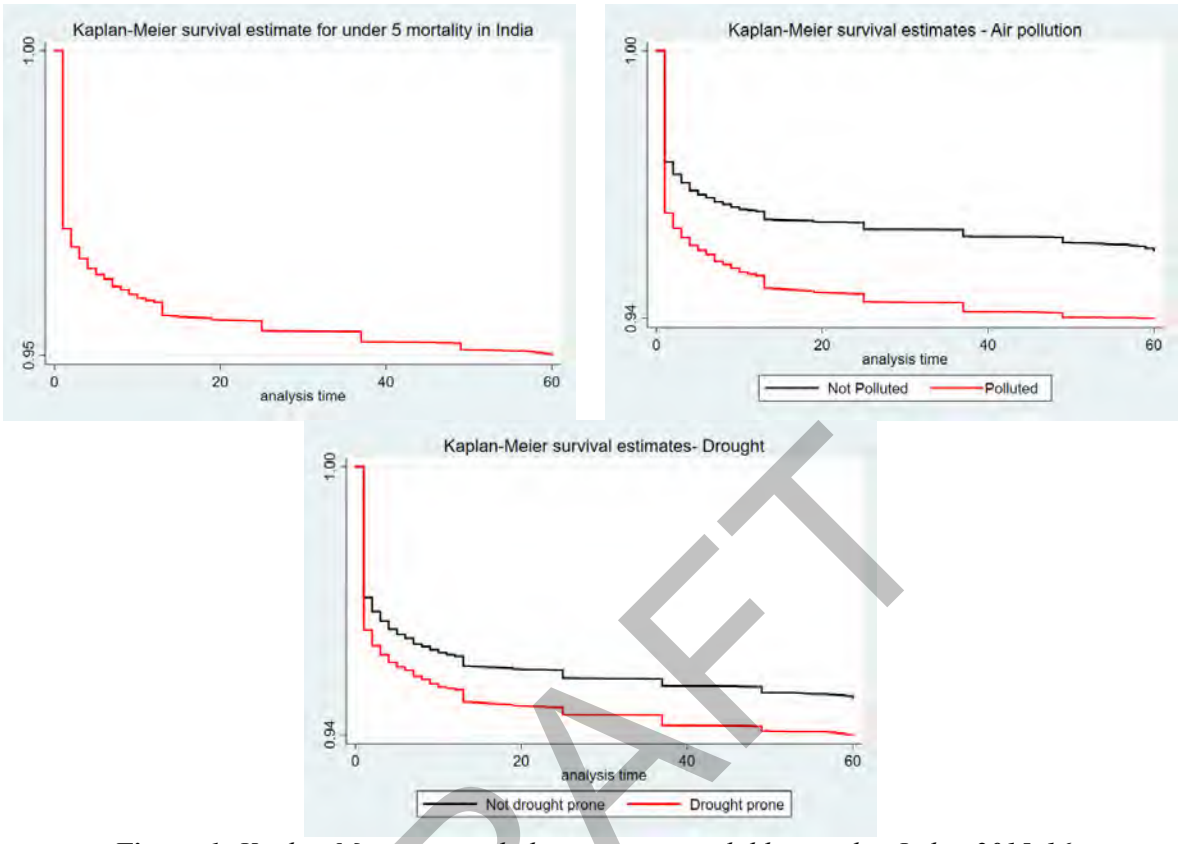



***Figure-1:*** *Kaplan-Meier survival plots comparing child survial in India, 2015-16*

DRAFT

**The impact of temperature on expressed sentiment by migration status: evidence from geo-located Twitter data"**

**Risto Conte Keivabu**

Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, 18057 Rostock, Germany

**Jisu Kim**

Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, 18057 Rostock, Germany

DRAFT

## Abstract

The escalating prevalence of extreme temperatures due to climate change is expected to impact various aspects of human life, with mental health being a critical component that has implications for overall well-being. In this study, we aim to contribute to the growing body of research examining the connection between temperature and expressed sentiment. Our analysis utilizes geo-referenced text data from Twitter users, combined with comprehensive meteorological information, to assess the effects of temperature on individuals. Our sample comprises 1,839 Twitter users based in the United States, including 979 natives and 860 migrants, who tweeted between 2012 and 2019. We combine the users' locations with finely-tuned meteorological data provided by Gridmet. Our findings reveal that both cold and hot days contribute to a decline in positive sentiment and an increase in negative sentiment among users. Interestingly, we do not detect any significant differences in the impact of heat or cold on native users compared to migrants. However, we do observe variations in impact of heat and cold within the migrant group, depending on the length of their stay in the United States. In conclusion, leveraging individual-level geo-referenced textual data offers valuable insights into the effects of extreme temperatures on people's mental states and the heterogenous impact in the population. Furthermore, it sheds light on how the anticipated rise in future temperatures could influence multiple dimensions of human well-being.

## Introduction

In recent years, the impact of climate change and heat exposure on human behavior and wellbeing has become even more critical (Adger et al., 2022; Vargas et al., 2023). Recent studies documented temperature to affect several domains of human activity such as productivity (Cai et al., 2018), cognitive abilities (Chang & Kajackaite, 2019), test scores (Park et al., 2020), mental health (Mullins & White, 2019), time use (Graff Zivin & Neidell, 2014), reproductive behavior (Hajdu & Hajdu, 2019; Wilde et al., 2017), decision making (Almås et al., 2019; Heyes & Saberian, 2019) and leisure activities (Fan et al., 2023). Moreover, increasingly researchers have inquired the impact of climate on human activities leveraging new sources of data.

Recently, studies have combined novel digital data with meteorological information to provide additional evidence on the effect of temperature on human online behavior and expressed sentiment. For example, studies investigated how climate affects expressed sentiment of Twitter and Facebook users in the United States (Baylis et al., 2018) and Weibo users in China (Wang et al., 2020) finding a decrease in positive sentiment and an increase in negative sentiment with exposure to extreme heat and cold. Additionally, an increase in hate speech with exposure to hot and cold days has been documented in a large sample of tweets in the United States (Stechemesser et al., 2022). Nevertheless, a gap in these studies is represented by the use of aggregate level data and a lack of insights on the heterogenous impacts of climatic variables on the online population residing within the same location.

In this article, we contribute to the growing literature interested in the impact of extreme temperature on online behavior exploring how migration status stratifies the impact of temperature on Twitter users. For this purpose, we leverage a unique dataset of geolocated tweets of individual users located in the United States that provides information on migration background (Kim et al., 2023). We combine this dataset with meteorological variables at the date in which the users have tweeted to estimate the impact of extreme temperatures on their sentiment. The analysis of a heterogeneous impact of extreme temperature by migration background could offer insights on adaptation to the local weather. For example, studies on the association between temperature and mortality found variation in adaptation to heat and cold based at different latitudes and climatic zones (de Freitas & Grigorieva, 2015; Medina-Ramón & Schwartz, 2007). Additionally, studies documented differences in human thermal perception that depend on physiological and psychological adaptations (Schweiker et al., 2018). On one hand, migrants could be less adapted then natives to the new climatic context to which they relocate and weather could be considered as an additional cost to their migration experience. On the other hand, having less experience with the local meteorological phenomena might refrain them from perceiving temperatures out of the local averages as unusual and feel concerned about it.

DRAFT

Followingly, we describe the data and methods used in the analysis, we present preliminary results and conclude with a discussion of the findings.

## Data and methods

*Data*

In the analysis, we rely on two main datasets. First, we collect data on tweets from a sample of Twitter users using the Twitter API. The sample is comprised of 979 native and 860 migrant users for which we have tweets collected between 2012 and 2019 comprising a total of 505,157 tweets. The migrant sample is defined as those users that: "have tweeted at least one geo-tagged tweet per month in one country (home country) for 12 consecutive months and one geo-tagged tweet per month for 12 consecutive months in another (destination) country". Moreover, the sample of native and migrants have been selected matching them on variables such as: number of followers, friends, tweets, account age, age, and gender. Also, we use the modal location of the tweets of each individual to allocate the city of residence to the users.

We connect the individual user location with daily meteorological information provided by GRIDmet. GRIDmet is a highly reliable source of meteorological data for the contiguous United States with a resolution of 4km and available from 1979 to 2023 (Abatzoglou, 2013). For our purpose, we collected the daily meteorological values at the grid cell at the center of the city in which the Twitter user is residing. The meteorological values we collected are maximum temperature, wind speed, shortwave downwelling radiation, humidity and precipitation.

*Variables*

Our outcome variables are measures of positive and negative sentiment expressed in the tweets. We used the RoBERTa algorithm on the textual data from the tweets to compute measures of positive and negative in the tweets (Liu et al., 2019). However, we disregarded tweets that contain only urls or emojis and translate non-English text to English. Alternative methods to compute sentiment exist but for the purpose of our textual data, RoBERTa has been shown to provide more robust estimates (Liu et al., 2019).

Our main exposure variable is maximum temperature. We measure maximum temperature on the day of the tweet using binary variables for the daily temperature range with intervals of 3°C ranging from < -6°C to > 33°C. The binary variables are preferred to a continuous measure of temperature to capture a non-linear association between temperature and sentiment as shown in previous studies (Wang et al., 2020).

Also, we collect additional meteorological information provided by Gridmet as precipitation, wind speed, shortwave downwelling radiation and humidity.

DRAFT

*Method*

$$1) \quad Y_{iutc} = TEMP_{tc}^{j} + X_{tc} + \delta_{mc} + \upsilon_{d} + \mu_{i} + \varepsilon_{iutc}$$

In the equation, $Y_{iut}$ denotes the outcomes, positive and negative sentiment score, measured for tweet *i* of user *u* at date *t* in city *c.* The temperature exposure is captured by TEMP composed by the categories j described above and measured at date *t* and city *c.* The temperature range 3 to 24°C is considered as the comfort zone and excluded from the analysis. We capture the causal effect of temperature on sentiment using a set of widely used fixed effects(Heyes & Saberian, 2019). Most importantly, we use user level fixed effects $\mu_{i.}$ that allows us to capture the within user variation determined by a day in a specific temperature bin on sentiment relative to a day in the comfort zone. Also, we include month-by-city fixed effects $\delta_{mc}$ to account for city specific seasonal differences and day of week fixed effects $\upsilon_{d}$ to rule out that our results are biased by possible differences in users sentiment within the week such as the Monday effect (K. Kim & Ryu, 2022). Also, we added a vector X of time-varying meteorological and individual level control variables captured at date t in city *c*. The meteorological control variables that we include are precipitation, wind speed, humidity and solar radiation for which we use continuous values. Standard errors are clustered based on the month-by-city unit of analysis to account for spatial and temporal autocorrelation[1].

Also, we explore a stratified impact of temperature in our population. First, we inquire heterogeneity in the impact of temperature between our sample of natives and migrants adding a running separate analysis based on our dummy identifier (0=Natives, 1=Migrants). Secondly, we stratify the analysis within the migrant population based on their length of stay in the United States. More precisely, we use a binary to divide migrants in two groups i.e.: those that resided more than 24 months in the US and those that resided less than 24 months[2].

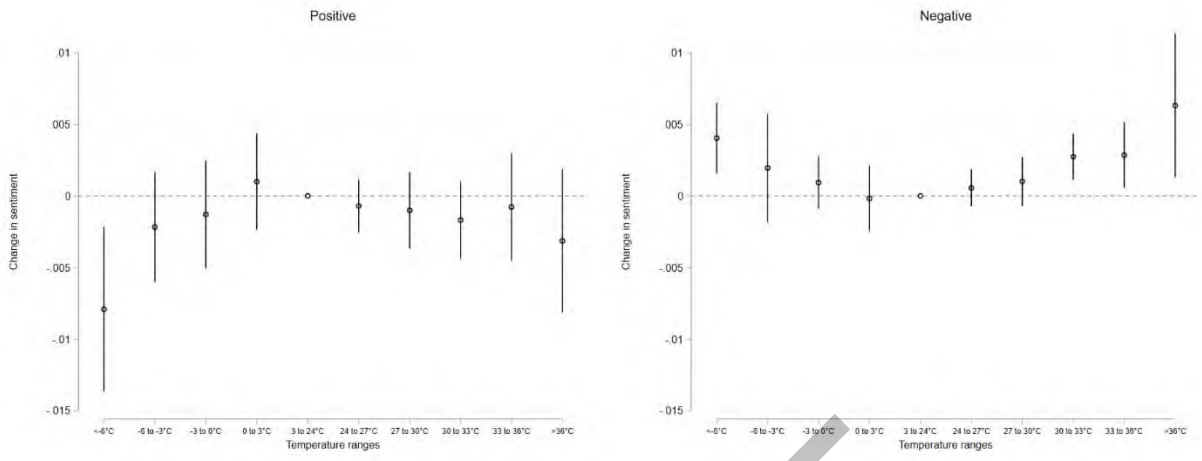
DRAFT

**Preliminary results**

In Figure 1, we present results on the impact of maximum temperature on positive and negative sentiment. We observe the relationship to be an inverted U-curve for positive sentiment and a U curve for negative sentiment. Respectively, extreme cold temperatures show to reduce positive sentiment with an increased effect size recorded at the coldest temperature bin (<-6°C). For hot temperatures(>36°C) we find a decrease in positive sentiment but the effects are not statistically significant at the 95% level. For negative sentiment, we observe a similar increase with hot and cold temperatures and a similar effect size at both extremes.

[1] Our estimates are robust to alternative clustering of standard errors.
[2] The choice of such threshold is based on the mean months of permanence in the sample of migrants that is 24 months.

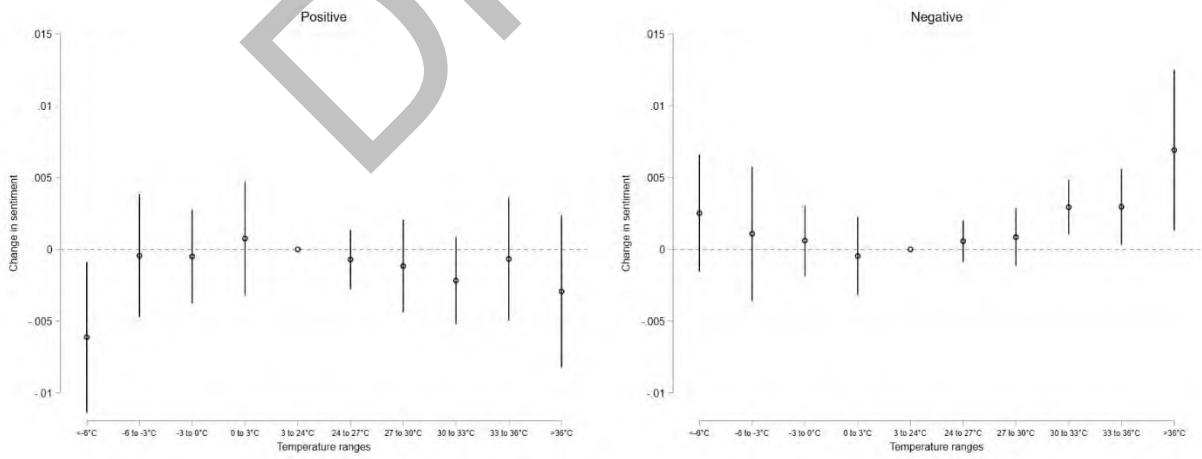
**Figure 1. Maximum temperature impact on positive and negative sentiment**




Note: in the figure we present results based on the equation 1 with 95% confidence intervals. On the left panel are presented results for positive sentiment and negative sentiment on the right.

In Figure 2 and 3, we explore heterogenous effects of temperature on sentiment by migration background. However, we do not observe any significant differences between the two groups as the coefficients for natives in Figure 2 and migrants in figure 3 have similar effect sizes and the confidence intervals are overlapping.

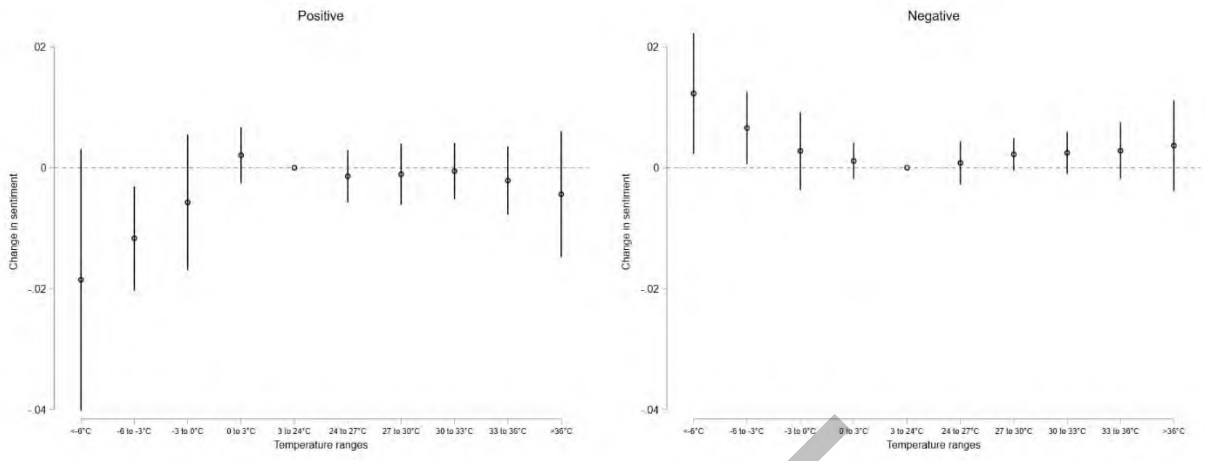
**Figure 2. Impact of temperature on native users**




Note: in the figure we present results based on the equation 1 for the native sample. Coefficients with 95% confidence intervals. On the left panel are presented results for positive sentiment and negative sentiment on the right.

DRAFT

**Figure 3. Impact of temperature on migrant user sentiment**

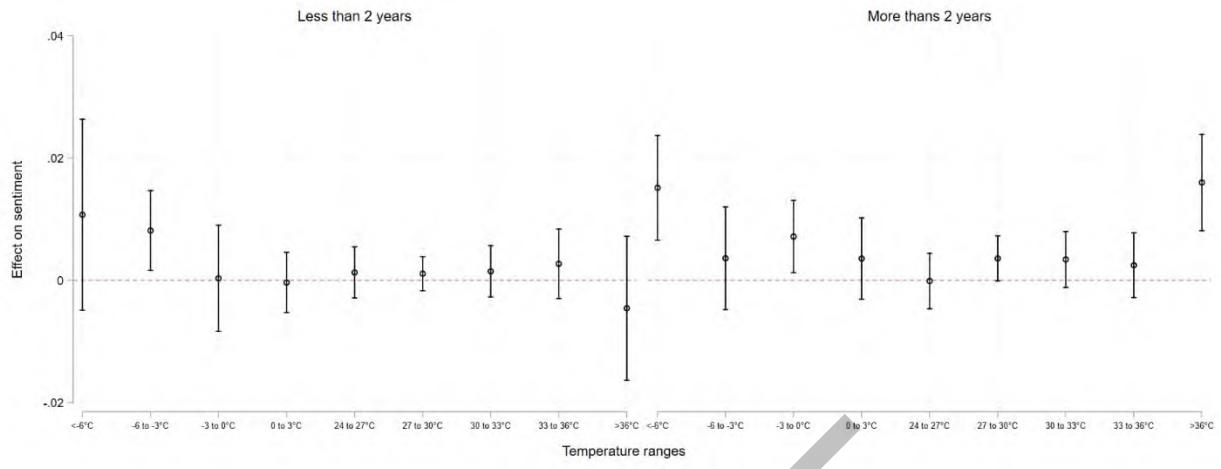



Note: in the figure we present results based on the equation 1 and for the migrant sample. Coefficients with 95% confidence intervals. On the left panel are presented results for positive sentiment and negative sentiment on the right.

Despite not finding any differences in the impact of temperature between natives and migrants we could observe some variation within migrants. In Figure 4, we observe a larger increase in negative sentiment with heat (>36°C) and cold(<6°C) for migrants that resided in the location for more than 2 years compared to those that resided less than 2 years. Possibly, the results could suggest migrants that resided for longer in the location to be more concerned about unusual weather events.

DRAFT

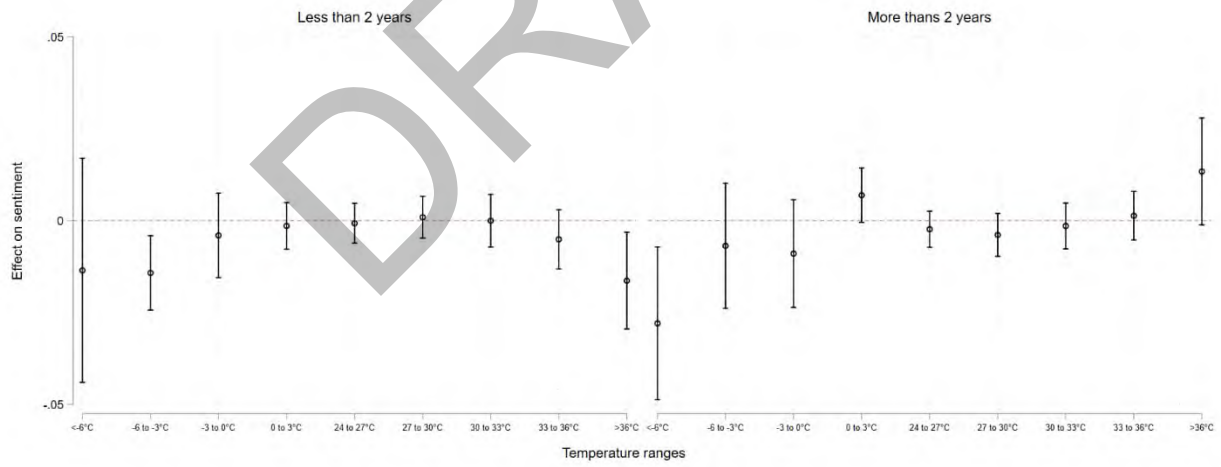
**Figure 4. Impact of temperature on negative sentiment of migrants by permanence time**




Note: in the figure we present results based on the equation 1 for the sample of migrants and an interaction with a dummy variable differentiating individuals that have been in the location for more than two years (0-1). With 95% confidence intervals.

Conversely, we observe in Figure 5 migrants that have resided less than two years in the location to experience a larger decrease in positive sentiment in their tweets when exposed to heat (>36°C). Possibly, the results could suggest a higher impact of such weather events on their mood.

**Figure 5. Impact of temperature on positive sentiment of migrants by permanence time**




Note: in the figure we present results based on the equation 1 for the sample of migrants and an interaction with a dummy variable differentiating individuals that have been in the location for more than two years (0-1). With 95% confidence intervals.

DRAFT

## Discussion and conclusion

In this article, we have analyzed the impact of temperature on the expressed sentiment of a sample of Twitter users residing in the United States. We provided three main findings. Extreme cold and heat reduce positive sentiment and increase negative sentiment of Twitter users as found in previous research (Baylis et al., 2018; Wang et al., 2020). Secondly, we did not observe any substantive differences in the impact of heat on cold between native and migrants. Finally, we found differences in the impact of temperature within migrants depending on their length of stay in the United States. Interestingly, we observe a larger increase of negative sentiment with heat and cold exposure for migrants that have resided in the location for more than two years suggesting a higher concern to the temperature extremes. Conversely, migrants that have recently relocated to the US experience a larger decrease in positive sentiment when exposed to heat that could be related to their higher sensitivity to such weather extremes.

Our study has three main limitations. First, a problem we share with existing studies is that we cannot infer how expressed sentiment online affects behavior or mental health offline. Secondly, the findings should be carefully generalized to other populations, as our sample of Twitter users is not representative of the U.S. general population or of other Twitter users. Nevertheless, its strength is the availability of information on individual's country of origin. Thirdly, we are not able to differentiate if the tweets contain negative or positive information related to the weather, climate change or to user specific state of mind. However, previous research has shown results to be robust when excluding weather specific tweets from the analysis of temperature effect on expressed sentiment of individuals(Baylis et al., 2018).

In following analysis, we will explore additional heterogeneities in the sample of migrants. For example, differences could exist between migrants depending on their country of origin and their location of relocation in the United States.

In conclusion, new geotagged digital data sources combined with meteorological data allow to explore the impact of temperature extreme on populations for which currently there is small evidence of a stratified effect.

DRAFT

## References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, *33*(1), 121–131. https://doi.org/10.1002/joc.3413

Adger, W. N., Barnett, J., Heath, S., & Jarillo, S. (2022). Climate change affects multiple dimensions of well-being through impacts, information and policy responses. *Nature Human Behaviour*, *6*(11), Article 11. https://doi.org/10.1038/s41562-022-01467-8

Almås, I., Auffhammer, M., Bold, T., Bolliger, I., Dembo, A., Hsiang, S. M., Kitamura, S., Miguel, E., & Pickmans, R. (2019). *Destructive Behavior, Judgment, and Economic Decision-making under Thermal Stress* (Issue 25785). National Bureau of Economic Research. https://doi.org/10.3386/w25785

Baylis, P., Obradovich, N., Kryvasheyeu, Y., Chen, H., Coviello, L., Moro, E., Cebrian, M., & Fowler, J. H. (2018). Weather impacts expressed sentiment. *PLOS ONE*, *13*(4), e0195750. https://doi.org/10.1371/journal.pone.0195750

Cai, X., Lu, Y., & Wang, J. (2018). The impact of temperature on manufacturing worker productivity: Evidence from personnel data. *Journal of Comparative Economics*, *46*(4), 889–905. https://doi.org/10.1016/j.jce.2018.06.003

Chang, T. Y., & Kajackaite, A. (2019). Battle for the thermostat: Gender and the effect of temperature on cognitive performance. *PLOS ONE*, *14*(5), e0216362. https://doi.org/10.1371/journal.pone.0216362

de Freitas, C., & Grigorieva, E. (2015). Role of Acclimatization in Weather-Related Human Mortality During the Transition Seasons of Autumn and Spring in a Thermally Extreme Mid-Latitude Continental Climate. *International Journal of Environmental Research and Public Health*, *12*(12), 14974–14987. https://doi.org/10.3390/ijerph121214962

Fan, Y., Wang, J., Obradovich, N., & Zheng, S. (2023). Intraday adaptation to extreme temperatures in outdoor activity. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-022-26928-y

Graff Zivin, J., & Neidell, M. (2014). Temperature and the Allocation of Time: Implications for Climate Change. *Journal of Labor Economics*, *32*(1), 1–26. https://doi.org/10.1086/671766

DRAFT

Hajdu, T., & Hajdu, G. (2019). Ambient temperature and sexual activity: Evidence from time use surveys. *Demographic Research*, *40*(12), 307–318. https://doi.org/10.4054/DemRes.2019.40.12

Heyes, A., & Saberian, S. (2019). Temperature and Decisions: Evidence from 207,000 Court Cases. *American Economic Journal: Applied Economics*, *11*(2), 238–265. https://doi.org/10.1257/app.20170223

Kim, J. S., Wang Sonne, S. E., Garimella, K., Grow, A., Weber, I. G., & Zagheni, E. (2023). *Online social integration of migrants: Evidence from Twitter* (WP-2023-012; 0 ed., p. WP-2023-012). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2023-012

Kim, K., & Ryu, D. (2022). Sentiment changes and the Monday effect. *Finance Research Letters*, *47*, 102709. https://doi.org/10.1016/j.frl.2022.102709

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Medina-Ramón, M., & Schwartz, J. (2007). Temperature, temperature extremes, and mortality: A study of acclimatisation and effect modification in 50 US cities. *Occupational and Environmental Medicine*, *64*(12), 827–833. https://doi.org/10.1136/oem.2007.033175

Mullins, J. T., & White, C. (2019). Temperature and mental health: Evidence from the spectrum of mental health outcomes. *Journal of Health Economics*, *68*, 102240. https://doi.org/10.1016/j.jhealeco.2019.102240

DRAFT

Park, R. J., Goodman, J., Hurwitz, M., & Smith, J. (2020). Heat and Learning. *American Economic Journal: Economic Policy*, *12*(2), 306–339. https://doi.org/10.1257/pol.20180612

Schweiker, M., Huebner, G. M., Kingma, B. R. M., Kramer, R., & Pallubinsky, H. (2018). Drivers of diversity in human thermal perception – A review for holistic comfort models. *Temperature*, *5*(4), 308–342. https://doi.org/10.1080/23328940.2018.1534490

Stechemesser, A., Levermann, A., & Wenz, L. (2022). Temperature impacts on hate speech online: Evidence from 4 billion geolocated tweets from the USA. *The Lancet Planetary Health*, *6*(9), e714–e725. https://doi.org/10.1016/S2542-5196(22)00173-5

Vargas, N. T., Schlader, Z. J., Jay, O., & Hunter, A. (2023). Prioritize research on human behaviour during extreme heat. *Nature Human Behaviour*, *7*(4), Article 4. https://doi.org/10.1038/s41562-023-01569-x

Wang, J., Obradovich, N., & Zheng, S. (2020). A 43-Million-Person Investigation into Weather and Expressed Sentiment in a Changing Climate. *One Earth*, *2*(6), 568–577. https://doi.org/10.1016/j.oneear.2020.05.016

Wilde, J., Apouey, B. H., & Jung, T. (2017). The effect of ambient temperature shocks during conception and early pregnancy on later life outcomes. *European Economic Review*, *97*, 87–107. https://doi.org/10.1016/j.euroecorev.2017.05.003

DRAFT

# An alternative to the Dirichlet-multinomial regression model for microbiome data analysis

Ascari Roberto[a], Migliorati Sonia[a], and Ongaro Andrea[a]

[a]Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca;
roberto.ascari@unimib.it, sonia.migliorati@unimib.it,
andrea.ongaro@unimib.it

**Abstract**

The prevalence of human microbiome data in biomedical research has increased due to the association observed between microbiome composition and several diseases. The Dirichlet-multinomial distribution is frequently used to analyze this type of data, but it often fails to adequately model real microbiome datasets due to the restrictive parameterization imposed on its covariance matrix. This work proposes a novel distribution to be considered in microbiome data analysis, which can be used to define an alternative regression model for multivariate count data (e.g., microbiome data). The new distribution is a structured finite mixture model with Dirichlet-multinomial components. We show how this mixture model can enhance microbiome data analysis by clustering patients into "enterotypes", a classification based on the taxa composition of gut microbiota. Finally, we consider an application based on a real gut microbiome dataset.

***Keywords:*** Bayesian inference, Count data, Discrete simplex, Mixture model, Multivariate regression

DRAFT

## 1. Introduction

The set of genes associated with the microbiota (i.e., bacteria, viruses, and unicellular eukaryotes living in the human body) is usually referred to as the human microbiome (1; 6). The relationship between human beings and microbiota is generally mutually beneficial, but changes in the gut microbiome can be negatively associated with health outcomes, such as diabetes, cardiovascular disease, obesity, autoimmune disease, and anxiety (4; 10; 12). However, the attention that microbiome research is experiencing is not solely due to the above-mentioned medical associations, but also to remarkable improvements in DNA sequencing technology, making microbiome data easier to be collected.

The Dirichlet-multinomial (DM) distribution is commonly used for analyzing microbiome data (3; 14). The DM, by assuming that the probability vector underlying a multinomial distribution follows a Dirichlet distribution, enriches the analysis of microbiome datasets by allowing for overdispersion. Nonetheless, its covariance structure still results too rigid, making the DM inadequate for modeling real microbiome datasets and hindering the description of co-occurrence and co-exclusion relationships between microbial taxa. This study proposes a new distribution generalizing the DM, the extended flexible Dirichlet-multinomial (EFDM), and a regression model based on it. The EFDM provides a better fit to real microbiome data while still allowing for a clear interpretation of its parameters. As a finite mixture model with DM components, the EFDM can account for latent groups in the data, enabling the identification of clusters sharing similar microbiota compositions.

## 2. Statistical models for microbiome data

Given a specific biological sample, microbiome data represent its composition in terms of $D$ bacterial taxa. More formally, we can define a microbiome dataset as a collection of $D$-dimensional vectors $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N\}$, where the $r$-th element $Y_{ir}$ of $\mathbf{Y}_i$ counts the number of occurrences of taxon $r$ in the $i$-th sample, $i = 1, \ldots, N$ and $r = 1, \ldots, D$. Since the $i$-th sample contains a total number $n_i$ of bacteria, microbiome observations satisfy $\sum_{r=1}^{D} Y_{ir} = n_i$. From a statistical point of view, this implies that the support of $\mathbf{Y}_i$ is the $D$-part discrete simplex $\mathcal{S}_{n_i}^{D} = \left\{\mathbf{y} = (y_1, \ldots, y_D)^\intercal : y_r \in \{0, 1, \ldots, n_i\}, \sum_{r=1}^{D} y_r = n_i\right\}$.

### 2.1 Distributions on the discrete simplex

The DM is a widespread generalization of the multinomial distribution that can be obtained by taking advantage of a compound approach (7). More specifically, one can assume that $\mathbf{Y}|\mathbf{\Pi} = \boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$ and that $\mathbf{\Pi} \in \mathcal{S}^D = \{\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^\intercal : \pi_r > 0, \sum_{r=1}^{D} \pi_r = 1\}$ (i.e., the $D$-part continuous simplex) follows a mean-precision parameterized Dirichlet distribution:

$$f_{\text{Dir}}(\boldsymbol{\pi}; \boldsymbol{\mu}, \alpha^+) = \frac{\Gamma(\alpha^+)}{\prod_{r=1}^{D} \Gamma(\alpha^+ \mu_r)} \prod_{r=1}^{D} \pi_r^{(\alpha^+ \mu_r)-1},$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{\Pi}] \in \mathcal{S}^D$, and $\alpha^+ > 0$ is a precision parameter. Within this framework, $\mathbf{Y}$ is marginally distributed as a DM, with probability mass function (p.m.f.)

$$f_{\text{DM}}(\mathbf{y}; n, \boldsymbol{\mu}, \alpha^+) = \frac{n!\Gamma(\alpha^+)}{\Gamma(\alpha^+ + n)} \prod_{r=1}^{D} \frac{\Gamma(\alpha^+ \mu_r + y_r)}{(y_r!)\Gamma(\alpha^+ \mu_r)}.$$

The mean vector of a DM distribution is $\mathbb{E}[\mathbf{Y}] = n\boldsymbol{\mu}$, so the parameter $\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}]/n$ can be thought of as a (scaled) mean vector. Moreover, its covariance matrix is

$$\mathbb{V}[\mathbf{Y}] = n\mathbf{M}\left[1 + \frac{n-1}{\alpha^+ + 1}\right], \tag{1}$$

where $\mathbf{M} = (\text{Diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal)$ coincides with the multinomial covariance matrix by considering $\boldsymbol{\mu} = \boldsymbol{\pi}$. On the one hand, Equation (1) highlights how the additional parameter $\alpha^+$ allows increased flexibility in the variability structure with respect to the standard multinomial distribution. On the other hand, it is clear that the DM enriches the multinomial by adding just a single parameter (namely $\alpha^+$) in the modelization of the covariance matrix, which may not be enough, especially with a large value of $D$.

We propose to take advantage of an alternative sound distribution defined on $\mathcal{S}^D$, namely the extended flexible Dirichlet (EFD, (9)). The EFD distribution can be defined as a structured finite mixture with Dirichlet components, entailing some constraints among the components' parameters to ensure the identifiability of the model.

Thanks to its mixture structure, the probability density function (p.d.f.) of an EFD-distributed random vector can be expressed as

$$f_{\text{EFD}}(\boldsymbol{\pi}; \boldsymbol{\mu}, \alpha^+, \mathbf{p}, \mathbf{w}) = \sum_{j=1}^{D} p_j f_{\text{Dir}}\left(\boldsymbol{\pi}; \boldsymbol{\lambda}_j, \frac{\alpha^+}{1 - w_r}\right), \tag{2}$$

where $\boldsymbol{\lambda}_j$ is the $j$-th component-specific mean vector defined by

$$\boldsymbol{\lambda}_j = \frac{1 - w_r}{1 - \mathbf{p} \cdot \mathbf{w}^\intercal}(\boldsymbol{\mu} - \mathbf{p} \circ \mathbf{w}) + w_r \mathbf{e}_j, \tag{3}$$

$\boldsymbol{\mu} \in \mathcal{S}^D$ is the overall mean vector of $\mathbf{\Pi}$, $\alpha^+$ is a real positive value, $\mathbf{p} \in \mathcal{S}^D$ is the vector of mixing weights, $\mathbf{e}_j$ is a vector with all elements equal to zero except for the $j$-th which is equal to one, and

DRAFT

$\mathbf{w} = (w_1, \ldots, w_D)^\intercal$ is a $D$-dimensional vector whose $r$-th element lies in the $\left(0, \min\left\{1, \frac{\mu_r}{p_r}\right\}\right)$ interval and controls the distance among the components' barycenters (9). Here, $(\mathbf{a} \circ \mathbf{b})$ denotes the Hadamard (i.e., element-wise) product between vectors $\mathbf{a}$ and $\mathbf{b}$.

If the parameter $\mathbf{\Pi}$ of the multinomial is supposed to be EFD distributed, an alternative discrete distribution for count vectors is defined, namely the extended flexible Dirichlet-multinomial (EFDM). The p.m.f. of the EFDM can be expressed as

$$f_{\text{EFDM}}(\mathbf{y}; n, \boldsymbol{\mu}, \alpha^+, \mathbf{p}, \mathbf{w}) = \sum_{j=1}^{D} p_j f_{\text{DM}}\left(\mathbf{y}; n, \boldsymbol{\lambda}_j, \frac{\alpha^+}{1 - w_r}\right), \tag{4}$$

where $\boldsymbol{\lambda}_j$ is defined as in Equation (3). It is interesting to note that when $\mathbf{p} = \boldsymbol{\mu}$ and $w_r = 1/(\alpha^+ + 1)$ for every $r = 1, \ldots, D$, then the DM distribution is recovered. Thus, the EFDM is a generalization of the DM distribution, including it as inner point.

Equation (4) shows that the EFDM is a finite mixture with DM components displaying a precise form for their (scaled) mean vectors $\boldsymbol{\lambda}_j$, $j = 1, \ldots, D$. Thanks to the properties of a finite mixture model, the overall EFDM mean vector can be expressed as:

$$\mathbb{E}[\mathbf{Y}] = n \sum_{r=1}^{D} p_r \boldsymbol{\lambda}_r. \tag{5}$$

It is possible to show that the EFDM covariance matrix is characterized by $2D - 1$ additional parameters with respect to the DM, namely $D - 1$ distinct elements in the vector of mixing weights $\mathbf{p}$, and the $D$ elements in the vector $\mathbf{w}$. This is the key element explaining the better ability of the EFDM in modeling a wide range of scenarios.

In Figure 1, the p.m.f. of the DM and EFDM distributions are compared through a discrete ternary diagram. In this example, $D = 3$ is selected and a clear bimodality of the EFDM can be noted. Indeed, even if the EFDM considers three non-empty components, its p.m.f. shows only two distinct modes due to two small values in $\mathbf{w}$, placing the second and the third clusters close to each other.

DRAFT

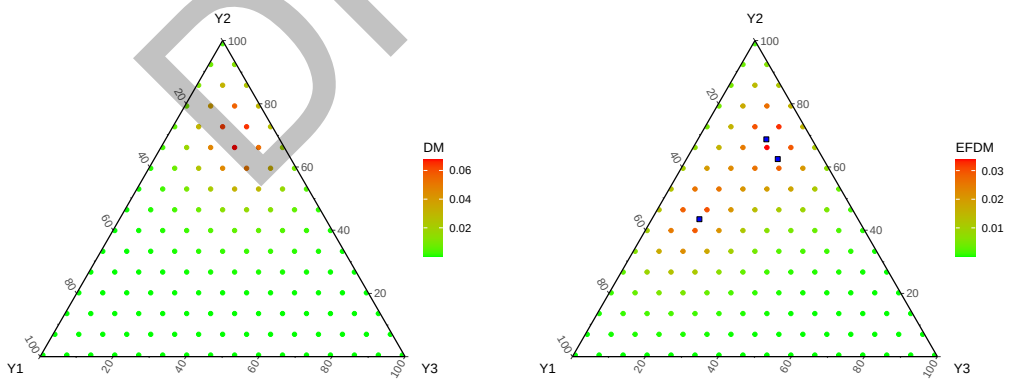



Figure 1: Probability mass function of the DM (left panel) and the EFDM (right panel) distributions with $n = 15$, $\boldsymbol{\mu}_{DM} = (0.13, 0.67, 0.2)^\intercal$, $\boldsymbol{\mu}_{EFDM} = (0.28, 0.54, 0.18)^\intercal$, $\alpha^+ = 150$, $\mathbf{w} = (0.348, 0.063, 0.063)^\intercal$, and $\mathbf{p} = (0.5, 0.25, 0.25)^\intercal$. Blue squares represent the component specific barycenters $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and $\boldsymbol{\lambda}_3$.

## 3. Regression models for microbiome data

To perform a regression analysis on a microbiome dataset, we start by considering a set of independent multivariate responses $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$ collected on $N$ subjects. Each response $\mathbf{Y}_i$ records the count (i.e., the frequency) of $D$ possible taxa among $n_i$ trials, so that the elements of $\mathbf{Y}_i$ sum to $n_i$. Furthermore, the $i$-th subject is associated with a $(K+1)$-dimensional vector of covariates $\mathbf{x}_i$. A parameterization of the EFDM model that is useful for regression analysis involves the parameters $\boldsymbol{\mu}, \mathbf{p}, \alpha^+, \tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_D)^\intercal$, where

$$\tilde{w_r} = \frac{w_r}{\min\left\{1, \frac{\mu_r}{p_r}\right\}} \in (0,1) \tag{6}$$

is a normalized version of $w_r$.

Two regression models can be defined: the EFDM regression (EFDMReg) and the DM regression (DMReg). EFDMReg assumes that each $\mathbf{Y}_i$ follows an EFDM$(n_i, \boldsymbol{\mu}_i, \alpha^+, \mathbf{p}, \tilde{\mathbf{w}})$ distribution, while DMReg assumes a DM$(n_i, \boldsymbol{\mu}_i, \alpha^+)$ distribution. In both models, by following a GLM-type approach (5), we can link the parameter $\boldsymbol{\mu}_i$ to the linear predictor by taking advantage of a proper link function, such as the multinomial logit link function:

$$g(\mu_{ir}) = \log\left(\frac{\mu_{ir}}{\mu_{iD}}\right) = \mathbf{x}_i^\intercal \boldsymbol{\beta}_r, \qquad r = 1, \ldots, D-1, \tag{7}$$

where $\boldsymbol{\beta}_r = (\beta_{r0}, \beta_{r1}, \ldots, \beta_{rK})^\intercal$ is a vector of regression coefficients for the $r$-th element of $\boldsymbol{\mu}_i$. The last taxon has been conventionally chosen as the baseline category, thus $\boldsymbol{\beta}_D = \mathbf{0}$, but any other taxa could have been selected as the baseline.

We shall adopt a Bayesian approach to inference, therefore a prior distribution for all parameters is needed. The parameterization of EFDMReg based on $\boldsymbol{\mu}, \mathbf{p}, \alpha^+$, and $\tilde{\mathbf{w}}$ defines a variation-independent parameter space, which allows assuming prior independence. As a consequence, we can elicit a prior distribution for each parameter separately. We prefer to adopt a weakly informative prior for each parameter, to induce minimum impact on the posterior distribution. In the following, we consider:

- $\boldsymbol{\beta}_r \sim N_{K+1}(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ is the $(K+1)$-vector with zero elements, and $\Sigma$ is a diagonal matrix with 'large' variance values;
- $\alpha^+ \sim Gamma(kg, g)$ for small values of the rate parameter $g$ so to induce a large variability around the prior mean $k$;
- $\tilde{w}_r \sim Unif(0,1)$, $r = 1, \ldots, D$;
- a uniform prior on the simplex for $\mathbf{p}$.

Inferential issues are dealt with by the Hamiltonian Monte Carlo (HMC) algorithm (8), which is a popular generalization of the Metropolis-Hastings. More specifically, the Stan modeling language (11) allows the implementation of an HMC method to obtain a simulated sample from the posterior distribution. To compare the fit of the models, we use the Watanabe-Akaike information criterion (WAIC) (13; 15), a Bayesian criterion that balances goodness-of-fit and model complexity. Models with lower WAIC values generally provide a better fit to the dataset.

## 4. An application to gut microbiome data

In this section, we show the results of the DMReg and EFDMReg models in a real-case scenario. In particular, we applied the DM and EFDM regression models to a microbiome dataset that was previously analyzed by Wu et al. (16) and Xia et al. (17). The dataset consisted of gut microbiome data from $N = 98$ healthy volunteers, with the counts of $D = 3$ bacterial genera (*Bacteroides*, *Prevotella*, and *Ruminococcus*) recorded. Arumugam et al. (2) used these bacteria to define three groups they called "enterotypes", which can provide information about the human body's ability to produce vitamins. Wu et al. conducted a cluster analysis on these data by using the partitioning around medoids (PAM) approach

DRAFT

and identified two of the three enterotypes defined by Arumugam et al, namely enterotypes 1 and 2. Moreover, these two clusters are characterized by different frequencies: 86 out of the 98 samples were allocated to the first enterotype, whereas only 12 samples were clustered into enterotype 2. This is due to the small number of subjects with a high abundance of Prevotella (i.e., only 36 samples showed a Prevotella count greater than 0).

In addition to bacterial data, we considered $K = 9$ covariates representing information on micronutrients in the habitual long-term diet, which were collected by using a food frequency questionnaire. These additional covariates were selected by Xia et al. using an $l_1$ penalized logistic normal multinomial regression approach. Table 1 shows the posterior mean and 90% credible set of each parameter involved in the DMReg and EFDMReg models. While the significant covariates are the same across the models, the EFDMReg model has a lower WAIC, indicating a better fit. This is due to the additional set of parameters involved in the mixture structure.

Table 1: Posterior mean and 95% CS for the parameters of the DMReg and EFDMReg models. Regression coefficients in bold are related to 90% CS's not containing the zero value.

| | | DMReg | | EFDMReg | |
|---|---|---|---|---|---|
| | | Post. Mean | 90% CS | Post. Mean | 90% CS |
| Bacteroides | Intercept | **2.208** | **(1.901, 2.514)** | **2.583** | **(2.209, 2.954)** |
| | Proline | -0.042 | (-0.297, 0.217) | -0.039 | (-0.282, 0.209) |
| | Sucrose | **-0.263** | **(-0.518, -0.008)** | **-0.261** | **(-0.503, -0.020)** |
| | Vitamin E, food fortification | -0.023 | (-0.316, 0.276) | -0.023 | (-0.302, 0.258) |
| | Beta cryptoxanthin | -0.064 | (-0.310, 0.190) | -0.056 | (-0.290, 0.184) |
| | Added germa from wheats | -0.156 | (-0.439, 0.135) | -0.159 | (-0.436, 0.121) |
| | Vitamin C | **0.318** | **(0.017, 0.690)** | **0.297** | **(0.024, 0.638)** |
| | Maltose | -0.016 | (-0.254, 0.227) | -0.005 | (-0.230, 0.224) |
| | Palmitelaidic trans fatty acid | 0.018 | (-0.241, 0.278) | 0.003 | (-0.238, 0.248) |
| | Acrylamide | 0.118 | (-0.138, 0.386) | 0.144 | (-0.102, 0.396) |
| Prevotella | Intercept | **-1.147** | **(-1.581, -0.739)** | **-0.715** | **(-1.211, -0.234)** |
| | Proline | -0.059 | (-0.483, 0.362) | -0.077 | (-0.495, 0.340) |
| | Sucrose | 0.015 | (-0.379, 0.401) | 0.002 | (-0.371, 0.376) |
| | Vitamin E, food fortification | 0.090 | (-0.296, 0.472) | 0.101 | (-0.277, 0.469) |
| | Beta cryptoxanthin | 0.274 | (-0.128, 0.678) | 0.297 | (-0.096, 0.682) |
| | Added germa from wheats | 0.263 | (-0.097, 0.620) | 0.204 | (-0.149, 0.547) |
| | Vitamin C | -0.125 | (-0.941, 0.541) | -0.113 | (-0.908, 0.530) |
| | Maltose | **0.650** | **(0.244, 1.063)** | **0.687** | **(0.278, 1.096)** |
| | Palmitelaidic trans fatty acid | **-0.522** | **(-0.928, -0.120)** | **-0.542** | **(-0.936, -0.147)** |
| | Acrylamide | **0.754** | **(0.396, 1.118)** | **0.748** | **(0.404, 1.101)** |
| | $\alpha^+$ | 1.519 | (1.148, 1.945) | 1.614 | (1.098, 2.285) |
| | $p_1$ | — | — | 0.714 | (0.253, 0.910) |
| | $p_2$ | — | — | 0.210 | (0.064, 0.408) |
| | $p_3$ | — | — | 0.076 | (0.000, 0.625) |
| | $\tilde{w}_1$ | — | — | 0.528 | (0.420, 0.636) |
| | $\tilde{w}_2$ | — | — | 0.602 | (0.491, 0.707) |
| | $\tilde{w}_3$ | — | — | 0.486 | (0.374, 0.596) |
| | WAIC | 1686.3 | | 1668.3 | |

The mixture structure of the EFDMReg model can be utilized to cluster observations into groups through a model-based approach. Specifically, each observation can be allocated to the mixture component that most likely generated it. The mixing weight estimates (0.714, 0.210, and 0.076) from Table 1 confirm the presence of two out of the three enterotypes defined by Arumugam et al. To illustrate the benefits of the EFDMReg model in microbiome data analysis, we compared the clustering profile

DRAFT

Table 2: Confusion matrix for clustering based on the EFDMReg model compared to the PAM algorithm.

| | | EFDMReg | |
|---|---|---|---|
| | | 1 | 2 |
| PAM | 1 | 78 | 8 |
| | 2 | 2 | 10 |

obtained by the EFDMReg model and the one obtained with the PAM approach used by Wu et al. Table 2 summarizes this comparison in a confusion matrix, which shows that the clustering generated by the EFDMReg model highly agrees with the one obtained by the PAM algorithm (i.e., for 91.7% of the observations). This percentage is obtained using the covariates selected by Xia et al. in a logistic normal multinomial regression model context. Clearly, the results could be improved by developing an ad hoc variable selection procedure for the EFDMReg model. In conclusion, the EFDM distribution and the EFDMReg model are new statistical tools to be used in the analysis of microbiome data (as well as any other kind of multivariate count data). The main gain of the EFDMReg, as well as the better fit than the standard DMReg model, lies in its ability to jointly perform regression and clustering observations into some biological-relevant groups.

DRAFT

## References


[1] Amato, K.: An introduction to microbiome analysis for human biology applications. American Journal of Human Biology, **29** (2017).

[2] Arumugam, M. et al: Enterotypes of the human gut microbiome. Nature, **473**, 174–180 (2011)

[3] Chen, J., Li, H.: Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. The Annals of Applied Statistics, **7**(1), 418–442 (2013)

[4] Koeth, RA. et al: Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nature Medicine, **19**(5) (2013)

[5] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall (1989)

[6] Morgan, XC., Huttenhower, C.: Human microbiome analysis. PloS Computational Biology, **8**(12) (2012)

[7] Mosimann, J.e.: On the compound multinomial distribution, the Multivariate $\beta$-distribution, and correlations among proportions. Biometrika, **49**, 65–82 (1962)

[8] Neal, R.M.: An improved acceptance procedure for the hybrid Monte Carlo algorithm (1994)

[9] Ongaro, A., Migliorati, S., Ascari, R.: A new mixture model on the simplex. Statistics and Computing, **30**(4), 749–770 (2020)

[10] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y.: A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, **490**, 55–60 (2012)

[11] Stan Development Team: Stan Modeling Language Users Guide and Reference Manual (2017)

[12] Turnbaugh, PJ. et al: A core gut microbiome in obese and lean twins. Nature, **457**, 480–484 (2009)

[13] Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing. **27**(5), 1413–1432 (2017)

[14] Wadsworth, WD., Argiento, R., Guindani, M., Galloway-Pena, J., Shelburne, SA., Vannucci, M.: An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinformatics, **18**(94) (2017)

[15] Watanabe, S.: A widely applicable Bayesian information criterion. Journal of Machine Learning Research. **14**(1), 867–897 (2013)

[16] Wu., G.D. et al.: Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science. **334**, 105–109 (2011)

[17] Xia, F., Chen, J., Fung, W.K., Li, H.: A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics. **69**(4), 1053–1063 (2013)

# Modelling ordinal response to treatment in a real-world cohort study

Marco Alfò[a], Maria Francesca Marino[b], and Silvia D'Elia[a]

[a]Dipartimento di Scienze Statistiche; marco.alfo@uniroma1.it,
silvia.delia@uniroma1.it
[b]Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti";
mariafrancesca.marino@unifi.it

**Abstract**

The Italian Thyroid Cancer Observatory foundation maintains a database including all records on patients with confirmed diagnosis of thyroid cancer reported by several specialized centers in the country. One of the objectives of this project is to monitor the evolution over time of the response to treatment, which is a synthesis of serum values and ultrasound imaging, measured over an ordinal 4-point scale. The response to treatment is measured at 12 months, 3 and 5 years since the initial treatment; patients may underwent additional treatments, between measurement occasions, which may alter the initial risk composition. Our analysis focuses on modelling the response at 12 months as a function of baseline risk classification, clinical and surgical information. Further, we aim at exploring the transition between response categories when we consider the 12 months and 3 years measurements, since these may be affected by additional treatments occurred in the meanwhile.

***Keywords:*** ordinal probit, multilevel study, unobserved heterogneity

DRAFT

## 1. The case study

The Italian Thyroid Cancer Observatory (ITCO) collects real-world practice data on patients with histologically confirmed diagnoses of differentiated, medullary, poorly differentiated or anaplastic, thyroid cancer. Data collection is performed via a web-based database made available since 2013 by the Thyroid Cancer Center at Sapienza University of Rome (the network's Coordinating Center). It has then been expanded to include data from several other centers in the country. It now includes more than 10,000 prospectively collected cases, registered at the time of primary treatment in one of the reporting ITCO centers. If the patient underwent surgery elsewhere, he/she is included when starting the follow-up in the reporting center (if within 12 months after primary treatment). Each record includes information on patient demographics and biometrics, diagnosis, pathology, surgical and radioactive iodine treatments, and information recorded at periodic follow-up examinations. For the purposes of this study, we focus on cases that match the following inclusion criteria:

1. histological diagnosis of differentiated thyroid cancer, including papillary, follicular, and poorly differentiated tumors;
2. registration in the ITCO database before February 1, 2023;
3. presence of clinical evaluation between 6 and 18 months after the primary treatment, when the response to the initial treatment is assessed.

Exclusion criteria are:

1. histological diagnosis of non-invasive follicular thyroid neoplasms with papillary-like nuclear features, unknown malignant potential tumors, medullary, or anaplastic thyroid cancer;
2. lack of complete information on the initial treatment or pathology.

For each case, several information are available, ranging from initial surgical treatment procedure, cervical lymph node dissection, radioactive iodine remnant ablation. The risk category was estimated by the study team according to the 2009 American Thyroid Association (ATA) guidelines (4), as modified by the 2015 release (6). The classification is based on data recorded at the time of initial treatment with a few exceptions in the presence of specific surgical or radioiodine remnant ablation (RRA) treatments, and for aggressive papillary thyorid carcinomas PTCs. The response to the initial treatment is calculated based on available clinical evaluation performed at 12 months (range 6-18 moths) since the initial treatment. It is an ordinal response with 4 categories, as described in the following, based on the evidence obtained from neck ultrasound [US] in all patients, and radioiodine scintigraphy in selected cases. Further information for defining the response is derived from basal/stimulated serum thyroglobulin (Tg) and anti-Tg antibody (TgAb) levels. Guidelines come from ATA when looking at patients undergoing thyroidectomy and RRA, and from the European Society for Medical Oncology (ESMO) as for the others. Ultrasound imaging was considered as indicative of an evidence of a disease in the presence of high suspicion lymph nodes, according to the European Thyroid Association guidelines stratification (7), while low suspicion lymph nodes were grouped in nonspecific findings (8).

## 2. The statistical model

Our purpose is to model the response to treatment at 12 months $Y_{ij}$, for patient $i = 1, \dots, n_j$, followed by center $j = 1, \dots, m$. Since the response is ordinal with $G = 4$ categories, we use an underlying random variable approach, see eg (3) for a thorough review. This dates back at least to the 50s, see e.g. (2) or (11) for early developments. The idea is that the *observed response* comes out of the discretization of an underlying latent variable $Y_{ij}^*$ described by the following linear regression model

$$Y_{ij}^* = \sum_{l=1}^{q} z_{ijl}\beta_l + u_j + \varepsilon_{ij} = \mathbf{z}_{ij}'\boldsymbol{\beta} + u_j + \varepsilon_{ij},$$

where $z(\mathbf{x}_{ij}) = \mathbf{z}_{ij}$ is a $q$-dimensional design vector based on the $p$-dimensional vector of individual features $\mathbf{x}_{ij}$, $i = 1, \dots, n_j$, $j = 1, \dots, m$. Given the hierarchical structure of the available data, with patients nested within treatment centers, a mixed effect specification is used. That is, each center is associated to a specific intercept $u_j$, $j = 1, \dots, m$. The usual assumption of conditional independence holds; further assumptions entail independence of $u_j$ and $\varepsilon_{ij}$, $i = 1, \dots, n_j$, $j = 1, \dots, m$, and *weak* exogeneity of the design vector with respect to the errors terms. Let us consider a set of strictly increasing thresholds, $-\infty = \alpha_0 < \alpha_1 \cdots < \alpha_{G-1} < \alpha_G = +\infty$; the underlying response variable (URV) approach is based on the assumption that $Y_{ij} = g$ if and only if $\alpha_{g-1} \leq Y_{ij}^* < \alpha_g$, $g = 1, \dots, G$. So that

$$\begin{aligned}\Pr\left(Y_{ij} \leq g \mid u_j, \mathbf{x}_{ij}\right) &= \Pr\left(Y_{ij}^* < \alpha_g \mid u_j, \mathbf{x}_{ij}\right) = \Pr\left(\mathbf{z}_{ij}'\boldsymbol{\beta} + u_j + \varepsilon_{ij} < \alpha_g \mid u_j, \mathbf{x}_{ij}\right) = \\ &= \Pr\left(\varepsilon_{ij} < \alpha_g - u_j - \mathbf{z}_{ij}'\boldsymbol{\beta} \mid u_j, \mathbf{x}_{ij}\right) = F_{\varepsilon|u,x}\left(\alpha_g - u_j - \mathbf{z}_{ij}'\boldsymbol{\beta}\right).\end{aligned}$$

The choice of non-decreasing thresholds $\alpha_g$, $g = 1, \dots, G$, and a constant parameter vector $\boldsymbol{\beta}$ are useful to ensure non-decreasing values for the cdf of the error term. Different choices for the cdf $F_{\varepsilon|u,x}$ may lead to different models for the ordered response; among these, we may recall the *cumulative* probit and logit models. Denoting by $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{G-1})'$ the vector of thresholds for the URV specification and according to all the hypotheses above, we may define the log-likelihood function as follows:

$$\ell\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi\right) = \prod_{j=1}^{m} \int_{\mathcal{U}} f_{Y|X,U}\left(\mathbf{y}_j \mid \mathbf{Z}_j, u_j\right) f_{U|X}\left(u_j \mid \mathbf{Z}_j\right) \mathrm{d}u_j,$$

DRAFT

where

$$u_j \overset{iid}{\sim} f_U(\cdot \mid \phi), \quad j = 1, \ldots, m,$$

and

$$f_{Y|X,U}(\mathbf{y}_j \mid \mathbf{Z}_j, u_j) = \prod_{i=1}^{n_j} \prod_{g=1}^{G} \theta_{ijg}^{d_{ijg}}.$$

Here, the parameter $\theta_{ijg}$ is defined as

$$\theta_{ijg} = \Pr(Y_{ij} = g \mid u_j, \mathbf{x}_{ij}) = \Pr(Y_{ij} \leq g \mid u_j, \mathbf{x}_{ij}) - \Pr(Y_{ij} \leq g - 1 \mid u_j, \mathbf{x}_{ij}),$$

while $d_{ijg}$ takes unit value if and only if $Y_{ij} = G$; that is, $d_{ijg} = \mathbf{1}(Y_{ij} = G)$, with $\mathbf{1}(\cdot)$ being the indicator function. The interested reader may refer to (1) for a comprehensive review of models for clustered ordinal responses. For a parametric assumption on the distribution of center-specific intercepts $u_j$, $j = 1, \ldots, m$, we may consider approximation techniques based on Gaussian quadrature and its extensions, see (9), (10). As it can be noticed, the distribution of the center-specific intercepts $u_j$, $j = 1, \ldots, m$, in the previous integral is conditional on the observed covariates, and this helps take into account potential dependence between center-specific observed and unobserved features. To handle this dependence, we approximate $f_{U|X}(\cdot \mid \cdot)$ by a discrete distribution specified on a finite set of locations $\{\zeta_1, \ldots, \zeta_K\}$ with associated masses given by $\pi(\zeta_k \mid \mathbf{X}_j)$, leading to the following log-likelihood function

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) = \prod_{j=1}^{m} \sum_{k=1}^{K} f_{Y|X,U}(\mathbf{y}_j \mid \mathbf{Z}_j, u_j = \zeta_k)\, \pi(\zeta_k \mid \mathbf{X}_j),$$

where

$$\pi(\zeta_k \mid \mathbf{X}_j) \propto \exp\left[\mathbf{w}(\mathbf{X}_j)'\boldsymbol{\gamma}\right]$$

and $\mathbf{w}(\mathbf{X}_j)$ is an appropriate design vector that describes the dependence of $u_j$ on $\mathbf{X}_j$, $j = 1, \ldots, m$, as in *concomitant variable* models, see (5).

DRAFT

## 3. Data analysis plan

After about 12 months since the initial treatment, the distribution of the $n = 5,840$ patients that satisfy the inclusion criteria by the surgical approach is as follows:

Table 1: Distribution of the response at 12 months by surgical approach

| | Surgical approach | | |
|---|---|---|---|
| Response | TT | nTT | LT |
| 1 | 196 | 4 | 2 |
| 2 | 329 | 16 | 9 |
| 3 | 2285 | 88 | 5 |
| 4 | 2583 | 105 | 218 |
| Total | 5393 | 213 | 234 |

The response is classified as

1 *structural incomplete*: chemical and imaging evidence of the disease;
2 *biochemical incomplete*: chemical values above the threshold, without clear imaging evidence;
3 *indeterminate*: chemical values below the threshold, but above zero, no clear imaging evidence;
4 *excellent*: no chemical or imaging evidence.

In addition to the surgical approach (total thyroidectomy - TT, near total thyroidectomy -nTT, thyroid lobectomy -LT), we use info on patients' demographic (age and sex), clinical (ATA risk category, histology, tumor size, radioiodine remnant ablation, etc.) and surgical features (central, lateral neck dissection, number of lymph nodes metastatic and removed,etc.).

As it can be noticed from the table above, the response is excellent in the majority of the cases, but still it remains indeterminate in quite a large portion of the sample. This poses relevant clinical questions about treatment effectiveness and the reasons for indeterminacy; clinicians therefore suggested the need to extend the follow-up window to include visits up to, at least, three years since the initial treatment.

We therefore analyzed the transitions from the state (the response values) observed at 12 months and at three years since the initial treatment. The aim is that of verifying whether indeterminacy decreases with time and, if so, if this may be affected by additional treatments underwent between the two periodic follow-up visits. For this purpose, we restrict our attention to cases classified as indeterminate at 12 months since the initial treatment, and analyze whether further empirical evidence registered in that occasion (eg suspicious residual tissue in thyroid bed found by neck US) and additional treatments (in particular RRA) may help explain the transition to the state observed at 3 years since initial treatment.

## References


[1] Agresti, A., Natarajan, R. (2001) Modeling clustered ordered categorical data: a survey. *International Statistical Review*, **69**, 345–371.

[2] Aitchinson, J., Silvey, S.D. (1957) The generalization of probit analysis to the case of multiple responses. *Biometrika*, **44**, 131–140.

[3] Boes, S., Winkelmann, R. (2006) Ordered response models. *Allgemeines Statistisches Arch*, **90**, 167–181.

[4] Cooper DS, Doherty GM, Haugen BR, Hauger BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, Sherman SI, Steward DL, Tuttle RM, Cancer ATAAGToTNaDT (2009). Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. *Thyroid*, **19**: 1167–1214.

[5] Dayton, C.M., MacReady, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, **83**, 173–178.

[6] Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. (2015) 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid*, **26**:1–133.

[7] Leenhardt L, Erdogan MF, Hegedus L, Mandel SJ, Paschke R, Rago T, Russ G. (2013) 2013 European thyroid association guidelines for cervical ultrasound scan and ultrasound-guided techniques in the postoperative management of patients with thyroid cancer. *European Thyroid Journal*, **2**:147–159.

[8] Lamartina L, Grani G, Biffoni M, Giacomelli L, Costante G, Lupo S, Maranghi M, Plasmati K, Sponziello M, Trulli F, Verrienti A, Filetti S, Durante C. (2016) Risk Stratification of Neck Lesions Detected Sonographically During the Follow-Up of Differentiated Thyroid Cancer. *Journal of Clinical Endocrinology & Metabolism*. **101**,3036–3044.

[9] Pinheiro, J.C., Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects models. *Journal of Computational and Graphical Statistics*, **4**,12-35.

[10] Pinheiro, J.C., Chao, E.C. (2006). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models, *Journal of Computational and Graphical Statistics*, **15**, 58–81.

[11] Snell, E.J. (1964) A scaling procedure for ordered categorical data, *Biometrics*, **20**, 555–573


DRAFT

# On the application of the symmetric graphical lasso for paired data

Saverio Ranciati[a] and Alberto Roverato[b]

[a]Dipartimento di Scienze Statistiche, Via Belle Arti, 41, Bologna, Italia;
saverio.ranciati2@unibo.it

[b]Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, Padova, Italia;
alberto.roverato@unipd.it

**Abstract**

Graphical lasso methods are not invariant to scalar multiplication of the variables. On the other hand, Gaussian graphical models are invariant to scalar multiplication, and thus it is common practice to apply graphical lasso after the observed variables are standardized to unit sample variances. We consider the symmetric graphical lasso method for learning Gaussian graphical models for paired data and show that this family of models is not invariant to scalar multiplication of the variables, but that in the special case where homologous variables have equal variances it still makes sense to standardise the variables. We then carry out an empirical analysis to assess the impact of standardization on the symmetric graphical lasso method.

***Keywords:*** brain network, fMRI data, Gaussian graphical model, penalized likelihood method.

DRAFT

## 1. Introduction

Let $Y_V$ be a multivariate Gaussian random vector whose entries are indexed by a finite set $V = \{1, \ldots, p\}$, and let $G = (V, E)$ be an undirected graph with vertex set $V$ and edge set $E$, where an edge is an unordered pair $\{i, j\} \in E$ of distinct vertices. In a Gaussian graphical model (GGM) every vertex of $G$ is associated with a variable in $Y_V$, and the distribution of $Y_V$ is said to be Markov with respect to $G$ if every missing edge between two vertices implies that the corresponding variables are conditionally independent given the remaining variables. If we denote by $\Sigma = (\sigma_{ij})_{i,j \in V}$ and $\Sigma^{-1} = \Theta = (\theta_{ij})_{i,j \in V}$ the covariance and the concentration matrix of $Y_V$, respectively, then in a GGM it holds that for every missing edge of the graph, i.e. $\{i, j\} \notin E$ with $i \neq j$, the corresponding entry of the concentration matrix is equal to zero, $\theta_{ij} = \theta_{ji} = 0$; we refer to (4) for a comprehensive account on GGMs.

GGMs have become a popular tool in applications involving the joint learning of multiple networks, that is in the case where the observations come from two or more groups sharing the same variables, and it is expected that there are similarities between the networks associated with groups. In this context, the literature has mostly focused on the case where the groups are independent so that every network is a distinct unit, disconnected from the other networks (7), and only more recently on the case where groups are not independent (8; 9; 5; 6). More specifically, (5) and (6) considered the case of paired data, where there are exactly two dependent groups. Paired data commonly arise in paired design studies, where each subject is measured under two different conditions, or time points, as well as in matched observational studies.

In the pair data framework, $Y_V$ is partitioned into two subvectors $Y_L = (Y_L, Y_R)^\top$ where $L = \{1, \dots, p/2\}$ and $R = \{p/2+1, \dots, p\}$ and every variable in $Y_L$ has an homologous variable in $Y_R$. We set $i' = i+p/2$ and assume, without loss of generality, that $Y_i$ is paired with $Y_{i'}$. Interest is for similarities between the edge structure of the two subnetworks of $Y_L$ and $Y_R$, but also on similarities between the parameter values of the two groups. Hence, (5) considered GGMs with additional restrictions of the form

$$\theta_{ij} = \theta_{i'j'} \quad \text{for some} \quad i, j \in L. \tag{1}$$

We remark that GGMs with equality restrictions on the concentration values were introduced by (3) with the name of RCON models, and thus, formally, the constrains in (1) identify a subfamily of RCON models. (5) approached the problem of learning the model from data by implementing a penalized likelihood method, which they called the *symmetric graphical lasso* (SGL), that comprises two penalty terms: a graphical lasso penalty (2, section 9.3) to induce sparsity in the graph structure, and then a fused lasso penalty to induce the equalities (1).

One fundamental issue in the application of graphical lasso methods is that they are not invariant to scalar multiplications of the variables, and for this reason it is common practice to standardise the data to have unit sample variances; see, among others, (2, p. 8). It is not clear, however, as this preliminary step may affect the result of the analysis when a symmetric graphical lasso is applied and in this paper we investigate this problem.

## 2. Normalization in paired data problems

We first approach the problem by considering the simplifying assumption that the variances of the variables in $Y_V$ are known. Formally, we assume that $\Delta = \mathrm{diag}(\Sigma)$ is known and set $Z_V = \Delta^{-\frac{1}{2}} Y_V$ so that $\mathrm{var}(Z_V) = \Delta^{-\frac{1}{2}} \Sigma \Delta^{-\frac{1}{2}}$ coincides with the correlation matrix of $Y_V$. Accordingly, the concentration matrix of $Z_V$ is given by $\mathrm{var}(Z_V)^{-1} = \Delta^{\frac{1}{2}} \Theta \Delta^{\frac{1}{2}} = (\theta_{ij}\sqrt{\sigma_{ii}\sigma_{jj}})_{ij \in V}$ so that one can see that

$$\theta_{ij} = 0 \quad \text{if and only if} \quad \theta_{ij}\sqrt{\sigma_{ii}\sigma_{jj}} = 0. \tag{2}$$

Because the missing edges of $G$ are given by the zero pattern of $\Theta$, then it follows immediately from (2) that the distributions of $Y_V$ and that of $Z_V$ belong to the same GGM. It is straightforward that the same is true when $\Delta$ is any diagonal matrix with positive entries and this shows that GGMs are invariant with respect to scalar multiplication of the variables. Different is the case where we are interested in equality restrictions because

$$\theta_{ij} = \theta_{i'j'} \quad \text{does not imply} \quad \theta_{ij}\sqrt{\sigma_{ii}\sigma_{jj}} = \theta_{i'j'}\sqrt{\sigma_{i'i'}\sigma_{j'j'}},$$

and this shows that RCON models are not invariant with respect to the scalar multiplication of the variables, in the sense that the distributions of $Y_V$ and $Z_V$ do not generally belong to the same RCON model defined by the restrictions in (1). We can conclude that standardisation is a problematic step in the application of the symmetric graphical lasso method because it does not maintain the equality constraints of the process generating the original data.

A feature of paired data is that for every $i \in L$ the variables $Y_i$ and $Y_{i'}$ are homologous and, thus, it is somehow natural to consider the specific case where their variances are equal. Indeed, if we assume that $\sigma_{ii} = \sigma_{i'i'}$ for every $i \in L$ then it is straightforward to see that $\theta_{ij} = \theta_{i'j'}$ if and only if $\theta_{ij}\sqrt{\sigma_{ii}\sigma_{jj}} = \theta_{i'j'}\sqrt{\sigma_{i'i'}\sigma_{j'j'}}$ and therefore, in this case, as far as the equality restrictions in (1) are concerned, the distributions of $Y_V$ and $Z_V$ belong to the same RCON model.

We turn now to the case where the entries of $\Delta$ are not known and a random samples of $n$ i.i.d. observations from $Y_V \sim N(0, \Sigma)$ is available. Then, the normalization step is commonly carried out by replacing $\Delta$ with $\mathrm{diag}(S)$, where $S = (s_{ij})_{ij \in V}$ is the sample covariance matrix. Recall that in GGMs it holds that $\hat{\sigma}_{ii} = s_{ii}$ is the m.l.e. of $\sigma_{ii}$, for every $i \in V$. On the other hand, we have seen above that for the family of RCON models of interest it makes sense to standardise the variables only when homologous variables have equal variances, and in this case the m.l.e. of the variances can be computed

DRAFT

as $\hat{\sigma}_{ii} = \hat{\sigma}_{i'i'} = (s_{ii} + s_{i'i'})/2$, for every $i \in L$. This suggests that in RCON models for paired data we have three possible options: no standardization (NS); classical standardization (CS), where we use a rescaling matrix with sample variances $s_{ii}$ on the diagonal; equal variances standardization (EVS), where we compute the m.l.e. of the variances under the assumption that $\sigma_{ii} = \sigma_{i'i'}$, and plug them into the computation of the rescaling matrix.

## 3. Empirical analysis on fMRI brain data

Although the application of graphical lasso to standardised data has become common practice, little is known about the impact of this operation on the result of the analysis, with the relevant exception of (1) where it is shown that the standardisation of the variables may have a strong effect on the result of inference with the graphical lasso. In this section, we consider RCON models for paired data and carry out an empirical analysis to assess the effect of standardisation on the result of the SGL method. We base our analysis on a set of real data so as to assess the impact of this rescaling in a relevant area of application. More specifically, we consider neuroimaging data previously analysed in (5). The data set we consider refer to $n = 404$ observations from $p = 70$ regions of the human brain, spatially paired regions from the left hemisphere ($Y_i$ with $i = 1, \ldots, 35$) and the right hemisphere ($Y_{i'}$ with $i' = 36, \ldots, 70$).

We have shown in Section 2. that in paired data problems it makes sense to standardise the variables only if the assumption of equal variances is satisfied and that, in this case, different types of standardisation are possible. We apply the SGL method to the residuals from a parametric filter on the original time series. These can be assumed to be i.i.d. realizations from a $p-$dimensional multivariate Gaussian distribution with zero mean vector and covariance matrix $\Sigma$. The covariance matrix is unknown and it is reasonable to assume all variables to be measured on the same scale, with equal variances, due to how the signals are recorded, pre-processed and then filter for temporal correlations. In this way, when applied to these data, the three different standardisation methods, i.e. (NS), (CS) and (EVS), should not affect or impact the results obtained by using the SGL method to estimate the concentration matrix from the data.

We have therefore a framework that allows us to quantify the effect of standardisation by comparing the models obtained from the application of the SGL method after the three different standardisation methods have been applied. Furthermore, in order to assess the uncertainty of the estimated effects, a subsampling procedure is implemented, as follows. From the original $404 \times 70$ data matrix we generate $B = 10$ subsamples by randomly sampling the observations into batches of size $\tilde{n} = 100$. For each of these 10 subsamples, as a first step we compute the sample covariance matrix S. Secondly, we rescale S via one of the three standardisation methods. In each subsample $b = 1, \ldots, B$, for each of the three options we fit the SGL model: the output of the algorithm is an estimated concentration matrix $\hat{\Theta}$ and the associated graph $\hat{G}$ and equalities in the concentration values.

We inspect the obtained graphs to understand the effect of the three approaches for standardization. In particular, we summarize how much these graphs are similar to each other, for all possible pairs of comparisons: $\{\hat{G}_{\text{NS}}, \hat{G}_{\text{CS}}\}$, $\{\hat{G}_{\text{NS}}, \hat{G}_{\text{EVS}}\}$, and $\{\hat{G}_{\text{CS}}, \hat{G}_{\text{EVS}}\}$. To quantify the similarity we use two different metrics: the F1 score, that is the harmonic mean of precision (fraction of true positives among predicted positives) and recall (fraction of true positives over the actual positives); and the Matthews Correlation Coefficient (MCC), which is the analogous of a correlation coefficient for binary classifiers. The F1 score ranges from 0 to 1 and higher values are associated with better matching between the two objects; the MCC varies from -1 to +1, with larger values (irrespective of the sign) characterizing situations where the two graphs agree on the structure identified. The two metrics are computed twice: (i) in terms of edge structure, that is considering how many edges are common to the two solutions compared; (ii) in terms of symmetric concentrations, which refers to the common equality constraints identified by the two fitted models being compared. We report a summary of the results for the analysed data in Table 1.

Considering first the edge structure, the average value of F1 is higher when comparing CS and EVS than the average values computed by juxtaposing each of the two options for standardization to NS. This means that the effect itself of rescaling - either assuming equal variances for all variables

DRAFT

Table 1: Average (*standard deviation*) F1 scores, in percentage, and MCCs across the 10 random subsamples; `NS`: no standardization; `CS`: classical standardization; `EVS`: equal variances standardization.

| Compared methods | Edge structure | | Symm. concentrations | |
|---|---|---|---|---|
| | F1 score | MCC | F1 score | MCC |
| NS - CS | 54.05 (*2.73*) | +0.51 (*0.05*) | 62.51 (*3.58*) | +0.62 (*0.04*) |
| NS - EVS | 51.52 (*5.34*) | +0.49 (*0.04*) | 39.82 (*20.29*) | +0.49 (*0.12*) |
| CS - EVS | 70.87 (*7.72*) | +0.67 (*0.11*) | 56.70 (*27.82*) | +0.66 (*0.17*) |

(`CS`) or assuming paired equalities (`EVS`) - produces two graphs similar to one another, and both these graphs appears to be less comparable with the one obtained without applying any rescaling at all. The insight is corroborated also by the computed average MCC values. A similar argument can be made with respect to the symmetries on the concentration matrices obtained by the SGL in each of the three scaling options: all average value of F1 are below 70%, with the similarity measure showing high variability when considering the comparisons that involve the EVS option. Some evidence of differences in the symmetries identified by the SGL for the three scaling methods is found also in the computed average MCC values.

We want to highlight that, although we expect data in this specific application to be observed on the same scale, the impact of the standardization is relevant on the recovered graphs, suggesting that either the assumption of equal variances is not met or the chosen option for rescaling actually matters. Moreover, even the comparison with the highest similarity (`CS` and `EVS`, with an average F1 score of $70.87\%$) points out there is still some effect of the choice of standardization method on the symmetric graphical lasso procedure, which deserves further investigation.

DRAFT

## 4. Discussion

In this short manuscript we tackled the often overlooked issue of rescaling data as a preliminary step to applying a graphical lasso procedure and, in particular, the SGL algorithm. We used fMRI data as an example of a setting where it is reasonable to assume variables to be measured on the same scale. We quantified the (dis)similarity of the recovered graphs, in terms of edge structure and symmetric concentrations, obtained by employing each rescaling option available and we observed a lack of robustness of the estimates to the approach adopted for standardizing.

We consider the problem of rescaling worthy of further considerations, and a potential venue for future research in order to understand the plausibility of equal variances assumption, and what is the effect of the rescaling when that assumption is not valid.

## References

[1] Carter, J.S., Rossell, D., Smith, J.Q.: Partial Correlation Graphical LASSO. arXiv preprint (2021) doi:10.48550/arXiv.2104.10099

[2] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press (2015)

[3] Højsgaard, S., Lauritzen, S.L.: Graphical Gaussian models with edge and vertex symmetries. J. Roy. Stat. Soc. B. **70**, 1005–27 (2008)

[4] Lauritzen, S.L.: Graphical Models. CRC Press (2015)

[5] Ranciati, S., Roverato, A., Luati, A.: Fused graphical lasso for brain networks with symmetries. J. Roy. Stat. Soc. C. **70**, 1299–1322 (2021)

[6] Roverato, A., Nguyen, D.N.: Model inclusion lattice of coloured Gaussian graphical models for

paired data. In: Salmerón, A., Rumí, R. (eds.) Proceedings of the 11th International Conference on Probabilistic Graphical Models, pp. 133-144. PMLR (2002)
[7] Tsai, K., Koyejo, O., Kolar, M.: Joint Gaussian graphical model estimation: A survey. Wires. Comput. Stat. **14**, e1582 (2022)
[8] Xie, Y., Liu, Y., Valdar, W.: Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. Biometrika. **103**, 493–511 (2016)
[9] Zhang, H.., Huang, X., Arshad, H.: Comparing Dependent Undirected Gaussian Networks. Bayesian. Anal. **1**, 1–26 (2022)

DRAFT

# Airports performances and sustainable practices. An empirical study on Italian data

Riccardo Gianluigi Serio[a], Maria Michela Dickson[a], Diego Giuliani[a], and Giuseppe Espa[a]

[a]Universita' degli studi di Trento; riccardo.serio@unitn.it, mariamichela.dickson@unitn.it, diego.giuliani@unitn.it, giuseppe.espa@unitn.it

**Abstract**

The transition to more environmentally sustainable production processes and managerial practices is an increasingly important topic. Many industries need to undergo radical change to meet environmental sustainability requirements; the tourism industry is no exception. In this respect, a particular aspect that needs further attention is the relationship between airport performances and investments in environmental sustainability policies. This work represents a first attempt to provide empirical evidences about this relationship. Through the application of a non-parametrical method, we first assess the efficiency of the Italian airports' industry. Secondly, we investigated the relationship between airports' performance and management commitment toward the ecological transition using a Tobit regression model. The results show that airports' adherence to formal multi-year ecological transition programs has a positive and consistent impact on their performance.

***Keywords:*** Environmental sustainability, DEA, Air transport sector, Tobit Regression.

DRAFT

## 1. Introduction

Today, the airport network of a country is considered a strategic asset for governments. Moreover, airports are the main engine of the tourist sector and several studies have shown their significant contribution to the economic development of the regions that host them (5; 6). The economic boost is due to the airport activities, which manifest themselves by reducing unemployment, increasing income per capita, enhancing productivity, favoring greater investment and trade as well as greater social and cultural development (18; 22; 24). Moreover, the performances of airports are a topic of interest for a vast array of stakeholders, including airlines, governments, passengers, and the residents of the served areas (9; 22). Another topic of growing importance is sustainability, intended as the societal goal to reduce human impact on the environment. In particular, the attention to transition management (i.e., the set of processes through which certain aspects of society change significantly over a short time horizon) to push sustainable growth is increasing sharply (21). In the airport sector, Airport Carbon Accreditation (ACA) is currently the only globally institutionally recognized certification for reducing the carbon footprint. This measure was launched in 2008 by Airport Council International (ACI), through 6 certification steps: "Mapping", "Reduction", "Optimization", "Neutrality", "Transformation" and "Transition". A fundamental prerequisite for accessing this program is the accomplishment, by an accredited institution, of compliance with ISO14064 (Greenhouse Gas Accounting). By reaching the latest level of accreditation (Transition), the airport proves to be in line with the 2015 Paris Agreement, that is, to actively

contribute to limiting the global average temperature rise to 1.5 °C and no more than 2 °C compared to preindustrial levels. Several works have been proposed to evaluate the performances of the airports (8). However, few have been interested in the environmental performances (16), while the relationship between efficiency and environmental sustainability is still an unexplored topic. The aim of this work is to study the relationship between the performance in terms of efficiency of airports and the investments towards the adoption of more sustainable practices. To achieve this goal, a two-step analysis was conducted on a dataset of Italian airports: first, we estimated the efficiency frontier through the Data Envelopment Analysis (DEA) approach. Subsequently, through a Tobit regression model, we investigated whether a relationship between efficiency and environmental sustainability exists. This was possible through the creation of a proxy variable that could assess which airports are further ahead in investments aimed at sustainability.

## 2. Literature review

Firms' efficiency, as well as the efficiency of the production process, are concepts widely studied in the transport sector, the agro-food sector, the large-scale retail trade, telecommunications, and the banking sector. The literature of benchmarking is mainly divided in two streams. One, focused on the use of parametric models, such as the stochastic frontier analysis (e.g., (1; 27)), where it is necessary to establish a priori a functional form for the relationship between input, output, and inefficiency. Notwithstanding, another research flow is involved in the adoption of nonparametric models for the study of the frontier, such as Data Envelopment Analysis (DEA) modeling (e.g., (4; 8; 19; 20)). DEA is a method that arises from the seminal work of (15) and (17), who were the first to try to measure the efficiency of a sample of production units, the so-called decision-making units (DMUs). This method, through a linear programming procedure, provides an efficiency score for each DMU with a limited number of necessary assumptions. It is based on an optimization function that defines weights to be attributed to the combination of inputs and outputs of each DMU such as to maximize the outputs (setting the level of inputs) or minimize the inputs satisfying at least a given level of output (hereinafter output-oriented and input-oriented DEA models). Furthermore, the DEA modeling framework can be divided according to the assumptions on the returns to scale. The model assuming constant returns to scale was developed by (12), while (7) added an assumption about the concavity of the frontier, allowing for variable returns to scale. The main difference between the two models lies in the definition of efficiency; in the Charnes model (henceforth CCR-I), the computed efficiency score is the overall efficiency of the DMU (i.e., OTE), taking into account both technical efficiency (i.e., the ability to produce a desired output using the minimum possible input) and allocative efficiency (i.e., the ability to use inputs optimally, choosing the most efficient combination of inputs). In the Banker's model (BCC-I), the score measures only the technical efficiency of the production unit, ignoring allocative efficiency (i.e., PTE). The DEA has been widely employed in the investigation of airport efficiency since the seminal work of (19). Some researchers, for example, (26), and (2), studied the efficiency of airports through an input oriented DEA (DEA-I), arguing that the main output, passenger traffic, was a phenomenon beyond managerial control, and therefore a difficult to maneuver lever. Others (e.g., (11; 19; 23)) have adopted an output-oriented DEA (DEA-O), assuming that most of the equipment for airport operation has the nature of fixed assets, therefore beyond the control of management (at least in the short run).

In the literature about the air transport sector, DEA models have been used in a two-step analysis, in the first step, the DEA is used to estimate the efficiency frontier, while in the second step regression models are used to explain the efficiency itself. A useful tool for estimating a linear relationship when the dependent variable is simultaneously censored on the left or on the right is the Tobit regression model (28). In practice, Tobin's model modifies the likelihood function in order to take into account the non-equiprobability in sampling for each observation depending on whether the latent dependent variable has fallen above or below the threshold determined by the censorship. In this way, it is possible to further discriminate the phenomena that influence the performances of the airports. Concerning the variables to be included in the models, several evidences have been highlighted in literature, such as for example the ownership of the airports (25), their sizes (11), the internationality of airport traffic (13) the location of the

DRAFT

airports (29) and their belonging to a group (3). However, the relationship between airport performance and environmental sustainability seems to remain particularly unexplored.

## 3. Data description

The dataset used in the present study reports information on almost all Italian airports (30 units) in 2019. The sample covers over 99% of all passengers transiting through Italian airports in the year, so it reports information very close to that of the real population. The data were collected merging various sources: economic and financial information from the AIDA platform, data on passengers, movements of aircraft and goods from the Italian Statistical Institute (ISTAT) archives; data on the runways, internal equipment, and size from the National Air Transport Authority (ENAC); and, finally, some variables have been constructed by the authors, as explained in the following. We computed both the variable returns to scale input-oriented model BCC-I (7), and the constant returns to scale input-oriented model, CCR-I (12) with 4 inputs (EMPLOYEES, CHINDESKS, RUNWAYMT, PRODCOSTS) and 4 outputs (TOTPAX, GOODS, TOTPLANES and TOTREVENUES). To ensure the discriminatory power of the DEA in this analysis, we followed the criteria suggested in the literature on the relationship between the number of DMUs, and inputs and outputs (e.g., (14)). Our model respects both the $N_{DMU_s} \geq 3 \times (inputs + outputs)$ rule, and the $N_{DMU_s} \geq (inputs \times outputs)$ rule. The variable EMPLOYEES contains the total number of employees, CHINDESKS reports the number of check-in desks available in the terminal, RUNWAYMT the runway meters, PRODCOSTS the total production costs (thousands / EUR), TOTPAX (%), GOODS (%), TOTPLANES (%), contain information on the total number of airports' passengers, goods and airplanes managed as percentage of the total in Italy in the year 2019, TOTREVENUES indicates revenues totals (thousands/EUR) for the year studied, and SURFACE describes the total surface of the airport $(m^2)$.

## 4. Results

In table 1 we report the results of the DEA. The DEA models suggest that 6 out of 30 airports are globally efficient (OTE = 1), while 12 out of 30 reach the purely technical efficiency frontier (PTE = 1). The average efficiency of Italian airports is relatively high. Indeed, the average efficiency exceeds 79% considering the OTE and is even higher for the PTE (87.9%). Following these results, we can see how Italian airports are well managed on average. By comparing the scores arising from the 2 models, we can discriminate between the relative inefficiency of the DMUs due to the management of operations (PTE) and the inefficiency due to its scale (SE). 6 out of 30 airports exhibit constant returns to scale and are efficient, suggesting that they are in their optimal production condition. However, the remaining DMUs operate under increasing returns, indicating that many airports could experience a more than proportional increase in performance from the increase in their production size. In order to investigate the relationship between performance and environmental sustainability, we constructed a variable able to capture the commitment toward the ecological transition shown by the management. Exploiting the 6 ACA certification steps, the variable SUSTAINABILITY was designed to assume values between 0 and 7. Level 0 includes all airports that do not currently exhibit any commitment towards sustainability that is higher than the duties established by Italian law. Level 1 includes all the companies that have publicly declared (through the website) a real commitment (e.g., reclamation investments in the area surrounding the airport, efficient water management, installation of solar panels), but which have not been admitted to in the ACA program. Scores from 2 to 7 are attributed to all companies that adhere to the ACA program based on the certification level reached by the airport (e.g., score 2 = "mapping", score 3 = "reduction", score 3 = "optimization", etc).

In table 2 we exhibit the results for the Tobit regressions (we report the marginal effects). The management's adoption of policies aimed at environmental sustainability has a positive impact on the performance of the airport. Further, the coefficient shows a positive sign, and the likelihood ratio test (LT) confirms the significance of the estimate, rejecting the null hypothesis that the SUSTAINABILITY

DRAFT

Table 1: Application of DEA CCR-I and BCC-I to Italian airports; this table reports the estimated overall efficiency under constant returns ($OTE_{crs}$), the pure technical efficiency under variable returns ($PTE_{vrs}$), the scale efficiency (SE) and the estimated returns to scale (RTS).

| N. | Airports | $OTE_{crs}$ | $PTE_{vrs}$ | SE | RTS |
|---|---|---|---|---|---|
| 1 | Bergamo-Orio Al Serio | 1.00 | 1.00 | 1.00 | Constant |
| 2 | Catania-Fontanarossa | 1.00 | 1.00 | 1.00 | Constant |
| 3 | Milano-Linate-Malpensa | 1.00 | 1.00 | 1.00 | Constant |
| 4 | Napoli-Capodichino | 1.00 | 1.00 | 1.00 | Constant |
| 5 | Roma-Ciampino-Fiumicino | 1.00 | 1.00 | 1.00 | Constant |
| 6 | Venezia-Tessera | 1.00 | 1.00 | 1.00 | Constant |
| 7 | Bologna-Borgo Panigale | 0.99 | 1.00 | 0.99 | Increasing |
| 8 | Lampedusa | 0.83 | 1.00 | 0.83 | Increasing |
| 9 | Perugia | 0.72 | 1.00 | 0.72 | Increasing |
| 10 | Grosseto | 0.65 | 1.00 | 0.65 | Increasing |
| 11 | Elba | 0.64 | 1.00 | 0.64 | Increasing |
| 12 | Bolzano | 0.50 | 1.00 | 0.50 | Increasing |
| 13 | Olbia-Costa Smeralda | 0.99 | 0.99 | 0.99 | Increasing |
| 14 | Palermo-Punta Raisi | 0.94 | 0.96 | 0.99 | Increasing |
| 15 | Genova-Sestri | 0.89 | 0.94 | 0.95 | Increasing |
| 16 | Treviso-Sant'Angelo | 0.86 | 0.93 | 0.92 | Increasing |
| 17 | LameziaTerme-ReggioCalabria-Crotone | 0.90 | 0.92 | 0.98 | Increasing |
| 18 | Firenze-Pisa | 0.91 | 0.91 | 1.00 | Increasing |
| 19 | Verona-Brescia | 0.86 | 0.86 | 0.99 | Increasing |
| 20 | Trieste-Ronchi dei Legionari | 0.79 | 0.82 | 0.96 | Increasing |
| 21 | Torino-Caselle | 0.82 | 0.82 | 0.99 | Increasing |
| 22 | Rimini-Miramare | 0.74 | 0.82 | 0.90 | Increasing |
| 23 | Cagliari-Elmas | 0.78 | 0.78 | 0.99 | Increasing |
| 24 | Bari-Brindisi-Foggia-Taranto | 0.78 | 0.78 | 0.99 | Increasing |
| 25 | Alghero-Fertilia | 0.74 | 0.76 | 0.97 | Increasing |
| 26 | Pescara | 0.68 | 0.76 | 0.89 | Increasing |
| 27 | Cuneo-Levaldigi | 0.50 | 0.66 | 0.75 | Increasing |
| 28 | Trapani-Birgi | 0.48 | 0.58 | 0.83 | Increasing |
| 29 | Ancona-Falconara | 0.52 | 0.57 | 0.90 | Increasing |
| 30 | Parma | 0.23 | 0.48 | 0.48 | Increasing |

DRAFT

effect on the efficiency scores is equal to 0. This is an important result and, to our knowledge, the first in identifying a clear impact of sustainable choices on airport performance. Both DEA models adopted for the study are input-oriented. The positive effect of an airport's admission to the ACA program could suggest that the improved efficiency is linked to an internal efficiency process (e.g., mapping excess CO2 emissions). If we consider the impact of sustainability on purely technical efficiency (BCC-I), the effect seems to be amplified, confirming an improvement in purely operational management. The EBITDA margin has a positive, statistically significant coefficient and with a considerable magnitude. This implies that the Italian airports, which transform a notable part of the sales volume into profits, are more efficient. The variables LCC and GROUP are not significant, suggesting that the airlines business model does not affect airports' efficiency and that independent airports or airports belonging to a group have, on average, similar performances. The same could be said for OWNERSHIP. Our results suggest that there is no evidence from data that private owned airports outperform public ones or vice-versa. The coefficient of LOGAREAPAX is significant but with the opposite sign considering the overall efficiency (which seems to benefit from more spaces dedicated to passengers), and purely technical efficiency. This may suggest that space management is crucial for the airport. Increasing the available space can lead to

Table 2: Regression results on CCR-I and BCC-I efficiency scores (standard errors in parentheses).

| Indep. variable | Dep. variable | |
|---|---|---|
| | OTE (CCR-I) | PTE (BCC-I) |
| SUSTAINABILITY | $0.034^{**}$ (0.015) | $0.059^{***}$ (0.019) |
| EBITDA | $0.268^{***}$ (0.033) | $0.245^{***}$ (0.039) |
| LCC | 0.027 (0.076) | -0.157 (0.100) |
| OWNERSHIP | 0.027 (0.040) | -0.018 (0.044) |
| GROUP | 0.006 (0.052) | 0.037 (0.061) |
| LOGAREAPAX | $0.040^{**}$ (0.019) | $-0.086^{***}$ (0.031) |
| Constant | $0.331^{**}$ (0.154) | $1.668^{***}$ (0.258) |
| Observations | 30 | 30 |
| Log Likelihood | 18.710 | 9.025 |
| Wald Test (df = 6) | $216.000^{***}$ | $120.200^{***}$ |
| Pseudo-$R^2$ | 0.779 | 0.668 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

the installation of more sale points and attractions for the passengers, and this can lead an airport to an overall efficiency resulting from the generated higher revenues. However, more space can also lead to potential threats in the operational management of resources.

## 5. Concluding Remarks

The present work constitutes an attempt to understand the relationship between airports efficiency and environmental sustainability. Specifically, the effect of airports joining programs aimed at reducing environmental impacts on overall efficiency and pure technical efficiency has been deepened through a DEA model and a Tobit regression analysis. Our results suggest that there is a statistically significant and positive association between Italian airport efficiency and environmental sustainability. In fact, according to our model, airport efficiency increases as management's commitment to sustainable policies grows. In particular, we assessed sustainability by discriminating between airports that do not show any declared commitment to ecological transition practices exceeding the obligations imposed by national law, and airports that adhere to the ACA program. This constitutes an important result, as it demonstrates that the time is now ripe for managers to include environmental performance improvement policies in their strategic choices, particularly in the airport sector.

DRAFT

## References


[1] Abrate, G., Erbetta, F. (2010). Efficiency and patterns of service mix in airport companies: An input distance function approach. TRANSPORT RES E-LOG, 46(5), 693-708.

[2] Adler, N. and Berechman, J. (2001). Measuring airport quality from the airlinesźviewpoint: an application of data envelopment analysis. Transp. Policy, 8(3):171-18.

[3] Adler, N., Liebert, V., and Yazhemsky, E. (2013a). Benchmarking airports from a managerial perspective. Omega, 41(2):442-45.

[4] Adler, N., Ulku, T., and Yazhemsky, E. (2013b). Small regional airport sustainability: Lessons from benchmarking. J. Air Transp. Manag., 33:22-3.

[5] ATAG (2008). Aviation benefits beyond borders.

[6] ATAG (2020). The economic and social benefits of air transport.

[7] Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and

scale inefficiencies in data envelopment analysis. Manage Sci., 30(9):1078-1092.
[8] Barros, C. P. and Dieke, P. U. (2007). Performance evaluation of italian airports: A data envelopment analysis. J. Air Transp. Manag., 13(4):184-19.
[9] Barros, C. P. and Peypoch, N. (2008). A comparative analysis of productivity change in italian and portuguese airports. Int. J. Transp. Econ., 35(2):205-216.
[10] Bell, R. A. and Morey, R. C. (1995). Increasing the efficiency of corporate travel management through macro benchmarking. J. Travel Res., 33(3):11-20.
[11] Carlucci, F., Cir'a, A., and Coccorese, P. (2018). Measuring and explaining airport efficiency and sustainability: Evidence from italy. SUSTDE., 10(2):400.
[12] Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. EUR J OPER RES, 2(6):429-444.
[13] Chow, C. K. W. and Fung, M. K. Y. (2009). Efficiencies and scope economies of chinese airports in moving passengers and cargo. J. Air Transp. Manag., 15(6):324-329.
[14] Cooper, W. W., Seiford, L. M., and Tone, K. (2007). Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software, volume 2. Springer.
[15] Debreu, G. (1951). The coefficient of resource utilization. Econometrica, 19(3):27
[16] Dimitriou, D., Voskaki, A., and Sartzetaki, M. (2014). Airports environmental management: Results from the evaluation of european airports environmental plans. Int. J. Inf. Syst. Supply Chain Manag., 7(1):1-14.
[17] Farrell, M. J. (1957). The measurement of productive efficiency. J R Stat Soc Ser A Stat Soc., 120(3):253.
[18] Gibbons, S. and Wu, W. (2020). Airports, access and local economic performance: evidence from china. J. Econ. Geogr., 20(4):903-937.
[19] Gillen, D. and Lall, A. (1997). Developing measures of airport productivity and performance: an application of data envelopment analysis. TRANSPORT RES E-LOG, 33(4):261-273.
[20] Gitto, S. and Mancuso, P. (2012). Two faces of airport business: A non-parametric analysis of the italian airport industry. J. Air Transp. Manag., 20:39-42.
[21] Gössling, S., Hall, C. M., Ekstr öm, F., Engeset, A. B., and Aall, C. (2012). Transition management: a tool for implementing sustainable tourism scenarios? J. Sustain. Tour., 20(6):899-916.
[22] Graham, A. (2008). Managing airports: An international perspective (3rd ed.). Elsevier.
[23] Martín, J. C. and Roman, C. (2001). An application of DEA to measure the efficiency of spanish airports prior to privatization. J. Air Transp. Manag., 7(3):149-157.
[24] Maughan, J., Raper, D., Thomas, C., and Gillingwater, D. (2001). Scan-uk the uk sustainable cities and aviation network. Air & Space Europe, 3(1-2):56-59.
[25] Oum, T. H., Yan, J., and Yu, C. (2008). Ownership forms matter for airport efficiency: A stochastic frontier investigation of worldwide airports. J Urban Econ., 64(2):422-435.
[26] Pels, E., Nijkamp, P., and Rietveld, P. (2003). Inefficiencies and scale economies of european airport operations. TRANSPORT RES E-LOG, 39(5):341-361.
[27] Scotti, D., Malighetti, P., Martini, G., and Volta, N. (2012). The impact of airport competition on technical efficiency: A stochastic frontier analysis applied to italian airport. J. Air Transp. Manag., 22:9-15.
[28] Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society, pages 24-36.
[29] Yoshida, Y. and Fujimoto, H. (2004). Japanese-airport benchmarking with the dea and endogenous weight tfp methods: testing the criticism of overinvestment in japanese regional airports. TRANSPORT RES E-LOG, 40(6):533-546.

DRAFT

# Quasi-experimental evidence on COVID-19 lockdown effects on Italian household food shopping basket composition and its sustainability

Beatrice Biondi[a] and Mario Mazzocchi[a]

[a]University of Bologna - Dept. of Statistical Sciences;

b.biondi@unibo.it; m.mazzocchi@unibo.it;

**Abstract**

Movement restrictions imposed by governments in response to the COVID-19 pandemic have helped controlling the spread of the disease, but they have also impacted lifestyles, such as dietary habits and food choices. In this study, we explore the effect of the first lockdown period and its aftermath on food and drink purchases, with a particular focus on the overall changes in healthiness and sustainability of dietary choices. We estimate a panel difference-in-differences model on weekly purchases from household scanner data, over a period of two years. We found a large increase in consumption of unhealthy, comfort food and drinks such as ice creams, sweet spreads and beer. Despite the relative change in composition of the food basket, the associated greenhouse gas emission levels are substantially unchanged.

***Keywords:*** Household scanner data, Panel difference-in-differences, COVID-19, food consumption

DRAFT

## 1. Introduction

The spread of COVID-19 disease has impacted people lives all over the world in many different ways. During forced lockdowns, the containment measures in place (e.g. stay at home restrictions, schools and workplaces closure and restaurants closure) changed the lifestyle of an entire population. Among other behaviours, dietary habits and food and drink consumption had to adjust to the new situation. The change in food consumption behaviours is mainly due to substitution for out-of-home consumption, stockpiling and hoarding behaviour, increased time availability for meal preparation, and psychological coping mechanisms potentially leading to an increase in comfort food consumption (1; 3).

The aim of the present study is therefore to explore how restrictions against COVID-19 spread affected food and drink at-home consumption, focusing on the first lockdown and the first post-pandemic unrestricted period. We consider the quantity consumed of some unhealthy foods and placebo goods, and analyse the overall (food basket) greenhouse gas emission (GHGEs) levels, and how they changed with respect to the pre-pandemic status quo.

# 2. Methods

## 2.1 Data

Our data consist in household-scanner recorded purchases of all food and drink purchases made by a representative sample of households in Italy over a period of two years, 2019 and 2020, provided by The Nielsen Company (Italy). The sample includes more than nine thousands unique households; each household in the Nielsen panel records all food and drink purchases brought home through an hand-held scanner and answers a questionnaire about socio-demographic characteristics once a year.
The dataset contains aggregated weekly purchases for all food and drink product categories. Each row in the dataset contains information about amount (in kg or litres), expenditure in Euros, number of items purchased. Prices were obtained by averaging household unit values (i.e. ratios between expenditures and volumes) by region and week, under the standard assumption that consumers within the same area and in the same time period face the same prices.

Data on food and drink purchased quantities were translated into GHG emissions based on conversion factors as provided in (2): they account for all the production phases, from field to farm gate, transport, processing, packaging, storage and supermarket operations. We first aggregated the 66 product-specific emission factors over our ECOICOP classification (19 food categories), and then calculate the total kg of carbon dioxide equivalent ($CO_2e$) GHGEs from food and drink purchases, per household per week.

## 2.2 Model

We compare food purchases in different periods of 2019 and 2020 by means of a Difference-in-Differences panel regression. The baseline period considers the pre-pandemic months (essentially January and mid-February), the lockdown period goes from March to mid-May, and the post-lockdown considers the period from mid-May until the end of September. The remaining period of the year is not considered because of its smaller informative power – being far away from the lockdown – and because of confounding effects related to new types of restrictions enacted.

Let's consider a generic output $Y$, and observations on the purchases of $n$-th household in week $t$:

$$Y_{nt} = \alpha_n + \mathbf{X}\beta + \gamma Year_{2020} + \delta_1 D_1 + \delta_2 D_2 + \zeta_1 D_1 Year_{2020} + \zeta_2 D_2 Year_{2020} + \varepsilon_{nt} \qquad (1)$$

where the vector $\mathbf{X}$ includes possible controls, $Year_{2020}$ is a dummy variable that equals one in the year 2020, $D_1$ is the dummy for the lockdown period, and $D_2$ is the dummy for the post-lockdown period, $a_n$ are household fixed effects, $\varepsilon_{nt}$ is the random component. We estimate several models, with different dependent variables and covariates:

- A set of $I$ regressions are estimated on disaggregated product categories, and consider *scaled* volumes, $Volume_s_{int}$, for the $i$-th product purchased by the $n$-th household in week $t$ [1]. The vector $\mathbf{X}$ includes the average regional weekly price for individual products, $P_{irt}$.
- Other models are estimated on aggregated data, for the $n$-th household in week $t$, the dependent variables considered are: $Volume_{nt}$, $Exp_{nt}$, $UV_{nt}$, which include total food quantity, total expenditure and aggregated unit value, respectively.
- The final set of estimated regressions is on emissions: $Emiss_{nt}$ is the weekly household GHGEs from food and drink purchases, and $Emiss_kg_{nt}$ considers the emissions per kg of purchased food. Here, we control for household-specific weekly aggregated unit values, $UV_{nt}$

[1] Since food products are very different in terms of average weekly purchased volumes, and rarely comparable in their unit of measure, we obtain comparable amounts by rescaling the volume variable as follows: $Volume_s_{int} = Volume_{int}/Volume_{i,baseline2019}$, where $i$ indicates the good, $n$ refers to the household and $t$ to the specific week; $Volume_{i,baseline2019}$ equals the average volume for product $i$ in the period between 31 December and 17 February 2019.

DRAFT

We are particularly interested in coefficients $\zeta_1$ and $\zeta_2$ in (1), the DiD coefficients for the lockdown and post-lockdown periods: they measure the differential level of $Y$ in 2020 in each period, i.e. $\zeta_1$ and $\zeta_2$ represent the lockdown and post-lockdown impact on $Y$, respectively.

## 3. Results and Discussion

Table 1 reports estimates of the relative changes in purchased quantities for a selection of food products. Hence, the coefficient reflects the change in purchased quantities with respect to the corresponding period of 2019. Considering the lockdown period, base ingredients (mainly flour, sugar and pastry ingredients) experienced the highest rise in consumption, confirming a shift towards home-made preparation. An increase in comfort food consumption is also observed, namely ice creams, beer, sweet spreads (jams, chocolate spreads and honey). Animal products, like pork meat and cheese, show relatively larger increases than other food groups, such as vegetables and pasta. Fish products purchases did not change during the lockdown period, whereas prepared dishes and festivity foods and wines decreased significantly. In the post-lockdown period, most variations became smaller, but still significantly higher than their baseline. Products like ice creams, for which out-of-home consumption is especially relevant, experience a significant decrease in purchases during the post-lockdown period, when restrictions were lifted.

Then, we translate the changes in purchases of the 19 food and drink categories into aggregate outcomes and their greenhouse gas emissions. Table 2 reports the results: estimates show a significant change in total purchased volumes during lockdown (more than 5 kg of food per household per week) and weekly household expenditure (around €12 per household per week). On average, there was a shift towards cheaper products, as the average unit value fell by €0.4 per kg. In terms of emissions, the larger amount of purchases during the lockdown period resulted in larger total emissions (nearly 40% more than the average emission levels in 2019). However, when controlling for the increased purchased volumes, no significant change in emission levels is observed. During the post-lockdown period, food purchases and expenditure remained slightly higher than their counterfactual level. Again, the impact in terms of GHGEs is merely related to the higher volumes, and emissions per kg of purchase foods were basically unchanged.

DRAFT

Table 1: Effects of lockdown and post-lockdown periods on purchases of selected food (relative to 2019 baseline period)

| | $\zeta_1$ | $\zeta_2$ |
|---|---|---|
| Base ingredients | $1.46^{**}$ | $0.24^{**}$ |
| | $(0.02)$ | $(0.02)$ |
| Icecreams | $1.10^{**}$ | $-0.21^{*}$ |
| | $(0.12)$ | $(0.11)$ |
| Beer | $0.54^{**}$ | $0.21^{**}$ |
| | $(0.04)$ | $(0.04)$ |
| Cheese | $0.51^{**}$ | $0.12^{**}$ |
| | $(0.01)$ | $(0.01)$ |
| Pork | $0.51^{**}$ | $0.25^{**}$ |
| | $(0.07)$ | $(0.06)$ |
| Sweet spreads | $0.50^{**}$ | $0.14^{**}$ |
| | $(0.02)$ | $(0.02)$ |
| Vegetables | $0.38^{**}$ | $0.05^{**}$ |
| | $(0.02)$ | $(0.01)$ |
| Pasta | $0.27^{**}$ | $0.12^{**}$ |
| | $(0.02)$ | $(0.02)$ |
| Water | $0.25^{**}$ | $0.05^{**}$ |
| | $(0.02)$ | $(0.02)$ |
| Fresh packaged fish | $-0.02$ | $0.38^{**}$ |
| | $(0.18)$ | $(0.17)$ |
| Prepared dish | $-0.05^{**}$ | $-0.01$ |
| | $(0.02)$ | $(0.02)$ |
| Festivity products | $-0.21^{**}$ | $-0.07^{**}$ |
| | $(0.04)$ | $(0.04)$ |
| Champagne and sparkling wine | $-0.24^{**}$ | $0.07$ |
| | $(0.08)$ | $(0.07)$ |

*Notes:* Estimates from panel model (1); Standard errors in brackets. Asterisks refer to estimates' significance at 0.01 (**) and 0.05 (*) level

DRAFT

Table 2: Effects of lockdown and post-lockdown periods on food baskets and household emissions

| | $\zeta_1$ | $\zeta_2$ |
|---|---|---|
| Total food quantity purchased (kg food hh$^{-1}$ week$^{-1}$) | $5.45^{**}$ | $1.41^{**}$ |
| | $(0.26)$ | $(0.12)$ |
| Total expenditure (€ hh$^{-1}$ week$^{-1}$) | $11.92^{**}$ | $1.91^{**}$ |
| | $(0.72)$ | $(0.25)$ |
| Aggregated unit value (€ kg food$^{-1}$) | $-0.41^{**}$ | $-0.14$ |
| | $(0.07)$ | $(0.09)$ |
| Total GHGEs (kg $CO_2$e hh$^{-1}$ week$^{-1}$) | $13.32^{**}$ | $3.59^{**}$ |
| | $(0.57)$ | $(0.21)$ |
| GHGEs per kg of purchased food (kg $CO_2$e kg food$^{-1}$) | $-0.01$ | $0.02^{**}$ |
| | $(0.01)$ | $(0.01)$ |

*Notes:* Estimates from panel model (1); Clustered standard errors in brackets. Asterisks refer to estimates' significance at 0.01 (**) and 0.05 (*) level

## 4. Conclusion

We exploit the detailed information available from Nielsen household scanner data to estimate the impact of COVID-19 restrictions on food purchases and their greenhouse gas emissions using a Difference-in-Difference model.

The lockdown period is associated, as expected, with higher amounts of purchased foods for home consumption, hence higher emissions. However, we find no evidence that the relative sustainability of food purchases has worsened. Similar findings emerge from the post-lockdown period, but to a much lower extent.

## References


[1] Biondi, B., Capacci, S., Mazzocchi, M.: Food purchasing behavior during the COVID-19 pandemic: Evidence from Italian household scanner data. Quaderni di Dipartimento, Serie Ricerche, pp. 44 (2021) ISSN 1973-9346.

[2] Hoolohan, C., Berners-Lee, M., McKinstry-West, J., Hewitt, C.N.: Mitigating the greenhouse gas emissions embodied in food through realistic consumer choices. Energy Policy. **63**, 1065-1074 (2013) doi: 10.1016/j.enpol.2013.09.046.

[3] Salazar-Fernàndez, C., Palet, D., Haeger, P.A., Romàn Mella, F.: The perceived impact of COVID-19 on comfort food consumption over time: The mediational role of emotional distress. Nutrients. **13**, 1910 (2021) doi: 10.3390/nu13061910.


DRAFT

# Path analysis in Ising models: an application to cyber-security risk assessment

Monia Lupparelli and Giovanni M. Marchetti

Department of Statistics, Computer Science, Applications "G. Parenti"
University of Florence, viale Morgagni 59, 50133 Firenze, Italy;
monia.lupparelli@unifi.it, giovanni.marchetti@unifi.it

**Abstract**

We propose a method for path analysis in undirected graph models for binary variables to quantify the strength of association in multiple paths joining a pair of vertices in the underlying graph. With special focus on Ising models, we provide a decomposition of a marginal pairwise parameter into the sum of measures uniquely associated to the edge set of the paths linking the related couple of vertices. The work has been stimulated by a collaboration with a consultant company working in cyber-security risk assessment in industrial Operational Technology systems.

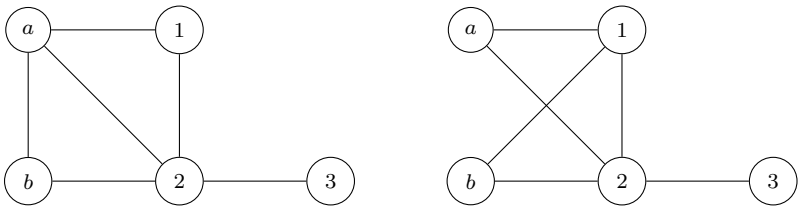
***Keywords:*** Industrial network, Odds-ratio, Risk mitigation, Undirected graph.

DRAFT

## 1. Introduction

Ising models, introduced by Ernst Ising in 1925 to represent the joint distribution of magnetic solid materials in a lattice, are increasingly used to study independence relationships between atomic variables which can assume "active" or "not active" status (Ising , 1925) . Variables are represented by the vertices of an undirected graph where a missing edge between two vertices corresponds to a conditional independence assumption for the related pair of variables. Ising models represent a special exception of hierarchical log-linear models with zero interaction parameters of order higher than two, and a set of odds-ratio parameters measure the association detected by the edges of the underlying graph. In ferromagnetism, modelling the transition of the variable status is a relevant aspect of interest and, nowadays, the same principle is borrowed to explore the relationships in a set of homologous binary variables which may show a synchronized behaviour, as symptoms or risk factors; see Kruis and Maris (2013) and Marsman et al. (2018). Recently, Ising models have been also employed to study the propagation of risk events in catastrophic systems and, in particular, in operational technology (OT) systems where studying and monitoring the spread of cyber-attacks is an aspect of primary interest (Denuit and Robert , 2022).

The set of paths joining a pair of vertices provides a full picture of the dependence between the related couple of variables and our opinion is that studying the relationships across multiple paths may provide insight to quantify the propagation of the variable status in Ising models. Despite path analysis has been introduced and is traditionally employed to study causal relationships in directed acyclic graphs (Wright , 1921), we acknowledge that the analysis of paths can be relevant in multivariate settings regardless the type of graphs and the causal interpretation of the associations. Path analysis in Gaussian undirected graph models has been introduced only more recently by Jones and West (2014) and some years later developed by Roverato and Castelo (2020). These Authors exploit the functional relationship between the variance-covariance and the precision matrices to derive a mapping interpretable in terms of path analysis. Path analysis in discrete undirected graphs seems to be barely studied.




(a) (b)

Figure 1: Ising graphical models for the random vector $X_V = \{X_a, X_b, X_1, X_2, X_3\}$; variables $X_a, X_b$ represent the pair of interest for the analysis of paths.

We propose a decomposition of the pairwise probability related to a couple of vertices into the sum of measures based on products of odds-ratios directly associated to the edge sets of multiple paths. In turns, this decomposition represents an alternative path-based mapping between the mean and the canonical parameters in Ising models. This approach represents a first attempt to study the status propagation in Ising models and in related fields of application.

## 2. The Ising model

Consider a finite set $V$, and let $X_V = \{X_j\}_{j \in V}$ be a vector of random binary variables with levels $i \in \mathcal{I}_V$, where $\mathcal{I}_V = \{0,1\}^{|V|}$ is a $2^{|V|}$ probability table. Let $\pi = (\pi_D)_{D \subseteq V}$ be the probability parameter for the joint distribution of $X_V$, where $\pi_D = P(X_D = 1_D, X_{V \setminus D} = 0_{V \setminus D})$ is the probability of the event associated to the cell $i_D$ of $\mathcal{I}_V$, with $D \subseteq V$. If $X_V$ follows a multivariate Bernoulli distribution $P(X_V; \pi)$, $\log \psi = (\log \psi_D)_{D \subseteq V}$ is the log-linear canonical parameter in the exponential family theory. The mapping between the probability and canonical parameter is based on the Möbius transformation,

$$\log \psi_D = \sum_{D' \subseteq D} (-1)^{|D \setminus D'|} \log \pi_{D'}, \qquad \log \pi_D = \sum_{D' \subseteq D} \log \psi_{D'}, \qquad D \subseteq V. \tag{1}$$

For any $j \in V$, the parameter $\psi_j$ is the odd for the variable $X_j$, for any $j, k \in V$, the parameter $\psi_{jk}$ is the odds-ratio for the variables $X_j, X_k$, both computed in the table $\mathcal{I}_V$. Given the product $\Psi = \prod_{D \subseteq V} \psi_D$, let $\Psi_D$ be any sub-product including only log-linear terms $\psi_E$ with $E \subseteq D$, for any $D \subseteq V$. The mean parameters for the joint distribution of $X_V$ are defined by the vector $\mu = (\mu_D)_{D \subseteq V}$, where $\mu_D = P(X_D = 1_D)$ is the probability of the realization of the event $i_D$ in the marginal table $\mathcal{I}_D = \{0,1\}^{|D|}$, with $D \subseteq V$. For any pair $D, E$ of disjoined subsets of $V$, consider the conditional probability

$$\pi_{D|E} = P(X_D = 1_D, X_{V \setminus D \cup E} = 0_{V \setminus D \cup E} | X_E = 1_E), \qquad D, E \subseteq V,\ D \cap E = \emptyset,$$

that the event $i_D$ realises in the slice of the table $\mathcal{I}_V$ defined by $i_E \in \mathcal{I}_E$.

In graphical models, a random vector $X_V$ is associated to a graph, in particular we consider an undirected graph $\mathcal{G} = (V, \mathcal{E})$, where $V$ is a set of vertices/nodes and $\mathcal{E}$ is a set of pairs of vertices; if the couple $\{j, k\} \in \mathcal{E}$, the vertices $j$ and $k$ of the graph are joined by an undirected edge; see Figure 1. The Ising model is an independence model for the vector $X_V$ over an undirected graph $\mathcal{G} = (V, \mathcal{E})$ (Ising , 1925). A pair of variables associates to vertices not joined by an edge are assumed to be independent given the status of other variables. A formal definition of an Ising model follows.

**Definition 1.** *Given a random vector $X_V$ of binary variables associated to an undirected graph $\mathcal{G} = (V, \mathcal{E})$, the Ising model is the family of Ising probability distributions $P_{\mathcal{G}}(X_V; \psi)$ where $\log(\psi_{jk}) = 0$ for every pair $\{j, k\} \notin \mathcal{E}$. The probability parameters are*

$$\pi_D \propto \left\{ \prod_{j \in D} \psi_j^{x_j} \times \prod_{j,k \in D : \{j,k\} \in \mathcal{E}} \psi_{jk}^{x_j x_k} \right\}, \qquad x_j, x_k \in \{0,1\}, \quad D \subseteq V. \tag{2}$$

DRAFT

An aspect of primary interest in Ising models is studying the transition of the active status through the edge set of the graph and the status propagation between two vertices reasonably runs through the set of paths connecting them. A path $\delta$ is defined by an ordered sequence $(j, k, \ldots, z)$ of distinct vertices, with $j, k, \ldots, z \in V$, such that any couple of adjacent vertices in the sequence is joined by an edge in $\mathcal{G}$. Repetition of vertices is not allowed in this definition of paths. Let $\Delta_{ab}$ be the set of all paths with endpoints $a, b \in V$ where the generic element of the set is a path $\delta = (\delta_1, \ldots, \delta_m)$ with $\{\delta_j, \delta_{j+1}\} \in \mathcal{E}$, for any couple of adjacent vertices in $\delta$. The set of paths joining vertices $a$ and $b$ for the graph in Figure 1(a) is $\{(a, b), (a, 1, 2, b), (a, 2, b)\}$. A path is said to be active when all variables along the path show the active status. The transition of the variable status realises through active paths.

This work aims to provide a decomposition of the probability $\mu_{ab}$ associated to vertices $a, b \in V$ with the intent to give insight on the strength association driving the transition of the variable status across paths. This decomposition is founded on the following lemma (Lupparelli and Marchetti , 2023).

**Lemma 1.** *Let $P_{\mathcal{G}}(X_V; \psi)$ be an Ising model. The marginal probability $\mu_{ab}$ is equivalent to*

$$\mu_{ab} = \frac{\sum_{\delta \in \Delta_{ab}} \pi_{\delta ab}}{\sum_{\delta \in \Delta_{ab}} \pi_{\delta|ab}}, \tag{3}$$

*where $\Delta_{ab}$ is the set of paths joining the pair $a, b \in V$ of vertices.*

Equation (3) does not completely addresses our scope since probability $\pi_{\delta ab}$ is not uniquely related to an active path $\delta \in \Delta_{ab}$, e.g., paths $(a, 1, 2, b)$ and $(a, 2, 1, b)$ in Figure 1(b) have the same probability $\pi_{12ab}$ to be active, despite they consider different sets of edges and different pairwise associations. Also, the probability $\pi_{\delta|ab}$ is a non-trivial function of $\psi$ used to model the presence of edges in Ising models.

## 3. Logistic regression models for events in collapsed tables

A logistic regression approach is proposed to parametrise the probability $\pi_{\delta|ab}$ of a baseline event in the conditional distribution of $X_{V\setminus ab}|\{X_a, X_b\}$.

**Definition 2.** *Given the set $X_V$ of variables associated to an undirected graph $\mathcal{G} = (V, \mathcal{E})$, for any $\delta \in \Delta_{ab}$ with $a, b \in V$, let $Y_\delta$ be a binary variable which takes level 1 if the event associated to the baseline cell $i_\delta \in \mathcal{I}_{V\setminus ab}$ realises, and 0 otherwise.*

Any $\delta \in \Delta_{ab}$ induces the partition $\{X_a, X_b, X_\delta, X_{\setminus\delta}\}$ of the random variables in $X_V$. The variable $Y_\delta$ is defined by collpasing the marginal table $\mathcal{I}_{V\setminus ab}$ with respect to the baseline cell related to the event $\{X_\delta = 1_\delta, X_{\setminus\delta} = 0_{\setminus\delta}\}$:

$$Y_\delta = 1 \quad \text{if } \{X_\delta = 1_\delta, X_{\setminus\delta} = 0_{\setminus\delta}\}, \quad \text{and} \quad Y_\delta = 0 \quad \text{otherwise.}$$

Each path $\delta \in \Delta_{ab}$ defines a probability table $\mathcal{I}_K$, with $K = \{\delta, a, b\}$, for a vector of three binary variables $X_K = (Y_\delta, X_a, X_b)$ following a trivariate Bernoulli distribution with probability parameter $\pi^\delta = (\pi^\delta_E)_{E \subseteq K}$ where

$$\pi^\delta_E = \pi_E, \qquad \delta \in E \subseteq K, \tag{4}$$

$$\pi^\delta_E = \sum_{E' \subseteq V\setminus ab : E' \neq \delta} \pi_{EE'}, \qquad \delta \notin E \subseteq K. \tag{5}$$

The log-linear parameter is $\log \theta^\delta = (\log \theta^\delta_E)_{E \subseteq K}$ with $\log \theta^\delta_E = \sum_{E' \subseteq E} (-1)^{|E \setminus E'|} \log \pi^\delta_{E'}$, for each $E \subseteq K$. We consider a logistic parameterisation for the conditional distribution of $Y_\delta|\{X_a, X_b\}$ to study the relationship between $\theta^\delta$ and $\psi$ parameters. For any $\delta \in \Delta_{ab}$, we have

$$\log \frac{P(Y_\delta = 1|X_a = 1, X_b = 1)}{1 - P(Y_\delta = 1|X_a = 1, X_b = 1)} = \beta_\delta + \beta_{\delta a} X_a + \beta_{\delta b} X_b + \beta_{\delta ab} X_a X_b, \quad \delta \in \Delta_{ab}. \tag{6}$$

The regression coefficients in equation (6) are function of parameter $\psi$ (Lupparelli and Marchetti , 2023).

DRAFT

**Theorem 1.** *Let $P_{\mathcal{G}}(X_V;\psi)$ be an Ising model. For any $\delta \in \Delta_{ab}$ with $a,b \in V$, consider the set $\{Y_\delta, X_a, X_b\}$ of binary variables and the logistic regression parameters in equation (6) for the conditional distribution of $Y_\delta|\{X_a, X_b\}$. The following equivalences hold:*

$$\beta_{\delta E} \;=\; \log\left(\frac{\prod_{j,k\in\delta}\psi_{Ejk}}{\theta^{\delta}_{E}}\right), \qquad E \subseteq \{a,b\}. \tag{7}$$

## 4. The path-based decomposition

Exploiting the results of Theorem 1 and of Lemma 1, a path-based decomposition of the marginal probability $\mu_{ab}$ is derived for any $a,b \in V$ (Lupparelli and Marchetti , 2023).

**Theorem 2.** *Let $P_{\mathcal{G}}(X_V;\psi)$ be an Ising model. Given the set $\Delta_{ab}$ of paths joining the pair $a,b \in V$ of vertices,*

$$\mu_{ab} = \sum_{\delta\in\Delta_{ab}} \omega_\delta \frac{\Psi_{\delta ab\setminus\omega_\delta}}{1^T\pi(\Delta_{ab})}, \tag{8}$$

*where*

$$\omega_\delta = \psi_{a\delta_1}\psi_{\delta_1\delta_2}\ldots\psi_{\delta_m b}, \qquad \delta\in\Delta_{ab}, \tag{9}$$

*$\Psi_{\delta ab\setminus\omega_\delta}$ is the sub-product of parameter $\psi_{jk}$, $j,k \in \delta ab$ with entries omitted in $\omega_\delta$, and $\pi(\Delta_{ab})$ is a column vector with generic element*

$$\frac{\Psi_{\delta ab}}{\Psi_{\delta ab}+\Theta^{\delta}_{ab}}, \qquad \delta\in\Delta_{ab},$$

*with $\Theta^{\delta}_{ab} = \prod_{E\subseteq ab}\theta^{\delta}_{E}$.*

For any $\delta \in \Delta_{ab}$, the decomposition in equation (8) includes the measure $\omega_\delta$ which is an interpretable parameter of multivariate dependence able to quantify the intensity of the propagation of the variable status since it is a product of the odds ratios associated to the edge set of the path. The term $\Psi_{\delta ab\setminus\omega_\delta}$ represents the residual strength of association to compute the joint probability of the active path event, but it does not directly influence the transition of the variable status along the path. Finally, $1^T\pi(\Delta_{ab})$ is a normalizing measure given by the sum over all paths $\delta \in \Delta_{ab}$ of the probability $\pi_{\delta|ab}$ that $X_a$ and $X_b$ are active only because of variables of a specific path $\delta$ are active. It is worth noticing that the decomposition of Theorem 2 can be also generalised to pairwise measures of association, e.g., the marginal odds-ratio or the dependence ratio, if the intent is to disentangle the marginal dependence across multiple paths.

## 5. An application

Inspired by the recent stream of the literature on attack graphs and on Ising models for cyber-security analysis (Denuit and Robert , 2022), the proposed method for path analysis is illustrated through an application for industrial networks with the focus on cyber-security in operational technology (OT) systems. In this context, vertices of a graph are used to represent industrial assets (e.g., servers, firewalls, computers, devise) corresponding to atomic variables which become active when they receive an attack and not active otherwise. The graph skeleton is a priori defined depending on the connections/interactions between items. A basic assumption in OT systems is that a cyber attack can run only through the paths of the graph. As an illustrative example, consider the graph in Figure 2 better discussed in Lupparelli and Marchetti (2023). For any pair of vertices $\{j,k\} \notin \mathcal{E}$, the probability of $X_j$ to be attacked/non-attacked is independent of the probability of $X_k$ to be attacked/non-attacked, given the status of their neighbors. An Ising model is assumed with the following log-linear parameters $\log\psi$ properly assigned on the basis

DRAFT

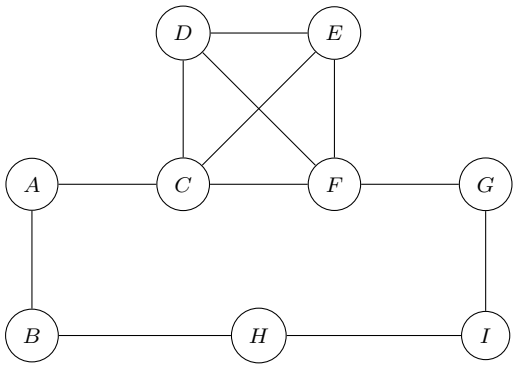



Figure 2: Ising model for a cyber-security case-study

of reasonings supported by the expertise of collaborators working in cyber-security field:

$$\begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{array} \begin{pmatrix} 0.1 & 0.2 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & 0.3 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 \\ . & . & -0.2 & 0.4 & 0.6 & 0.9 & 0 & 0 & 0 \\ . & . & . & 0.1 & 0.3 & 0.2 & 0 & 0 & 0 \\ . & . & . & . & 0.1 & 0.5 & 0 & 0 & 0 \\ . & . & . & . & . & 0.2 & 0.7 & 0 & 0 \\ . & . & . & . & . & . & -0.6 & 0 & 0.8 \\ . & . & . & . & . & . & . & 0.1 & 0.9 \\ . & . & . & . & . & . & . & . & 0.2 \end{pmatrix}.$$

the elements in the main diagonal represent the propensity of each item to be attacked depending on its vulnerabilities; the off-diagonal terms are related to the graph edges and reveal the risk of a local attack propagation which depends on the intensity and on the type of interaction between the items, on their own vulnerabilities and of their neighbours.

Suppose that item $B$ received an attack which spread to item $F$. We are interested in measuring the propagation risk of this attack over all paths starting from $B$ to $F$. The same reasoning holds for an attack going from $F$ to $B$. The path-based decomposition of the probability $\mu_{BF}$ provides insights to assess and to mitigate, at the same time, the risk of the attack propagation. Under the Ising model, $\mu_{BF} = 0.64$. Table 1 includes the list of six paths running between vertices $B$ and $F$ with the related risk measures $\omega_\delta$ of attack propagation computed using the decomposition in Theorem 2. Path $(6)$ has the highest risk since it includes edges with high scores showing great interconnections between vertices. The table includes the conditional probability $\pi^*_{\delta|BF}$, normalized over all the possible paths, that, since $B, F$ have beed attacked, the attack propagation occurred through path $\delta \in \Delta_{BF}$; paths (1) and (3) show the highest probability. Notice that paths (1) and (3), sharing the same vertices with different orderings, have the same probability $\pi^*_{\delta|BF}$ but different propagation risks $\omega_\delta$; conversely, paths (3) and (4) have the

| Path | size | $\omega_\delta$ | $\psi_{\delta_1\delta_2}$ | $\psi_{\delta_2\delta_3}$ | $\psi_{\delta_3\delta_4}$ | $\psi_{\delta_4\delta_5}$ | $\psi_{\delta_5\delta_6}$ | $\pi^*_{\delta\|BF}$ |
|---|---|---|---|---|---|---|---|---|
| (1) $(B, A, C, D, E, F)$ | 6 | 9.03 | 1.22 | 2.23 | 1.49 | 1.35 | 1.65 | 0.33 |
| (2) $(B, A, C, D, F)$ | 5 | 4.95 | 1.22 | 2.23 | 1.49 | 1.22 | - | 0.07 |
| (3) $(B, A, C, E, D, F)$ | 6 | 8.17 | 1.22 | 2.23 | 1.82 | 1.35 | 1.22 | 0.33 |
| (4) $(B, A, C, E, F)$ | 5 | 8.17 | 1.22 | 2.23 | 1.82 | 1.65 | - | 0.12 |
| (5) $(B, A, C, F)$ | 4 | 6.69 | 1.22 | 2.23 | 2.46 | - | - | 0.04 |
| (6) $(B, H, I, G, F)$ | 5 | 24.54 | 2.23 | 2.46 | 2.23 | 2.01 | - | 0.11 |

Table 1: Path-based analysis for attack propagation risks in a cyber-security case-study

DRAFT

same risks, since they show similar intensity of interconnection along the path, but different probabilities. A naive risk assessment for the network system focused on the pair of items $B$ and $F$ could be obtained by averaging the propagation risk $\omega_\delta$ with the probabilities $\pi^*_{\delta|BF}$ over all the paths, i.e,

$$R_{BF} = \sum_{\delta \in \Delta_{BF}} \omega_\delta \times \pi^*_{\delta|BF} = 9.95.$$

We also explore the effect of risk mitigation in the network structure. Suppose that we are able to protect both the item $B$ and the interconnection represented by the edge $H - I$ such that the parameters become $\log \psi_B = -0.1$ and $\log \psi_{HI} = 0.3$, respectively. Then, the propagation risks for paths (1)-(5) are unchanged since the edge $H - I$ appears only in the last path where the risk reduces from 24.54 to 13.46. Similarly, the normalized probabilities of paths (1)-(5) slightly change, whereas the conditional probability that path (6) is active goes from 11% to 6%. The naive risk assessment becomes $R_{BF} = 8.49$. Another possible action could be devoted to protecting the interconnection represented by the edge $A-C$ which is crucial for five paths. If the log-linear parameter can be reduced to $\log \psi_{AC} = 0.4$, then the propagation risks $\omega_\delta$ for paths (1)-(5) reduce to 6.04, 3.32, 5.47, 5.47 and 4.48, respectively. The related conditional probabilities slightly decreases with the exception of the last one that increases from 11% to 16%. The overall risk is almost the same, i.e. $R_{BF} = 8.43$. In this setting, the mitigation of the pairwise parameters related to the edges provides a reduction of the risk propagation parameter $\omega_\delta$; however this action might not reduce the overall risk which also depends on main effect parameters revealing the propensity of each item to be attacked/non-attacked.

## 6. Conclusion

Cyber-security risk assessment rises serious challenges for future research developments involving also computer science expertises. The industrial network is typically huge and not sparse, then any methodology needs to be scalable and able to handle very large tables. From this side, the proposed decompositions directly work on the rectangular log-linear parameter space, rather than on the probability space, and this represents a technical gain since log-linear parameters are variation independent. A relevant issue consists in the assignment of model parameters that should be dynamically learned and inferred from the activity of the network which rapidly and continuously changes over the time, as well as the vulnerability of vertices and of their interactions.

DRAFT

## References


[1] Denuit, M. and Robert, C.Y.: Networks: an introduction. Annals of Actuarial Science, **31**, 183–209 (2022)

[2] Ising, E.: Beitrag zur Theorie des Ferromagnetismus. Zeitschrift fur Physik A Hadrons and Nuclei, **31**, 253–258 (1925).

[3] Jones, B. and West, M.: Covariance decomposition in undirected Gaussian graphical models. Biometrika, **92**, 4, 779–786 (2005)

[4] Kruis, J. and Maris, G.: Three representations of the Ising model. Scientific Report, **6** (2013)

[5] Lupparelli, M. and Marchetti G.M.: Path-dependent parametric decompositions in Ising models. (*Submitted*)

[6] Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L.J., van der Maas, H.L.J. and Maris, G.: An Introduction to network psychometrics: relating Ising network odels to Item Response Theory models. Multivariate Behavioral Research, **53**, 15–35 (2018)

[7] Roverato, A. and Castelo, R.: Path weights in concentration graphs. Biometrika, **107**, 705–722 (2020)

[8] Wright, S.: Correlation and causation. Journal of Agricultural Research, **20**, 557–585 (1921)

# Enhancing Markowitz model: inspection of correlations and tail covariances

Gloria Polinesi[a]

[a]Department of Economics and Social Sciences - Universitá Politecnica delle Marche;
g.polinesi@staff.univpm.it

**Abstract**

Information contained in the correlation matrix of the financial products plays a crucial in order to construct portfolios as well as the tail effects of asset returns.

Structure hidden in the correlation matrix can be revealed appealing to hierarchcal clustering algorithms and spectral methods individually, or through a combination of them. Furthermore, covariance as a mesure of portfolio risk does not distinguish downside from upside risk. .

The work shows the state of the art of asset allocation models which enhances Markowitz portfolios focusing on Minimum Spanning Tree and Random Matrix Theory used to extract information from correlations and on different measures of portfolio risk accounting for asymmetry in the risk.

***Keywords:*** Financial time series, correlation matrices, tail dependence, cluster analysis, random matrix theory, asset allocation.

DRAFT

## 1. Introduction

Time series are one of the many instruments to represent data and this type of data is present in a variety of fields, from brain activity to financial area.

Looking for similarities between time series requires peculiar techniques due to their nature. Clustering algorithms, spectral methods and correlation based graph allow leading information from the structural organization of correlation matrix of the return time series whereas correlation matrices could be represented as complete graphs where the notion of hierarchy lacks [7]. The key role of information contained in correlation matrix of financial asset returns and the asymmetry in the portfolio risk - difference between downside and upside risk- is crucial in order to construct portfolios taylor made for investors.

The aim of the work is to show the state of the art of asset allocation models which enhances Markowitz portfolios focusing on the combination of hierachical clustering and spectral methods, i.e, Random Matrix Theory (RMT), used to extract information from correlations; and on different measures of portfolio risk accounting for asymmetry in the assets distribution.

The literature relative to clustering algorithms on correlation matrices of stock returns time series is very amply and it stems from the seminal paper of [19]. Specifically, this author investigates the correlation coefficient matrix to detect the hierarchical organization present inside the stock market: distance matrix based on correlation matrix is used to determine the Minimal Spanning Tree (MST) connecting the $N$ stocks. With only a few exceptions Mantegna shows that groups are homogeneous with respect to industry and often also subindustry sectors,

meaning that stocks belonging to the same sector or subsector are driven by the same economic factors.

The work of [34] shows how a MST calculated on a correlation matrix "filtered" from random noise through the RMT approach can improve the performance of the optimal portfolios. Similar papers, based on the applications of RMT to asset management, are [16], [31], [32], [5].

[11] extend the application of the RMT approach in [34] to Exchange Traded Fund returns (ETFs) but including a third dimension of risk, the network centrality, into the Markowitz objective function.

Furthermore, [21] provide a measure of systemic vulnerability of portfolio due to the asset centrality by combining in the Markowitz model tha classical covariance and the left-tail covariance. This new measure of portfolio risk overcome the drowbacks of the covariance which does not take into account asimmetry in the risk. Other works as [18] and [13] overcome the limitation of covariance since they are able to capture the downside risk through quantile-based measures.

The remainder of this paper is organized as follows. Sect. 2. introduces hierarchical clustering algorithms and extensions to "filter" correlation matrix of financial assets. Sect. 3. describes asimmetry measures of portfolio risk . Sect. 4. is devoted to portfolio optimization strategies. Finally, Sect. 5. draws some conclusions.

DRAFT

## 2. Methods for filtering correlation matrices

This Sect. describes hierarchical and not hierarichical methods used for filtering correlation matrices of return time series. These techniques reduce the number of parameters cleaning "measurement noise" due to the finite length of time series.

Mantegna's first work dates back to the 1999 and this author investigates the correlation matrix to detect the hierarchical organization of stocks in financial market. In a ultrametric space, the MST connecting stocks reveals a topological arrangement of financial market. Indeed, MST, defined as the minimum structure in terms of sum of distances between nodes, groups stocks homogeneous with respect to the economic sector of underlying companies.

Following this stream of literature, [35] confirm that nodes share information according to the communities they belong to whose are organized in a nested structure and hierarchical clustering algorithms allow to detect this complex structure.

[33] also qualify the minimal spanning tree as the corresponding representation of a fully-connected system (network) where sparseness replaces completeness in a proper way.

Steps in order to draw MST can be summarized as follow, starting from the correlation matrix of the time series of N asset returns, computed as difference of the logarithm of prices in the time horizon T [1].

$$r_i(t) = logP_i(t) - logP_i(t-1) \tag{1}$$

The elements of correlation matrix for each pair of assers

$$c_{ij} = \frac{E(r_i r_j) - E(r_i)E(r_j)}{\sigma_i \sigma_j} \tag{2}$$

converted in distance elements:

$$d_{ij} = \sqrt{2 - 2c_{ij}} \tag{3}$$

MST is drawn from the distance matrix in Eq. (3), which allows to shrink links connecting financial products from $\frac{N(N-1)}{2}$ (total number of parameters in the distance matrix) to $N-1$. In general, MST is able to detect the hierarchical organization in sectors and subsectors of stocks, but it is known in literature that result changes if frequency of data changes. [3] demonstrate

[1] Price observations can be daily, weekly, monthly or yearly.

that decreasing the time horizon (i.e. from daily to intraday frequency) correlation between pairs of stocks decreases by affecting the hierarchical organization: minimal spanning tree moves from a clustered and structured set to a simpler set.

Random matrix theory (RMT) represents the main non-hierarchical approach in order to investigate the structure of the financial correlations. The first results in the financial context can be found in [10], [15] and [28]. Specifically, RMT compares eigenvalues, $\lambda_k < \lambda_{k+1}$, of an empirical correlation matrix (Eq. 1) to eigenvalues of a random Wishart matrix $\mathbf{R} = \frac{1}{T}AA^T$. The $\mathbf{A}$ contains N time series of length T whose elements are independent, identically distributed random variables with zero mean and variance $\sigma^2 = 1$.
The random correlation matrix of this set of variables in the limit $T \to \infty$ is the identity matrix, when T is finite, the correlation matrix is in generally different from the identity matrix.
The theory of random matrices shows that in the limit $N \to \infty$ and $T \to \infty$ with a fixed ratio $Q = \frac{T}{N} \geq 1$ and $\sigma^2 = 1$, the eigenvalue spectral density is given by:

$$f(\lambda) = \frac{T}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{4}$$

where $\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}$. Information can be extract from eigenvalues that are higher then $\lambda_+$ (deviating eigenvalues) and it involves correlations between assets belonging to the same industry or geographical area; while, the "bulk" of eigenvalues agree with RMT reveals the random correlations [29].

Authors use RMT to filter correlation matrix and then construct MST because, in order to extract the structure hidden in large correlation matrices, trees are easier to interpret with respect to the inspection of large matrices [6]. By using this procedure, [23] obtain a clusterization per strategies of the Hedge Funds returns, according to the definition of strategies provided by [17].

DRAFT

## 3. Asimmetry measures of portfolio risk

The mean variance approach proposed by [22] to measure portfolio risk does not account for asymmetry in the risk. This is due to the fact that covariance is a measure of portfolio risk based on moments and, as consequence, does not distinguish downside from upside risk.

The quantile-based tail measures — value-at-risk (Var), expected shortfall (ES), extreme downside correlations (EDC) and p-tail risk (see, for example [18], [13]) — overcome the limitation of covariance in that they are able to capture the downside risk. However, their main drawback is that they are rather insensitive to the shape of the tail distribution since they strongly depend on the a priori choice of the confidence level and/or quantiles, thereby accounting mainly for the frequency of the realizations - not their values - [14].

In contrast to quantile-based approaches, the extreme downside hedge (EDH) of [13] is estimated by regressing asset returns on some measure of market tail risk. The latter, however, suffers from the above-mentioned drawback. Nevertheless, this approach is an attempt to use the values of asset returns to measure the tail risk.

Moreover, [21] propose a new variability measure for left-hand tail risk that overcomes the mentioned drawback based on the stratification procedure by [20]. This measure has the advantage that it is defined endogenously, without any a priori choice, and it captures risk from the asset volatility and co-movements as well as from the left-hand tail distribution of the asset returns, while preserving an elementary expression. In detail, the procedure in [20] can be adapted to the gross returns to identify a new informative set of parameters for each asset time series. This informative set is used to define the left-tail-covariance-like matrix via the cosine similarity and the new variability measure of portfolio risk obtained as a convex combination between classical portfolio variance and the left-tail-covariance.

## 4. Portfolio optimization strategies

Recent literature on financial time series has investigated techniques to exploit clustering information for developing a new approach to portfolio selection (see, among others, [30], [24], [25], [4], and [36]).

The idea underlying this approach is to substitute the original correlation matrix of the classical Markowitz model with an ultrametric correlation-based clustering matrix with the twofold objective of filtering the correlation matrix and improve the portfolio robustness to measurement noise.

The use of "improved covariance" matrix estimators as an alternative to the sample estimator is considered an important approach for enhancing portfolio optimization [26]. Indeed, most of the portfolio strategies alternative to the Markowitz model are based on RMT filtering procedure and ultrametric matrices associated to the hierarchical clustering algorithms applied to the empirical correlation matrices.

[29] use the filtering correlation matrix resulting from random matrix approach in the minimum variance model proposed by Markowitz showing that for these portfolios the realized risk is more closer to the predicted one. Moreover, [6] demonstrate that RMT is found to greatly reduce the difference between the predicted and realized risk of a portfolio, leading to an improved risk profile for a fund of hedge funds.

[34] construct portfolio by solving Markowitz solution but cleaning correlation matrix of stock returns through random matrix approach, single linkage and average linkage. Clustering algorithms combined with RMT filtering procedure improve the reliability of the Markowitz portfolio in terms of the ratio between predicted and realized risk. Also [35] prove that using the empirical correlation matrix leads to a dramatic underestimation of the real risk. In fact, they demonstrate that the risk of the optimized portfolio obtained using a "filtered" correlation matrix is more stable, although the real risk is always larger than the predicted one.

[11] describe how the use of filtered covariance matrices (through RMT and MST) and the dimension of systemic risk represented by eigenvector centrality between assets in the original Markowitz solution is outperformed in terms of standard portfolio performance measures.

Previous works consider Pearson correlation coefficient, other researches focus their attention on different correlations. For example, [2] apply an active portfolio management framework where interconnectedness between assets are based on mutual information measure, [8] and [9] propose to group time series of returns according to assets behaviour in tail (loss events) and then construct portfolios in order to manage risk during crisis scenario. [12] pick satellite (or pheripheral) assets according to their Adaptive Lasso Quantile Regression (ALQR) coefficients, that provide the information concerning the dependence between core portfolio and satellites at different tail events.

The recent study by [27] reveals that for a realistic stress test, special attention should be given to tail risk in individual returns and also tail correlations. In this context, the work of [21] interpretates the portfolio risk obtained by combining classical and left-tail covariances as the weighted average of the centralities of a suitably defined node weighted network [1].

## 5. Conclusion

The work shows the state of the art of asset allocation models which enhances Markowitz portfolios focusing on Minimum Spanning Tree and Random Matrix Theory used to extract information from correlations and on different measures of portfolio risk accounting for asymmetry in the risk.

These tools employed in the asset allocation models allow higher performing investments in terms of risk by enhancing the classical mean-variance portfolio model.

Filtering techniques combined with hierarchical clustering algoritms and tail movements of assets are tailored to manage risk of portfolio decreasing exposure to systemic vulnerability of

DRAFT

the financial products.

## References


[1] Abbasi, A., Hossain, L. Hybrid centrality measures for binary and weighted networks. In R. Menezes, A. Evsukoff and M. C. Gonzàlez (Eds.), Complex networks. Berlin: Springer-Verlag (2013)

[2] Baitinger, E., Papenbrock, J.: Interconnectedness risk and active portfolio management. Journal of Investment Strategies **6 (2)**, 63–90 (2017)

[3] Bonanno, G., Lillo, F., Mantegna, R. N.: High-frequency cross-correlation in a set of stocks. Quantitative Finance **1**, 96–104 (2001)

[4] Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N., Mantegna, R. N.: Networks of equities in financial markets. The European Physical Journal B **38 (2)**, 363–371 (2004)

[5] Bun, J., Bouchaud, J., Potters, M.: Cleaning large correlation matrices: tools from random matrix theory. Physics Reports **666**, 1–109 (2017)

[6] Conlon, T., Ruskin, H. J., Crane, M.: Random matrix theory and fund of funds portfolio optimisation. Physica A: Statistical Mechanics and its Applications **382 (2)**, 565–576 (2007)

[7] De Prado, M. L.: Building diversified portfolios that outperform out of sample. The Journal of Portfolio Management **42 (4)**, 59–69 (2016)

[8] Durante, F., Pappadà, R., Torelli, N.: Clustering of financial time series in risky scenarios. Advances in Data Analysis and Classification **8 (4)**, 359–376 (2014)

[9] Durante, F., Pappadà, R., Torelli, N.: Clustering of time series via non-parametric tail dependence estimation. Statistical Papers **56 (3)**, 701–721 (2015)

[10] Galluccio, S., Bouchaud, J., Potters, M.: Rational decisions, random matrices and spin glasses. Physica A: Statistical Mechanics and its Applications **259 (3-4)**, 449–456 (1998)

[11] Giudici, P., Polinesi, G., Spelta, A.: Network models to improve robot advisory portfolios. Annals of Operations Research, **313**, 965–989 (2022)

[12] Härdle, W. K., Lee, D. K. C., Nasekin, S., Petukhina, A.: Tail event driven asset allocation: evidence from equity and mutual funds' markets. Journal of Asset Management **19 (1)**, 49–63 (2018)

[13] Harris, R., Nguyen, L. H., Stoja, E.: Systematic extreme downside risk. Journal of International Financial Markets, Institutions and Money **61**, 128–142 (2019)

[14] Kuan, C., Yeh, J., Hsu, Y.: Assessing value at risk with CARE, the conditional autoregressive expectile models. Journal of Econometrics **150 (2)**, 261–270 (2009)

[15] Laloux, L., Cizeau, P., Bouchaud, J., Potters, M.: Noise dressing of financial correlation matrices. Physical Review Letters **83 (7)**, 1467 (1999)

[16] León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., Nino, J.: Clustering algorithms for risk-adjusted portfolio construction. Procedia Computer Science **108**, 1334–1343 (2017)

[17] Lhabitant, F., Learned, M.: Hedge fund diversification: How much is enough? The Journal of Alternative Investments **5 (3)**, 23–49 (2002)

[18] Liu, F., Wang, R.: Assessing value at risk with CARE, the conditional autoregressive expectile models. Mathematics of Operations Research **46 (3)**, 1109–1128 (2021)

[19] Mantegna, R. N.: Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems **11 (1)**, 193–197 (1999)

[20] Mariani, F., Ciommi, M., Chelli, F. M., Recchioni, M. C.: An iterative approach to stratification: poverty at regional level in Italy. Social Indicators Research, **161**, 873–903 (2022)

[21] Mariani, F., Polinesi, G., Recchioni, M. C.: A tail-revisited Markowitz mean-variance approach and a portfolio network centrality. Computational Management Science **19 (3)**, 425–455 (2022)


DRAFT

[22] Markowitz, H.: Portfolio selection. The Journal of Finance **7 (1)**, 77–91 (1952)
[23] Miceli, M., Susinno, G.: Ultrametricity in fund of funds diversification. Physica A: Statistical Mechanics and its Applications **344 (1-2)**, 95–99 (2004)
[24] Onnela, J., Chakraborti, A., Kaski, K., Kertesz, J., Kanto, A.: Dynamics of market correlations: taxonomy and portfolio analysis. Physica Review E **68 (5)**, 056110 (2003)
[25] Onnela, J., Chakraborti, A., Kaski, K., Kertesz, J., Kanto, A.: Asset trees and asset graphs in financial markets. Physica Scripta **48**, (2003)
[26] Pantaleo, E., Tumminello, M., Lillo, F., Mantegna, R. N.: When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators. Quantitative Finance **11 (7)**, 1067–1080 (2011)
[27] Paraschiv, F., Reese, S. M., Skjelstad, M. R.: Portfolio stress testing applied to commodity futures. Computational Management Science **17 (2)**, 203–240 (2020)
[28] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Stanley, H. E.: Universal and nonuniversal properties of cross correlations in financial time series. Physical Review Letters **83 (7)**, 1471 (1999)
[29] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., Stanley, H. E.: Random matrix approach to cross correlations in financial data. Physical Review E **65 (6)**, 066126 (2002)
[30] Puerto, J., Rodríguez-Madrena, M., Scozzari, A.: Clustering and portfolio selection problems: a unified framework. Computers & Operations Research **117**, 104891 (2020)
[31] Raffinot, T.: Hierarchical clustering-based asset allocation. The Journal of Portfolio Management **44 (2)**, 89–99 (2017)
[32] Ren, F., Lu, Y., Li, S., Jiang, X., Zhong, L., Qiu, T.: Dynamic portfolio strategy using clustering approach. PloS One **12 (1)**, e0169299 (2017)
[33] Spelta, A., Araújo, T.: The topology of cross-border exposures: beyond the minimal spanning tree approach. Physica A: Statistical Mechanics and its Applications **391 (22)**, 5572–5583 (2012)
[34] Tola, V., Lillo, F., Gallegati, M., Mantegna, R. N.: Cluster analysis for portfolio optimization. Journal of Economic Dynamics and Control **32 (1)**, 235–258 (2008)
[35] Tumminello, M., Lillo, F., Mantegna, R. N.: Correlation, hierarchies, and networks in financial markets. Journal of Economic Behavior & Organization **75 (1)**, 40–58 (2010)
[36] Wang, G., Xie, C., Stanley, H. E.: Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks. Computational Economics **51 (3)**, 607–635 (2018)

DRAFT

# Objective and subjective dimension of economic well-being: an approach based on statistical matching

Daniela Marella[a], Vincenzina Vitale[a], and Pierpaolo D'Urso[a]

[a]Department of Social and Economic Sciences, Sapienza University of Rome;
daniela.marella@uniroma1.it,
vincenzina.vitale@uniroma1.it,pierpaolo.durso@uniroma1.it

**Abstract**

Interest in the well-being measurement is constantly increasing worldwide. In addition to objective measures, interest often regards the degree to which people are satisfied with their economic conditions and a joint subjective and objective perspective to analyze the economic well-being is adopted. Data for statistical analysis is often available from different samples, with each sample containing measurements on only some of the variables of interest. In this context, statistical matching appears as a very useful technique.

In this paper the statistical matching between variables measuring the objective and the subjective economic well-being, available in different samples, is performed and its uncertainty is evaluated.

***Keywords:*** statistical matching, economic well-being, likelihood ridge, uncertainty

DRAFT

## 1. Introduction

Interest in the well-being measurement is constantly increasing worldwide in order to reduce social inequalities and to improve people's quality of life [12]. In addition to objective measures, interest often regards the degree to which people are satisfied with their economic conditions and a joint subjective and objective perspective to analyze the economic well-being is adopted. Household income is generally used as proxy for objective economic well-being as well as what people think about their economic situation is used as measure of subjective economic well-being.

In Italy, reliable information on households income is provided by EU Statistics on Income and Living Conditions (EU-SILC, for short). On the other hand information on subjective concepts such as emotional well-being, social participation and trust in institutions, complementary to EU-SILC are provided by Aspect of Daily Life survey (ADL, for short). Both surveys are carried out by ISTAT.

This constitutes a serious problem since joint information on objective and subjective economic well-being are used by policy makers for analyzing the impact of policy strategies. In this context, statistical matching appears as a very useful technique for the integration of multiple independent sources referring to the same, not overlapping, target population, as an alternative to implementing new surveys or the extension of the questionnaires of existing one.

In this paper the approach proposed in [3] to deal with the statistical matching is applied. The paper is organized as follows. In Sect. 2, the statistical matching problem for categorical variables and the concept of uncertainty are briefly recalled. Furthermore, the reduction of uncertainty by means of auxiliary information is discussed. In Sect. 3, such a methodology is applied to EU-SILC and ADL datasets.

## 2. Uncertainty in statistical matching for categorical data

Suppose $A$ and $B$ are two independent samples of $n_A$ and $n_B$ independent and identically distributed records from the same joint probability mass function (*pmf*) $p(x, y, z; \boldsymbol{\theta})$ referring to the categorical variables $(X, Y, Z)$ with $I$, $J$, and $K$ categories, respectively. In sample $A$, we only observe $X$ and $Y$, in sample $B$ we only have observations of $X$ and $Z$. The common variable $X$ is called matching variable.

At a *micro* level, the aim is to reconstruct a complete synthetic dataset with joint observations on all the variables of interest. At a *macro* level, the aim is the estimation of the joint *pmf* $p(x, y, z; \boldsymbol{\theta})$, see [4]. Because of the lack of joint information on all the three variables, $p(x, y, z; \boldsymbol{\theta})$ is not directly identifiable, unless under strong assumptions, which are generally hard to confirm.

Alternative approaches have been proposed in the literature to overcome the *identification problem*. The first approach assumes the conditional independence assumption between $Y$ and $Z$ given $X$ (CIA) [9]. A second approach assumes the existence of external information regarding the statistical relationship between $Y$ and $Z$, *i.e.*, a third sample $C$ in which $(X, Y, Z)$ has been observed [11] or proxy variables for $Y$, $Z$ as in [13]. A third approach consists in analyzing the uncertainty regarding $p(x, y, z; \boldsymbol{\theta})$. Under this approach, several alternative models for the joint distribution of $(X, Y, Z)$, compatible with the distributions of $(X, Y)$ and $(X, Z)$ in the samples $A$ and $B$ are considered.

In a parametric setting, the consequence of the identification problem is that only ranges of values containing all the pointwise estimates obtainable by each model compatible with the available sample information can be detected. Intervals defined by these ranges are known in the literature as *uncertainty intervals*, see [8; 10; 3]. Uncertainty in a nonparametric setting is analyzed in [1]. In [6] statistical matching under informative sampling designs is investigated. Finally, in [7] statistical matching under non-ignorable sampling and nonresponse is discussed.

When the variables of interest are categorical, uncertainty is dealt with in [3], where parameters uncertainty is estimated according to the maximum likelihood principle under the multinomial assumption of the joint $pmf$ of $(X, Y, Z)$. Notice that, if we perfectly know $P(X = i, Y = j) = \theta^*_{ij.}$ and $P(X = i, Z = k) = \theta^*_{i.k}$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, the parameter $\boldsymbol{\theta}^* = \{\theta_{ijk}\}$ lies in the following reduced parameter space

$$\Theta_{SM} = \left\{ \sum_k \theta_{ijk} = \theta^*_{ij.}, \sum_j \theta_{ijk} = \theta^*_{i.k}, \, \theta_{ijk} \geq 0, \sum_{ijk} \theta_{ijk} = 1 \right\}. \tag{1}$$

The parameter estimate which maximizes the likelihood function is not unique and the set of plausible maximum likelihood estimates for (1), is called *likelihood ridge*. It is defined as

$$\widehat{\Theta}_{SM} = \left\{ \sum_k \theta_{ijk} = \widehat{\theta}_{ij.}, \sum_j \theta_{ijk} = \widehat{\theta}_{i.k}, \, \theta_{ijk} \geq 0, \sum_{ijk} \theta_{ijk} = 1 \right\}, \tag{2}$$

where

$$\widehat{\theta}_{ij.} = \frac{n^A_{ij.}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B}, \quad \widehat{\theta}_{i.k} = \frac{n^B_{i.k}}{n^B_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B}, \tag{3}$$

and $n^A_{ij.}$ denotes the number of observations in sample $A$ with $(X = i, Y = j)$. Similar definitions hold for the remaining quantities in (3). All the distributions in the likelihood ridge are equally informative, given the data.

Uncertainty measures can be defined by analyzing the characteristics of the parameter space $\widehat{\Theta}_{SM}$. Since each parameter $\theta_{ijk}$ lies in an interval $\theta^L_{ijk} \leq \theta_{ikk} \leq \theta^U_{ijk}$ whose upper and lower bounds can not be generally expressed in closed form, we use as uncertainty measure (UM) for each $\theta_{ijk}$ the interval length, that is $UM(\theta_{ijk}) = [\theta^U_{ijk} - \theta^L_{ijk}]$. Then, an overall uncertainty measure, proposed in [10], can be defined as follows

$$UM(\boldsymbol{\theta}) = \frac{\sum_{ijk} (\theta^U_{ijk} - \theta^L_{ijk})}{M} \tag{4}$$

DRAFT

where $M$ is the number of uncertain parameters. In this context, extra-sample information (constraints) leading to a restriction of the parameter space $\Theta_{SM}$ to a subspace $\Omega \subset \Theta_{SM}$ can be very useful for reducing the overall uncertainty. Clearly, the more informative is the constraint, the greater is the reduction in uncertainty (4).

In Sect. 3, the likelihood ridge is analyzed under two scenarios: (i) without constraints, referring to the estimation of the unconstrained parameter space $\Theta_{SM}$ (unconstrained scenario); (ii) under constraints defined as structural zero, referring to the estimation of the constrained parameter space $\Omega$ (constrained scenario).

## 3. Application to EU-SILC and ADL

ADL multi-purpose survey belongs to the National Statistical Plan and is carried out annually by ISTAT. It collects information on the daily life of both individuals and households covering different social aspects referring to the individual's quality of life, the degree of satisfaction with one's living conditions, the economic situation. Subjective motivations and opinions contribute, together with objective information on social life, health, work, and lifestyles to define social information. For the 2019 edition, the sample consists of 19536 households. EU-SILC survey's main objective is to provide data on income, poverty, social exclusion, and living conditions. For the 2019 edition, the sample consists of about 22000 households. In this application we consider the following variables:

1. $Y$ =*Total Disposable Household Income*, henceforth *Income*, discretized in four ordinal categories : 1 [0, €16100); 2 [€16100, €25400); 3 [€25400, €40800); 4 [€40.800, $\infty$). The categorization is based on the quartiles.
2. $Z$ =*Satisfaction with the economic situation over the past 12 months*, henceforth *Sitec*, based on an ordinal four-point scale assigning 1 to *very much*; 2 to *quite a lot*; 3 to *a little*; 4 to *not at all*.

With regard to the matching variables the literature highlights two main criteria for their selection. First of all, there must be both homogeneity in their statistical content and similarity in the distributions of the variables across the two surveys. Secondly, the variables must be significant in explaining variations in the target variables.

In the application, the variables *Household size* (*Ncomp*), *Geographical area of residence* (*Area*) and *Occupational status* (*Condlav*), have been considered as possible matching variables and have been harmonized across the two datasets. The Hellinger distance is equal to 7%, 9% and 10% for *Ncomp*, *Area* and *Condlav*, respectively. Furthermore, all the variables are statistically significant in explaining variations in both $Y$ and $Z$. Then, $X$ = *Ncomp* is chosen as matching variable.
The exploration of the likelihood ridge, for the unconstrained and constrained scenario, has been done by using EM algorithm considering 100000 random starting points, see [3]. The restricted parameter space is defined by imposing the following structural zeros in the contingency table: $\theta_{i,1,1} = \theta_{i,1,2} = \theta_{i,4,3} = \theta_{i,4,4} = 0 \; for \; i = 1, \ldots, 4$; *i.e* we impose that a households with income below the first quartile cannot express an high or quite high degree of satisfaction with their economic resources, as well as households with income greater or equal to the last quartile cannot be totally or partially dissatisfied.

Table 1 shows the simulation extremes of the likelihood ridge $[\theta^L_{ijk}, \theta^U_{ijk}]$ in 100000 runs of EM for both scenarios, the average values $\bar{\hat{\theta}}_{ijk}$ over the 100000 estimates and the uncertainty measure $UM(\theta_{ijk})$. Finally, last column reports the estimate under the CIA assumption ($\hat{\theta}^{CIA}_{ijk}$).

The main finding in Table 1 refers to the reduction of ranges for some of the estimated probabilities when auxiliary information is introduced (see columns 7 and 11). In some cases, the introduction of structural zeros is so informative that EM algorithm converges to a single estimate. Notice that, a reduction in the range can correspond to a higher concentration of its density. For instance, the density of $\theta_{122}$ (cell 6) approximated with the frequency distribution of the 100000 simulations, is reported in Fig 1 for the unconstrained (histogram on the left) and the constrained scenario (histogram on the right). As expected, the latter distribution is more concentrated, due to the extra-sample information. The overall uncertainty (4) is 0.036 under the unconstrained scenario. It reduces to 0.021 where structural zero are introduced.

DRAFT

Table 1: Range of probability estimates $[\theta^L_{ijk}, \theta^U_{ijk}]$, mean values $\overline{\hat{\theta}}_{ijk}$ and uncertainty measure $UM(\theta_{ijk})$ over 100000 runs of EM under the unconstrained scenario and the constrained scenario. The last column refers to CIA estimate ($\hat{\theta}^{CIA}_{ijk}$)

| | Variables | | | Unconstrained Likelihood ridge | | | | Constrained Likelihood ridge | | | | CIA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ncomp | Income | Sitec | $\theta^L_{ijk}$ | $\theta^U_{ijk}$ | $\overline{\hat{\theta}}_{ijk}$ | $UM$ | $\theta^L_{ijk}$ | $\theta^U_{ijk}$ | $\overline{\hat{\theta}}_{ijk}$ | $UM$ | $\hat{\theta}^{CIA}_{ijk}$ |
| 1 | 1 | 1 | 1 | 0.000 | 0.016 | 0.009 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.008 |
| 2 | 1 | 2 | 1 | 0.000 | 0.016 | 0.005 | 0.016 | 0.000 | 0.015 | 0.010 | 0.015 | 0.002 |
| 3 | 1 | 3 | 1 | 0.000 | 0.015 | 0.002 | 0.015 | 0.000 | 0.015 | 0.004 | 0.015 | 0.001 |
| 4 | 1 | 4 | 1 | 0.000 | 0.012 | 0.000 | 0.012 | 0.000 | 0.015 | 0.001 | 0.015 | 0.000 |
| 5 | 1 | 1 | 2 | 0.011 | 0.162 | 0.087 | 0.151 | 0.000 | 0.000 | 0.000 | 0.000 | 0.089 |
| 6 | 1 | 2 | 2 | 0.001 | 0.116 | 0.062 | 0.115 | 0.107 | 0.122 | 0.112 | 0.015 | 0.022 |
| 7 | 1 | 3 | 2 | 0.000 | 0.052 | 0.029 | 0.052 | 0.039 | 0.054 | 0.050 | 0.015 | 0.007 |
| 8 | 1 | 4 | 2 | 0.000 | 0.019 | 0.011 | 0.019 | 0.005 | 0.019 | 0.018 | 0.015 | 0.006 |
| 9 | 1 | 1 | 3 | 0.001 | 0.111 | 0.054 | 0.110 | 0.120 | 0.120 | 0.120 | 0.000 | 0.054 |
| 10 | 1 | 2 | 3 | 0.000 | 0.100 | 0.037 | 0.100 | 0.000 | 0.000 | 0.000 | 0.000 | 0.012 |
| 11 | 1 | 3 | 3 | 0.000 | 0.052 | 0.016 | 0.052 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 |
| 12 | 1 | 4 | 3 | 0.000 | 0.019 | 0.006 | 0.019 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 |
| 13 | 1 | 1 | 4 | 0.000 | 0.042 | 0.021 | 0.042 | 0.044 | 0.044 | 0.044 | 0.000 | 0.020 |
| 14 | 1 | 2 | 4 | 0.000 | 0.042 | 0.014 | 0.042 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 |
| 15 | 1 | 3 | 4 | 0.000 | 0.036 | 0.005 | 0.036 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 |
| 16 | 1 | 4 | 4 | 0.000 | 0.017 | 0.001 | 0.017 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 |
| 17 | 2 | 1 | 1 | 0.000 | 0.012 | 0.001 | 0.012 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 |
| 18 | 2 | 2 | 1 | 0.000 | 0.013 | 0.003 | 0.013 | 0.000 | 0.013 | 0.003 | 0.013 | 0.003 |
| 19 | 2 | 3 | 1 | 0.000 | 0.013 | 0.005 | 0.013 | 0.000 | 0.013 | 0.004 | 0.013 | 0.001 |
| 20 | 2 | 4 | 1 | 0.000 | 0.013 | 0.003 | 0.013 | 0.000 | 0.013 | 0.006 | 0.013 | 0.001 |
| 21 | 2 | 1 | 2 | 0.000 | 0.040 | 0.024 | 0.040 | 0.000 | 0.000 | 0.000 | 0.000 | 0.062 |
| 22 | 2 | 2 | 2 | 0.001 | 0.078 | 0.044 | 0.077 | 0.001 | 0.078 | 0.041 | 0.077 | 0.044 |
| 23 | 2 | 3 | 2 | 0.003 | 0.099 | 0.055 | 0.096 | 0.010 | 0.097 | 0.052 | 0.086 | 0.013 |
| 24 | 2 | 4 | 2 | 0.001 | 0.078 | 0.044 | 0.077 | 0.066 | 0.079 | 0.072 | 0.013 | 0.010 |
| 25 | 2 | 1 | 3 | 0.000 | 0.039 | 0.012 | 0.039 | 0.008 | 0.040 | 0.029 | 0.031 | 0.037 |
| 26 | 2 | 2 | 3 | 0.000 | 0.072 | 0.023 | 0.072 | 0.000 | 0.073 | 0.026 | 0.073 | 0.023 |
| 27 | 2 | 3 | 3 | 0.000 | 0.081 | 0.030 | 0.081 | 0.000 | 0.078 | 0.033 | 0.078 | 0.008 |
| 28 | 2 | 4 | 3 | 0.000 | 0.075 | 0.023 | 0.075 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 |
| 29 | 2 | 1 | 4 | 0.000 | 0.026 | 0.003 | 0.026 | 0.000 | 0.031 | 0.011 | 0.031 | 0.014 |
| 30 | 2 | 2 | 4 | 0.000 | 0.031 | 0.008 | 0.031 | 0.000 | 0.031 | 0.009 | 0.031 | 0.008 |
| 31 | 2 | 3 | 4 | 0.000 | 0.031 | 0.011 | 0.031 | 0.000 | 0.031 | 0.012 | 0.031 | 0.003 |
| 32 | 2 | 4 | 4 | 0.000 | 0.031 | 0.008 | 0.031 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 |
| 33 | 3 | 1 | 1 | 0.000 | 0.006 | 0.000 | 0.006 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 |
| 34 | 3 | 2 | 1 | 0.000 | 0.007 | 0.001 | 0.007 | 0.000 | 0.006 | 0.000 | 0.006 | 0.004 |
| 35 | 3 | 3 | 1 | 0.000 | 0.007 | 0.002 | 0.007 | 0.000 | 0.007 | 0.001 | 0.007 | 0.002 |
| 36 | 3 | 4 | 1 | 0.000 | 0.007 | 0.004 | 0.007 | 0.000 | 0.007 | 0.006 | 0.007 | 0.002 |
| 37 | 3 | 1 | 2 | 0.000 | 0.014 | 0.008 | 0.014 | 0.000 | 0.000 | 0.000 | 0.000 | 0.027 |
| 38 | 3 | 2 | 2 | 0.000 | 0.024 | 0.014 | 0.024 | 0.000 | 0.018 | 0.006 | 0.018 | 0.056 |
| 39 | 3 | 3 | 2 | 0.001 | 0.054 | 0.029 | 0.053 | 0.000 | 0.018 | 0.011 | 0.018 | 0.029 |
| 40 | 3 | 4 | 2 | 0.004 | 0.076 | 0.041 | 0.072 | 0.073 | 0.080 | 0.074 | 0.007 | 0.024 |
| 41 | 3 | 1 | 3 | 0.000 | 0.014 | 0.005 | 0.014 | 0.000 | 0.014 | 0.010 | 0.014 | 0.016 |
| 42 | 3 | 2 | 3 | 0.000 | 0.024 | 0.008 | 0.024 | 0.000 | 0.024 | 0.014 | 0.024 | 0.030 |
| 43 | 3 | 3 | 3 | 0.000 | 0.050 | 0.018 | 0.050 | 0.018 | 0.052 | 0.032 | 0.034 | 0.018 |
| 44 | 3 | 4 | 3 | 0.000 | 0.055 | 0.026 | 0.055 | 0.000 | 0.000 | 0.000 | 0.000 | 0.016 |
| 45 | 3 | 1 | 4 | 0.000 | 0.012 | 0.001 | 0.012 | 0.000 | 0.014 | 0.004 | 0.014 | 0.006 |
| 46 | 3 | 2 | 4 | 0.000 | 0.017 | 0.002 | 0.017 | 0.000 | 0.018 | 0.004 | 0.018 | 0.011 |
| 47 | 3 | 3 | 4 | 0.000 | 0.019 | 0.006 | 0.019 | 0.000 | 0.019 | 0.011 | 0.019 | 0.006 |
| 48 | 3 | 4 | 4 | 0.000 | 0.019 | 0.009 | 0.019 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 |
| 49 | 4 | 1 | 1 | 0.000 | 0.004 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 |
| 50 | 4 | 2 | 1 | 0.000 | 0.006 | 0.001 | 0.006 | 0.000 | 0.001 | 0.000 | 0.001 | 0.003 |
| 51 | 4 | 3 | 1 | 0.000 | 0.007 | 0.002 | 0.007 | 0.000 | 0.001 | 0.000 | 0.001 | 0.003 |
| 52 | 4 | 4 | 1 | 0.000 | 0.007 | 0.004 | 0.007 | 0.005 | 0.007 | 0.007 | 0.002 | 0.004 |
| 53 | 4 | 1 | 2 | 0.000 | 0.011 | 0.006 | 0.011 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 |
| 54 | 4 | 2 | 2 | 0.000 | 0.020 | 0.011 | 0.020 | 0.000 | 0.002 | 0.001 | 0.002 | 0.044 |
| 55 | 4 | 3 | 2 | 0.000 | 0.046 | 0.024 | 0.046 | 0.000 | 0.002 | 0.001 | 0.002 | 0.042 |
| 56 | 4 | 4 | 2 | 0.009 | 0.080 | 0.045 | 0.071 | 0.084 | 0.085 | 0.084 | 0.002 | 0.046 |
| 57 | 4 | 1 | 3 | 0.000 | 0.011 | 0.004 | 0.011 | 0.000 | 0.011 | 0.008 | 0.011 | 0.006 |
| 58 | 4 | 2 | 3 | 0.000 | 0.020 | 0.007 | 0.020 | 0.001 | 0.020 | 0.015 | 0.019 | 0.023 |
| 59 | 4 | 3 | 3 | 0.000 | 0.045 | 0.016 | 0.045 | 0.026 | 0.047 | 0.034 | 0.021 | 0.026 |
| 60 | 4 | 4 | 3 | 0.001 | 0.057 | 0.031 | 0.056 | 0.000 | 0.000 | 0.000 | 0.000 | 0.031 |
| 61 | 4 | 1 | 4 | 0.000 | 0.010 | 0.001 | 0.010 | 0.000 | 0.011 | 0.003 | 0.011 | 0.002 |
| 62 | 4 | 2 | 4 | 0.000 | 0.015 | 0.002 | 0.015 | 0.000 | 0.018 | 0.005 | 0.018 | 0.008 |
| 63 | 4 | 3 | 4 | 0.000 | 0.019 | 0.005 | 0.019 | 0.000 | 0.019 | 0.012 | 0.019 | 0.009 |
| 64 | 4 | 4 | 4 | 0.000 | 0.019 | 0.011 | 0.019 | 0.000 | 0.000 | 0.000 | 0.000 | 0.010 |

DRAFT

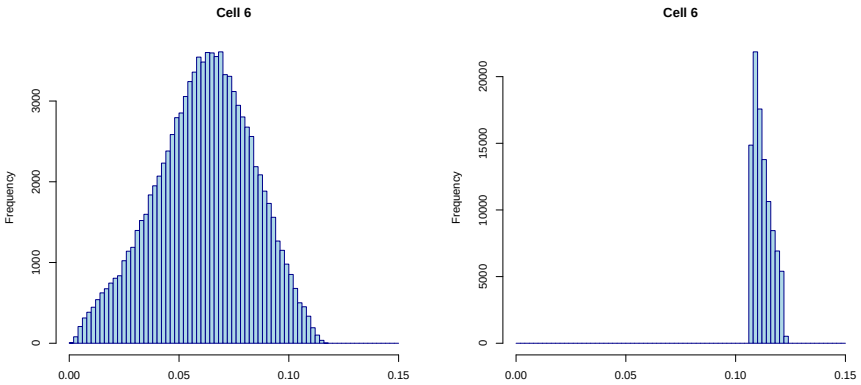



Figure 1: Likelihood ridge for cell 6 for scenario 1 (histogram on the left) and scenario 2 (histogram on the right).

Finally, the Goodman-Kruskal (Gamma) association index is computed over the 100000 estimated contingency tables under scenario 2. The histogram of the Gamma sample distribution has been reported in Fig. 2, from which it can be argued that there is a strong positive association between income and self-assessment of economic status. The negative sign is due to the reverse order of the categories of the *Sitec* variable with respect to the concept of satisfaction.

DRAFT

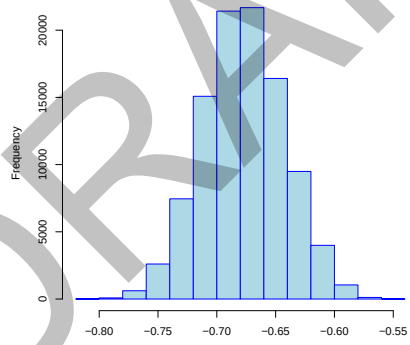



Figure 2: The values of Goodman-Kruskal Gamma association index over 100000 replications under scenario 2

With regard to the *micro* approach, by choosing a plausible joint *pmf* from the likelihood ridge, is possible to reconstruct a synthetic data set in which all variables of interest are observed. Further developments of the present work will focus on statistical matching in a multivariate context to pursue the goal of identifying the overall determinants of economic well-being. This can be achieved by embedding the methodological framework of Bayesian networks in the statistical matching as proposed in [2; 5].

## References


[1] Conti P.L., Marella D., Scanu M.: Statistical matching analysis for complex survey data with applications. *J Am Stat Assoc.* 111, **516**, 1715–1725 (2016)

[2] Conti P.L., Marella D., Vicard P., Vitale V.: Multivariate statistical matching using graphical modeling. *Int J Approx Reason* **130**, 150–169 (2021b)

[3] D'Orazio M., Di Zio M., Scanu M.: Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *J Off Stat* **22**, 137–157 (2006a)

[4] D'Orazio M., Di Zio M., Scanu M.: Statistical Matching: Theory and Practice. Chichester: Wiley, (2006b)

[5] Endres E., Augustin T.: Statistical matching of discrete data by Bayesian networks JMLR: Workshop and Conference Proceedings, **52**, 159–170 (2016)

[6] Marella, D., Pfeffermann, D.: Matching information from two independent informative sampling. *J Stat Plan Inference* **203**, 70-81 (2019)

[7] Marella, D., Pfeffermann, D.: Accounting for Non-ignorable Sampling and Non-response in Statistical Matching. *International Statistical Review*, https://doi.org/10.1111/insr.12524 (2022)

[8] Moriarity C., Scheuren F.: Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. *J Off Stat* **17**, 407–422 (2001)

[9] Okner B.: Constructing a new data base from existing microdata sets: the 1966 merge file. *Ann Econ Soc Meas* **1**, 325–342 (1972)

[10] Rässler S.: Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York (2002)

[11] Singh A.C., Mantel H., Kinack M., Rowe G.: Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Surv Methodol* **19**, 59–79 (1993)

[12] Stiglitz J., Sen A. Fitoussi J.P.: The measurement of economic performance and social progress revisited: Reflections and Overview, Sciences Po publications 2009-33, Sciences Po. (2009)

[13] Zhang, L.-C.: On proxy variables and categorical data fusion. *J Off Stat* **31**, 783–807 (2015)


DRAFT

# Comparison of traffic flow data sources for air pollution modelling

Theresa Smith[a] and Nick McCullen[b]

[a]Department of Mathematical Sciences, University of Bath, Bath, United Kingdom;
T.R.Smith@bath.ac.uk
[b]Department of Architecture and Civil Engineering, University of Bath, Bath, United Kingdom;
N.J.McCullen@bath.ac.uk

**Abstract**

An understanding of the relationship between traffic and pollution is crucial to enable policy makers and planners to manage the effects of pollution in urban areas. Typically air pollution modelling uses expensive, intensive monitoring, but in this article we explore the potential of using readily available surrogates for traffic conditions from crowd-sourced data. We design a travel delay metric to compare traffic flow estimates from a gold-standard automatic number-plate recognition (ANPR) survey in Bath, UK to data from the Google Maps traffic API. We observe very good agreement between the ANPR-based and Google-based delay ratios with two caveats: 1) we observe some evidence that the Google-based measures have a slight time delay and 2) Google-based measures are subject to truncation in low-traffic settings compared to ANPR-based delay measures.

***Keywords:*** air pollution, traffic modelling, ANPR, Google Maps

DRAFT

## 1. Introduction

Air pollution accounted for 6.4 million deaths in 2015 with more than half (4.2 million) caused by poor ambient air quality (1). These deaths, attributed to diseases such as cancer and cardiovascular diseases, are associated with pollutants such as carbon monoxide (CO), nitrogen dioxide ($NO_2$) and other nitrogen oxides ($NO_x$), as well as fine particulate matter (PM 2.5). Traffic is a major source of pollution in cities, along side other sources such as industry and household heating emissions. A recent report by the UK Department of Transportation on the contribution from different sources of pollution found that transport is responsible for a third of $NO_x$ emissions in the UK, with cars and taxis having the greatest share (2).

An understanding of the relationship between traffic and pollution is crucial to enable policymakers and planners to try to control the effects of pollution in urban areas. This understanding is usually obtained by modelling the emissions of traffic and the resulting pollution. The results of the modelling are then used by local governments in urban areas to develop regional policies based on measures such as restricting certain classes of vehicles from a city centre to reduce pollution concentrations. However, the data required for accurate models needs to be very comprehensive, requiring information on factors such as fleet age, typical average speeds, road classification and engine types as well as a detailed understanding of traffic conditions, especially congestion. Typically, this requires intensive on-the-ground monitoring, which is financially impractical for resource-constrained small local authorities.

The objective of the research we present here is to explore the potential of using readily available surrogates for traffic conditions from crowd-sourced data from mobile devices in place of expensive,

intensive monitoring. We compare the readily available data against an intensive survey over a two week period in the City of Bath, United Kingdom and find generally good agreement between the two. We also demonstrate the correlation of our novel metric with measured road-side air pollution levels.

## 2. Comparisons of Traffic Measures

In this section, we introduce two types of data collected during the survey period. Then we will define the concept of a 'delay ratio' and describe how we calculate this metric for our two data sources. Finally, we explore the relationships between these two measures.

### 2.1 Data Sources

Data on traffic levels is available at varying cost and resolution. Some sources of data give the flow of traffic in terms of average vehicle speeds at a point or along a segment or route; whereas, others consist of vehicle counts, which can be converted into either volume of cars passing a sensor in a given time window or speeds along a route if pairs of vehicle observations are combined, which requires recognising and recording the individual vehicles.

Data at the individual vehicle level can be collected by various methods, most often either human surveys or Automatic Number Plate Recognition (ANPR) cameras. The data from ANPR cameras can be cross referenced against a database of vehicles and their engine specifications to get detailed information about which vehicles passed which cameras at any given time, alongside the pollutants they are expected to emit. However, such methods are expensive and thus limited in their spatial density as well as the time for which they can be installed.

The Bath and Northeast Somerset Council carried out ANPR monitoring in the first two weeks of November 2017 to inform decision making about which vehicle types to charge in a Clean Air Zone later introduced in 2021. Cameras at 40 sites covering the main entry and exit points to the city were used to understand traffic patterns. The city publicly released anonymised versions of the ANPR data, with unique vehicle IDs replacing number plates. Observations of unique vehicle IDs at sequential cameras were then used to calculate journey times along selected origin-destination (OD) routes. Because vehicles are not continually observed, we cannot determine whether the journey between two cameras was direct or involved a stop off (e.g., at a petrol station or school). To focus on plausibly direct journeys, we exclude very long journey times (30 minutes in the examples below).

Crowd-sourced data from mobile devices in vehicles themselves are a promising alternative source of traffic data. These sources include commercial data from GPS navigation devices installed in vehicles (5), ride share taxi movement data (6), and Google's Traffic API (3). Google provide their Google Distance Matrix API to access their traffic data, which is based on the movements of all Android phones where the users have not disabled such access to their personal movements. Requests for live travel time data from Google's API between a selected OD pair can be made using various programming interfaces, including libraries in JavaScript and Python, and gives a proxy for vehicle flow rather than absolute counts. While this aggregate information contains no data about vehicle fleet composition, it is easily accessible, requires no installation of physical infrastructure and can be used instantly at any given location.

### 2.2 Defining a metric to compare ANPR and Google Maps

To compare the ANPR and Google data, we defined a new proxy for congestion called the *delay ratio*. While the ANPR can be used to measure both traffic volume and traffic flow speed between OD pairs, the Google data can only measure traffic flow. Although the Google data cannot be used to measure congestion directly, we can use increased journey times as a proxy for congestion.

We define the delay ratio, $D$, as the ratio of the travel time in a 15-minute window to the typical travel time on an OD route. This is a dimensionless measure where $D = 1$ means the journey times in the 15-minute window are typical, and proportional increases indicate delayed journeys and congestion.

DRAFT

For example, a value of $D = 1.2$ means the journey between specified end points takes 20% longer than expected. For the ANPR data, we assign each journey to a 15 minute interval based on the time the vehicle was recorded at the end of the route. We take the median journey time for each interval within the 15 minute window as the 'typical' travel time. This is then re-scaled by the median travel time over the whole two-week period to construct a delay index. For the analysis in this paper, we exclude 15 minute windows where fewer than 10 vehicles were observed. The travel times from Google were obtained by querying the API every 15 minutes. Compared to the ANPR data, we cannot know as precisely what the 15 minute travel time represents, and as we will see later, there are likely some important differences in terms of lags in Google's estimates of travel time in a given window.

## 3. Comparisons of delay metrics

To compare the delay metrics, we used data from a three mile route between two major roundabouts in Bath, shown in Figure 1.

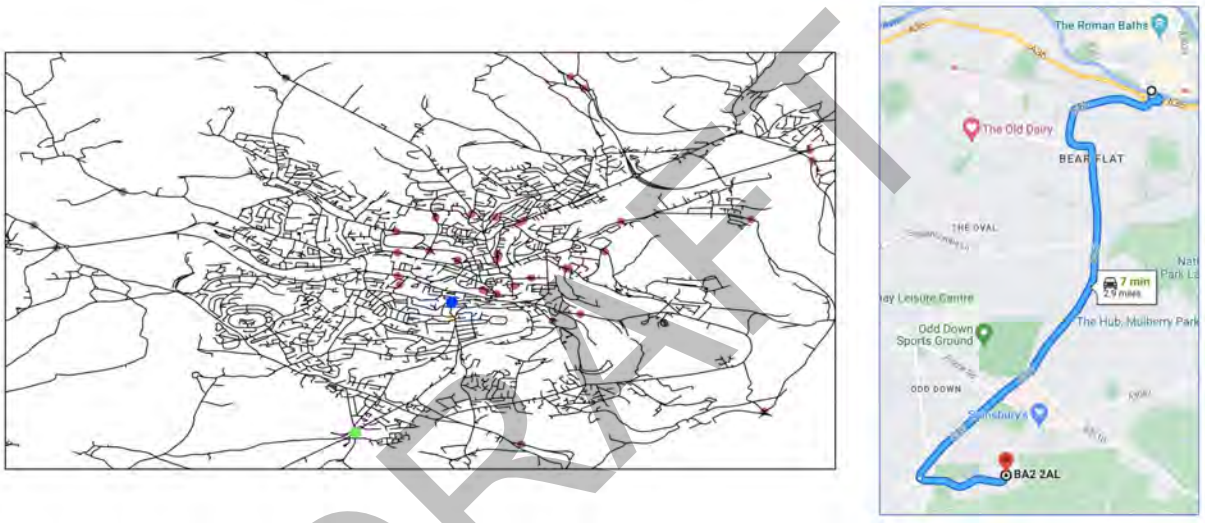
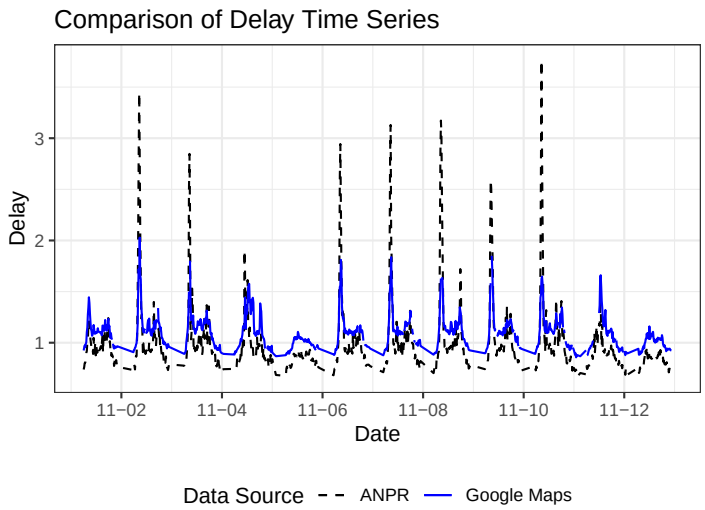



Figure 1: The left hand panel shows the major roads in Bath with the coloured dots indicating the camera locations in the ANPR survey. The blue and green dots are the locations of the cameras at the origin and destination, respectively, of the route studied in this section. The right hand panel shows the journey in Google Maps.

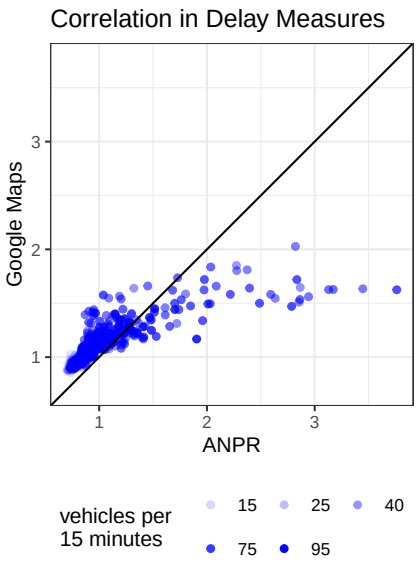
Exploratory analysis of the two types of delay metrics shown in Figure 2 indicates generally good agreement between the delay metrics defined by the two data sources. However, both the time series (Figure 2a) and the scatter plot (Figure 2b) show that the the ANPR-based measure of delay have larger extremes than the Google-based measure of delay. We note that the observations of extreme delays in the ANPR data cannot be explained by low traffic volumes in the corresponding time windows. Instead, we hypothesise that the Google Maps travel times are smoothed in some fashion that moderates extreme values. During light traffic conditions, when travel times are *below* the typical values, there is also a discrepancy between the two metrics, with Google-based values again being modulated towards 1, but this discrepancy is arguably less relevant to pollution modelling.

Figure 2a also indicates a small phase shift in the delay measurements, with the peaks in the Google Maps-based delay metric occurring just after the ANPR-based peak. We verified this by calculating the cross correlation function between the two time series across lags of up to 5 time windows (75 minutes). We found that the ANPR-based delay metric at time $t$ is most strongly correlated with the Google-baed delay metic at time $t + 1$ (i.e., the subsequent 15 minute interval). One potential explanation is that there is a gap between when users transmitting their location data to Google experience traffic congestion and when Google updates the estimated travel times for a user about to start the same journey.

DRAFT




(a) Comparison of delays metrics over 15 minute windows.




(b) Scatter plot of delay metrics.

Figure 2: Visual comparisons of delay metrics. In the right hand plot, the level of transparency indicates the number of vehicles that contributed to the ANPR delay calculations for a given 15 minute time window.

DRAFT

## 4. Relationships to Pollution

To explore the relationship between our proposed Google-based delay metric and air pollution, we queried journey times over short segments near four permanent air pollution monitoring stations over a period of several weeks in Autumn 2018 to Spring 2019. Figure 3 shows the delay ratio as defined above for east-bound traffic on the London Road, which is traffic entering the city, and nitrogen dioxide measurements from a road-side station that is part of the UK Department for Environment Food & Rural Affairs' Automatic Urban and Rural Network of air pollution monitoring stations.

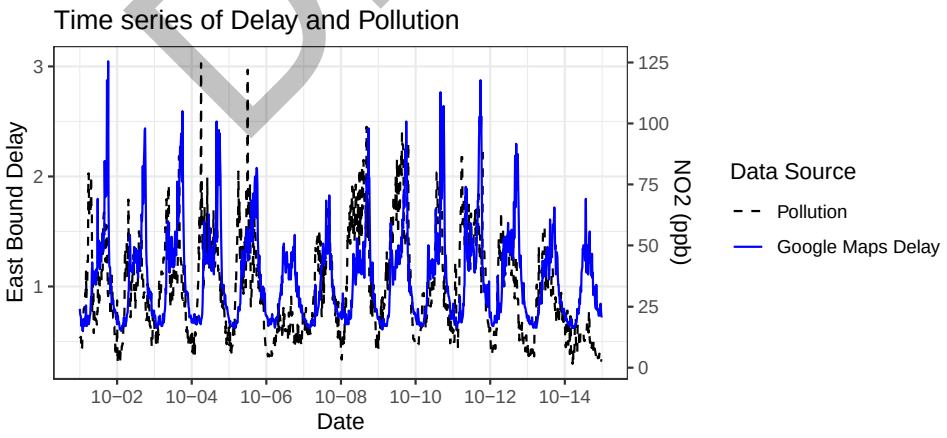



Figure 3: Comparison of the time series of Google Maps-based delay in East Bound traffic and $NO_2$ on the London Road in Bath in October 2018.

Figure 3 shows the comparison of delay and pollution for two weeks of data in October 2018. While there are some clear correlations between the traffic flow as measured by Google Maps and $NO_2$ levels, we again see that the two quantities are out of phase, with the peak in delays often counterintuitively coming after the peak in pollution. We also see some cases (e.g., October 6) where there is very weak

correlation between traffic flow and pollution levels. This behaviour is expected because many other environmental conditions, particularly wind speed and direction, will determine road-side pollution levels in a particular location alongside the amount of pollution being generated by traffic. State-of-the art air quality models would include emissions from traffic within physically motivated chemical transport models that describe the dispersion of pollutants over time as well as the advection of gasses and particles by the wind (4).

## 5. Conclusions

Overall, we observe very good agreement between the ANPR-based and Google-based delay ratios with two caveats: 1) we saw that the Google-based measures have a slight time delay compared and 2) Google-based measures are subject to truncation in low-traffic settings, leading under-estimates compared to ANPR-based delay ratios. Furthermore, we demonstrated that the Google-based measures show a strong association with a key pollutants in Bath, although the lagging issue persists. Taken together, these results indicate that crowd-sourced traffic flow data could be used a cheap yet effective proxy for intensive traffic monitoring in real-time air pollution modelling.

## References


[1] Cohen, A. J., Brauer, M., Burnett, R., et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet (2017) https://doi.org/10.1016/S0140-6736(17)30505-6

[2] Department of Transportation (2022) "Transport and environment statistics 2022" available at https://www.gov.uk/government/statistics/transport-and-environment-statistics-2022/transport-and-environment-statistics-2022

[3] Google Maps Platform (2023) "Distance Matrix API" available at https://developers.google.com/maps/documentation/distance-matrix/

[4] Sokhi, R.S., Moussiopoulos, N., Baklanov, A., et al. Advances in air quality research – current and emerging challenges. Atmospheric chemistry and physics (2022). https://doi.org/10.5194/acp-22-4615-2022

[5] TomTom N.V. (2023) "Flow Segment Data" available at https://developer.tomtom.com/traffic-api/documentation/traffic-flow/flow-segment-data

[6] Uber Technologies, Inc. (2023) "Uber Movement" available at https://movement.uber.com


DRAFT

# Data analysis of photogrammetry-based mapping: the seacucumbers in the Giglio Island as a case-study

Gianluca Mastrantonio[a], Daniele Ventura[b], Edoardo Casoli[b], Arnold Rakaj[c], Giovanna Jona Lasinio[d], and Alessio Pollice[e]

[a]DISMA - Politecnico di Torino
[b]DBA - Università di Roma La Sapienza
[c]BIO - Università di Roma Tor Vergata
[d]DSS - Università di Roma La Sapienza
[e]DiEF - Università degli Studi di Bari Aldo Moro

DRAFT

**Abstract**

Holothurians are marine invertebrates that produce ecosystem services increasing the local productivity of benthic communities, by e.g. releasing trapped nutrients, reducing stratification and increasing nutrification of sediments. In this work, we propose a point-process model that aims at better understanding the species distribution accounting for the effects of some ecological covariates.

***Keywords:*** Point process, Bayesian inference, Sampling effort

## 1. Introduction

Holothurians, commonly known as sea cucumbers, are benthic marine invertebrates belonging to the Phylum Echinodermata, Class Holothurioidea, represented by more than 1500 species worldwide constituting globally one of the most common organisms associated with soft bottoms [11]. Also, among temperate waters they play a key role as bioturbators, in fact, Holothurians offer ecosystem services that increase local productivity. Bioturbation of sediments by Holothurians releases nutrients trapped in the sediments to benthic ecosystems [11]. When Holothurians occur with high densities, they can reduce microalgal production and enhance the availability of nutrients, such as ammonium, increasing the gross productivity of benthic communities. Sediment digestion by Holothurians may be responsible for up to 50% of the dissolution of calcium carbonate in reef systems [9]. Holothurian bioturbation also reduces stratification and nutrification of sediments and can directly increase oxygen levels in the sediment [9]. The digestion of carbonate sands and metabolic activity of Holothurians are also a cause of turnover of inorganic carbon. The ecological role of Holothurians as bioturbators is thus pivotal in facilitating the availability of nutrients and oxygen for other organisms in both tropical and temperate waters [5].

Local fisheries exploit the species inhabiting coral reefs in tropical and subtropical regions, and their high value in Asian markets has encouraged global overfishing and associated declines. In response to the market demand, the fishery of sea cucumbers has shifted to new target species as the Mediterranean ones [13]. The current lack of information about the sea cucumbers' population dynamics, habitat use, and spatial ecology can lead to inadequate management [7].

Using underwater photogrammetry-based imagery (i.e., High spatial resolution ∼0.5 cm/pixel orthophotomosaics and Digital Surface Models), data on the presence, and relative coordinates, of Holothurians were collected in 5 temporal windows. We envision these data as a realization of a log-Gaussian Cox process, which is estimated under a Bayesian framework using the Integrated nested Laplace approximations (INLA) [8], implemented in the INLA software [2], available at `www.r-inla.org`. We propose 4 different models. In the first two, only the locations of Holothurians are modeled: while the first model assumes that there is a log-Gaussian process for each time-window, in the second the log-Gaussian process is shared. In the last two models we assume that Holothurians can be hidden by the presence of Posidonia oceanica meadows. Then the presence of Posidonia oceanica is described as a spatial process used as a predictor in the log-Gaussian Cox models introduced above, in a similar way as in the preferential sampling approach [1].

In what follows, we show some preliminary results that highlight the role of the natural Posidonia oceanica meadow and of some environmental covariates in the spatial distribution of the Holothurians.

## 2. Structure-from-Motion photogrammetry data and image processing

The recent widespread of Structure from Motion (SfM) photogrammetry for marine research [6] provides new tools and approaches to accurately map the distribution and behavior of many organisms in shallow aquatic environments including Holothurians that typically have high contrast against pale sandy sediments. Structure from Motion (SfM) photogrammetry allows extracting a plethora of fine-scale variables of benthic habitats by analyzing 3D digital models of the underwater environment, directly derived from overlapping two-dimensional camera images taken from different points of view [10]. SfM-derived ortho-mosaics can be analyzed in several ways, providing an excellent tool for accurate measurements that are difficult or even impossible to get in situ with traditional methods. They can provide also valuable information for fine-scale assessment and monitoring of benthic communities[12]. Some attempts have been carried out to map Holothurians in the tropical reef by using aerial imagery acquired by lightweight Unmanned Aerial Systems [13]. However, underwater photogrammetry-based imagery has never been used to assess Holothurians abundance and density among Mediterranean waters. In this study, to accurately map the seabed and the presence of two species of sea cucumbers (Holothuria tubulosa and H. poli) we used ultra-fine scale orthophoto mosaics derived by SfM processing according to the methodology proposed by [12].

Fieldwork was carried out in 5 temporal windows (seasons) on the east side of Giglio Island (central Tyrrhenian Sea, Italy), inside one of the four restoration sites located in the shipyard areas identified after the Costa Concordia shipwrecking. At the end of the wreck removal operations (July 2014) and after the seabed cleaning phase (April 2018), the natural environmental conditions were restored, ensuring the suitability of the area for transplant. Indeed, since 2019 transplanting operation has begun inside the shipyard area by using iron stakes to install into the dead matte bed detached fragments of P. oceanica. The area displayed a mosaic of habitat, including hard and soft bottom communities as well as natural (high-density meadow) and transplanted (low-density meadow) as P. oceanica beds. A diver propulsion vehicle equipped with spirit level, compass, depth gauge and a GoPro Hero10 Black action camera was used for image acquisition. Approximately 6000 still images were acquired by pointing the camera down towards the substrate at a near-nadir position, 4 m above the seabed, ensuring a resolution useful for micro-scale assessments.

Image processing was carried out through the 3D reconstruction software Agisoft Metashape v 1.6.2 (Agisoft LLC, Russia) generating digital surface models (DSMs) and orthophoto mosaics. ArcGIS 10.6 was used to manually digitize each visible Holothurian in the orthomosaic at each sampling event (seasons). As many as 984, 1478, 938, 769 and 668 sea cucumbers were respectively detected for each of the 5 seasons. DSMs implemented in the Spatial analyst extensions were used to estimate other seabed features such as slope, aspect, vector ruggedness (VRM) and rugosity. An object image analyses (OBIA) approach was adopted to speed up the classification of the orthophoto mosaics, leading to the definition of five cover classes: sandy bottoms, hard bottoms covered by photophilic algae, Posidonia oceanica, dead 'matte' of Posidonia oceanica and transplanted Posidonia oceanica.

DRAFT

| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
|---|---|---|---|---|---|
| M1 | Post. Mean | -0.001 | 0.017 | 1.207 | -41.914 |
| | 95% CI | (-0.016, 0.014) | ( 0.003, 0.031) | (-0.320, 2.734) | (-47.967, -35.860) |
| M2 | Post. Mean | 0.006 | 0.010 | 1.065 | -37.91 |
| | 95% CI | (0.005, 0.008) | (0.008, 0.012) | (0.101, 2.029) | (-42.324, -33.512) |
| M3 | Post. Mean | -0.008 | 0.315 | 3.132 | -59.497 |
| | 95% CI | (-0.101, 0.085) | (0.228, 0.402) | (1.243, 5.021) | (-67.990, -51.003) |
| M4 | Post. Mean | -0.009 | 0.314 | 3.033 | -59.108 |
| | 95% CI | (-0.102, 0.085) | (0.227, 0.402) | (1.133, 4.934) | (-67.563, -50.652) |

Table 1: $\log \lambda_t(\mathbf{x})$ - Posterior means and 95% credible intervals of the regression parameters in $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$.

## 3. Log-Gaussian Cox process models for the presence of Holothurians

Let $\mathbf{X}_t$ be a point process on the space $\mathcal{S} \in \mathbb{R}^2$ at times $t \in \{1, 2, 3, 4, 5\}$, and $\mathbf{x}_t = (\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{n_t,t})'$ be its realization, where $\mathbf{x}_{i,t} = (x_{i,1,t}, x_{i,2,t})'$ represents the presence of an Holothurian and its location. The idea is to model $\mathbf{X}_t$ as a log-Gaussian Cox process, with the following intensity function (letting $\mathbf{x}_{i,t} = \mathbf{x}$ for convenience):

$$\log \lambda_t(\mathbf{x}) = \beta_{0,t} + \mathbf{z}_t(\mathbf{x})\boldsymbol{\beta} + w_t(\mathbf{x}) \tag{1}$$

where $\beta_{0,t}$ is a time-specific intercept, $\mathbf{z}_t(\mathbf{x})$ is a vector of possibly time-dependent covariates, while the regressors $\boldsymbol{\beta}$ are the same for all $t$. The component $w_t(\mathbf{x})$ is the realization of a time-specific second-order stationary Gaussian process $W_t(\mathbf{x})$ with mean 0 and covariance function $C(h; \boldsymbol{\theta})$.

As an alternative model, we assume that the data at different time points share the same Gaussian process, possibly rescaled, hence

$$\begin{aligned} \log \lambda_1(\mathbf{x}) &= \beta_{0,1} + \mathbf{z}_1(\mathbf{x})\boldsymbol{\beta} + w_1(\mathbf{x}) \qquad \text{if } t = 1 \\ \log \lambda_t(\mathbf{x}) &= \beta_{0,t} + \mathbf{z}_t(\mathbf{x})\boldsymbol{\beta} + \delta_t w_1(\mathbf{x}) \qquad \text{if } t > 1 \end{aligned} \tag{2}$$

As a third type of model we implemented a modified version of the preferential sampling [1]. The idea is that the ability to observe the presence of Holothurians highly depends on the concentration of P. oceanica in the area. For this reason, letting $y(\mathbf{x})$ be the logistic transformation of the fractional P. oceanica coverage, we extend the first model assuming the following:

$$\begin{aligned} y(\mathbf{x}) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \rho v(\mathbf{x}) + \epsilon(\mathbf{x}) \\ \log \lambda_t(\mathbf{x}) &= \beta_{0,t} + \mathbf{z}_t(\mathbf{x})\boldsymbol{\beta} + w_t(\mathbf{x}) + v(\mathbf{x}) \end{aligned} \tag{3}$$

where $v(\mathbf{x})$ is a realization of a Gaussian process independent of $W_t(\mathbf{x})$ and $\epsilon(\mathbf{x})$ is an i.i.d. Gaussian nugget effect. The same idea is used to extend the second modeling approach, thus obtaining:

$$\begin{aligned} y(\mathbf{x}) &= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \rho v(\mathbf{x}) + \epsilon(\mathbf{x}) \\ \log \lambda_1(\mathbf{x}) &= \beta_{0,1} + \mathbf{z}_1(\mathbf{x})\boldsymbol{\beta} + w_1(\mathbf{x}) + v(\mathbf{x}) \qquad \text{if } t = 1 \\ \log \lambda_t(\mathbf{x}) &= \beta_{0,t} + \mathbf{z}_t(\mathbf{x})\boldsymbol{\beta} + \delta_t w_1(\mathbf{x}) + v(\mathbf{x}) \qquad \text{if } t > 1 \end{aligned} \tag{4}$$

Notice that models in (3), and (4) merge raster data for P. oceanica coverage with point data for Holoturian locations. This is possible since INLA allows to map data with different spatial support to a common estimation mesh by defining suitable projection matrices, thus avoiding the so-called change of support problem (COSP) [3].

## 4. Application

In this section, we show some preliminary results. In all models, we use as covariates the two spatial coordinates, the water temperature and the sea bottom rugosity, hence we have:

$$\mathbf{z}_t(\mathbf{x}) = (x_1, x_2, z_{\text{temp},t}(\mathbf{x}), z_{\text{rug}}(\mathbf{x}))'.$$

DRAFT

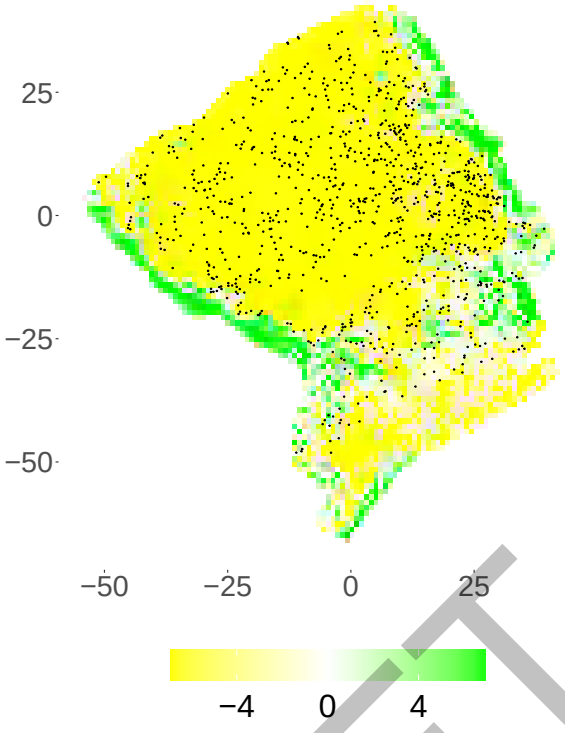



Figure 1: Graphical representation of the logistic transformation of the fractional P. oceanica coverage ($y(\mathbf{x})$). Dots correspond to the locations of holothurian at $t = 1$. Values below zero are depicted using a yellow gradient, while values above zero are illustrated using a green gradient.

| | | $\gamma_1$ | $\gamma_2$ | $\rho$ |
|---|---|---|---|---|
| M3 | Post. Mean | 0.002 | -0.004 | 0.928 |
| | 95% CI | (0.001, 0.002) | (-0.004, -0.003) | (0.359, 1.440) |
| M4 | Post. Mean | 0.002 | -0.004 | 0.994 |
| | 95% CI | (0.001, 0.002) | (-0.004, -0.003) | (0.390, 1.597) |

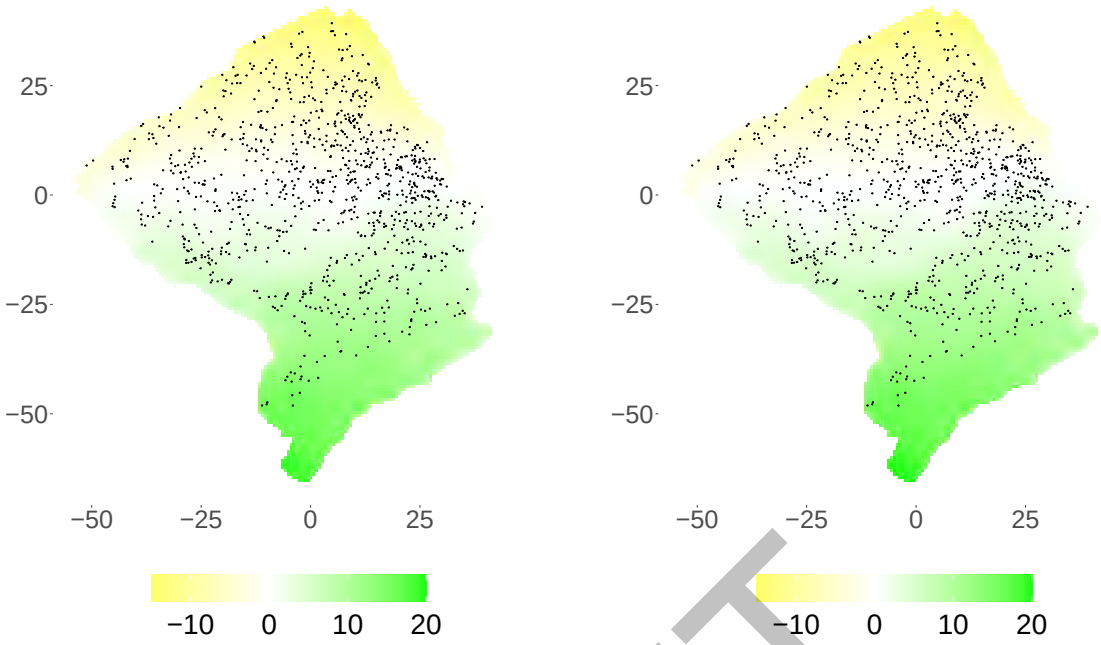
Table 2: $y(\mathbf{x})$ - Posterior means and 95% credible intervals of the regression parameters $\gamma_1$ and $\gamma_2$ and of the interaction parameter $\rho$.

DRAFT

To differentiate between the results of the four approaches, we indicate as M1, M2, M3, and M4 the models given by equations (1), (2), (3), and (4), respectively. The water temperature and the sea bottom rugosity were chosen as an example, but the appropriate selection of the available explanatory variables will contribute to an extended version of this work. Figure 1 shows evidence of an inverse relation between the recorded presence of Holothurians and the concentration of natural P. oceanica meadow that can hide the presence of Holothurians.

The INLA R package was used to estimate the four models and to obtain the WAIC [4]. It should be noted that the WAIC can't be used to compare models M1 and M2 with M3 and M4, since they consider different variables. Tables 1 and 2 contain the posterior means and 95% credible intervals of the regressive parameters in $\boldsymbol{\beta}$, $\gamma_1$ and $\gamma_2$ and of the interaction parameter $\rho$. Figure 2 shows the maps of the estimated shared spatial component $v(\mathbf{x})$ for models M3 and M4.

By analyzing the models output, there is evidence pointing toward models M2 and M4 being a better description of the data with respect to M1 and M3, which consider different Gaussian processes for every time point $t$. This is shown by the posterior estimates of the regressive parameters in Table 1 being quite similar across models and the posterior estimates of $\delta_t$'s in M2 and M4 being very close to the value 1 for all values of $t$. Moreover, WAIC values (not reported) are more favorable to model M2 than to M1 and to




(a) M3 (b) M4

Figure 2: Graphical representation of the estimated shared spatial component $v(\mathbf{x})$ for models M3 (left) and M4 (right). The color graduation is the same across the two pictures. Values below zero are depicted using a yellow gradient, while values above zero are illustrated using a green gradient. Dots correspond to the locations of holothurian at $t = 1$.

DRAFT

model M4 than to M3. It should be noted that, even though EDA shows an apparent inverse relationship between the presence of Holothurians and the concentration of P. oceanica, the positive estimated value of the coefficient $\rho$ implies a direct relationship between the two, meaning more Holothurians with more P. oceanica. This pushes toward the expected conclusion that P. oceanica meadows tend to hide the presence of Holothurians in the pictures, thus inducing a form of preferential sampling [1]. Moreover, the range of variation of the estimated shared spatial component $v(\mathbf{x})$ is much higher than the one of the residual spatial variation $w_t(\mathbf{x})$ across the four models for all values of $t$. Given that $\delta_t$'s have posterior estimates close to 1, this proves that $v(\mathbf{x})$ is more relevant than $w_t$ in explaining the spatial variation of the presence of sea cucumbers.

## 5. Conclusion

The preliminary results of this work show evidence that the proportion of Posidonia is relevant information to be used to describe the presence of Holothurians by a preferential sampling approach. In the future, we will enrich the model by using the Holothurians size as a mark, hence defining a marked point-process, and we will also investigate if and how the mark is affected by the Posidonia. We will also investigate how the other available covariates can be used to explain the presence of Holothurians and if their spatial distribution can be described by a point-process with clustering structure.

## References

[1] Diggle, P.J., Menezes, R., Su, T.: Geostatistical inference under preferential sampling. Appl. Statist. **59**(2), 191–232 (2010)

[2] Lindgren, F., Rue, H.: Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software, **63**(19), 1–25 (2015)
[3] Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M. (Eds.). (2010). Handbook of Spatial Statistics (1st ed.). CRC Press. https://doi.org/10.1201/9781420072884
[4] Gelman, A., Hwang, J., and Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat Comput **724**, 997–1016 (2014)
[5] Hammond, L. S.: Patterns of feeding and activity in deposit-feeding Holothurians and echinoids (Echinodermata) from a shallow back-reef lagoon, Discovery Bay, Jamaica. Bulletin of Marine Science, **32**(2), 549–571 (1982)
[6] Marre, G., Holon, F., Luque, S., Boissery, P., Deter, J.: Monitoring marine habitats with photogrammetry: a cost-effective, accurate, precise and high-resolution reconstruction method. Frontiers in Marine Science, **276** (2019)
[7] Pasquini, V., Porcu, C., Marongiu, M. F., Follesa, M. C., Giglioli, A. A., Addis, P.: New insights upon the reproductive biology of the sea cucumber Holothuria tubulosa (Echinodermata, Holothuroidea) in the Mediterranean: Implications for management and domestication. Frontiers in Marine Science, **9** (2022)
[8] Rue, H., Martino, S. Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **71**, 319–392 (2009)
[9] Schneider, K., Silverman, J., Kravitz, B., Rivlin, T., Schneider-Mor, A., Barbosa, S., ..., Caldeira, K.: Inorganic carbon turnover caused by digestion of carbonate sands and metabolic activity of Holothurians. Estuarine, Coastal and Shelf Science, **133**, 217–223 (2013)
[10] Ternon, Q., Danet, V., Thiriet, P., Ysnel, F., Feunteun, E., Collin, A.: Classification of underwater photogrammetry data for temperate benthic rocky reef mapping. Estuarine, Coastal and Shelf Science, **270**, 107833 (2022)
[11] Uthicke, S.: Nutrient regeneration by abundant coral reef Holothurians. Jour. Exp. Mar. Biol. Ecol. **265**(2), 153–170 (2001)
[12] Ventura, D., Mancini, G., Casoli, E., Pace, D. S., Lasinio, G. J., Belluscio, A., Ardizzone, G.: Seagrass restoration monitoring and shallow-water benthic habitat mapping through a photogrammetry-based protocol. Journal of Environmental Management, **304**, 114262 (2022)
[13] Williamson, J. E., Duce, S., Joyce, K. E., Raoult, V.: Putting sea cucumbers on the map: projected holothurian bioturbation rates on a coral reef scale. Coral Reefs, **40**, 559–569 (2021)

DRAFT

# Understanding forest damage in Germany: Finding key drivers to help with future forest conversion of climate sensitive stands

Nicole Augustin[a], Heike Puhlmann[b], and Simon Trust[b,c]

[a]School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK; nicole.augustin@ed.ac.uk
[b]Forest Research Institute Baden-Württemberg, Wonnhaldestrasse 4, 79100, Freiburg, Germany; heike.puhlmann@forst.bwl.de
[c]Abteilung Umwelt, FICHTNER Water & Transportation GmbH, Linnestr. 5, 79110 Freiburg; Germany; simon-trust@gmx.de

DRAFT

**Abstract**

Recently climate change has contributed to the decline in forest health, and yearly European forest health monitoring data are increasingly being used to investigate the effects of climate change on forests in order to decide on forest management strategies for mitigation. Forests in Germany have been badly affected and climate change now appears to be the major cause of defoliation (Eickenscheidt et al., 2019; Augustin et al., 2009). Thus, large scale forest conversions to more mixed forests with drought and heat resistant species are planned in some areas of Germany. This talk will cover the statistical aspects of a modelling project which has been informing decisions regarding this future forest conversion. Model selection is a challenge because of spatial confounding and the large number of correlated time varying environmental predictors. In addition there are computational challenges due to the large number of parameters and large sample sizes. A generalized additive mixed model is used for estimating spatio-temporal trends of defoliation, an indicator for tree health. Defoliation is modelled as a function of site characteristics (topography, soil and climate) with the aim of identifying the main factors associated with tree damage. The minimal model contains a space-time smoother and an AR1 process for temporal correlation. To eliminate predictors with negligible effects in the remaining set of predictors we use stability selection. Variable selection using integrated backward selection is carried out repeatedly with resampled data yielding selection inclusion frequencies. The final set of predictors are the predictors with selection inclusion frequencies above a certain threshold.

***Keywords:*** forest health, spatio-temporal model, generalised additive mixed model, stability selection, integrated backward selection

## 1. Introduction

Recently climate change has contributed to the decline in forest health, and yearly European forest health monitoring data are increasingly being used to investigate the effects of climate change on forests in order to decide on forest management strategies for mitigation. Forests in Germany have been badly affected and climate change now appears to be the major cause of defoliation (Eickenscheidt et al., 2019; Augustin et al., 2009). There is a strong association between drought stress and defoliation (an indicator

of tree vitality) of the main species. Thus, large scale forest conversions to more mixed forests with drought and heat resistant species are planned in some areas of Germany.

Here we model forest health data from the Terrestrial Crown Condition Inventory (TCCI), a forest health monitoring survey which has been carried out yearly in the forests of Baden-Württemberg, Germany since 1983. Forests in this area have been badly affected by climate change: In 2019 46% of the forest area in Baden-Württemberg was considered to be significantly damaged: that is trees had more than 60% defoliation in the crown. In particular tree species that were previously thought to be climate-resilient were also badly affected by heat, drought and storms in recent years. The available predictor variables cover different aspects contributing to forest damage with site characteristics (topography, soil, competition) contributing to pre-disposition. The primary damaging factors are effects of climate change (drought, heat, late frost) and secondary damaging factors (pathogens, insects). These factors interact with each other.

The aim of the modelling is to identify the main factors associated with tree damage. The results will help to formulate hypothesis regarding the causes of damage and identify areas which are still suitable for the main species. A generalized additive mixed model is used for estimating spatio-temporal trends of defoliation, an indicator for tree health. Defoliation is modelled as a function of site characteristics (topography, soil and climate) with the aim of identifying the main factors associated with tree damage.

Model selection is a challenge because of the large number of correlated time varying environmental predictors with non-linear effects of many predictors due to there often being an optimal range for a variable. For example, there is an optimal range of Julian day of budburst, and budburst below or above this range is associated with increased defoliation. In addition we have spatial confounding. This happens when spatial regression models use spatial random effects or smooths to account for residual spatial correlation in the response variable and result in fitted values that are smoothed across the spatial domain of the data (Dupont et al., 2022). This can lead to unrealiable effect as collinearity between covariates and spatial effects can lead to significant bias. In addition there are computational challenges due to the large number of parameters and large sample sizes. We use a pragmatic approach to address these challenges in the context of our complex modelling task.

DRAFT

## 2. Data

We model forest health data from the Terrestrial Crown Condition Inventory (TCCI), a forest health monitoring survey which has been carried out yearly in the forests of Baden-Württemberg since 1983. The survey is in alignment with the International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests and thus uses the same survey protocol (Eichhorn et al., 2017). The TCCI includes sampling points from the large-scale monitoring programme (level I) and longterm intensive monitoring areas (level II). The background and description of these two parts of the survey is given in Damman et al. (2001); de Vries et al. (2003) and Eichhorn et al. (2017). The level I data are essentially yearly repeated measures on a regular spatial grid with different subsets of locations missing, depending on the year resulting in sampling points on either a $4 \times 4$, $8 \times 8$ or $16 \times 16$ km grid. At each sampling grid point 24 sample trees with minimum height of 60 cm and a dominant and subdominant position within the forest stand are randomly selected using a protocol ensuring good spatial coverage within a 50 m radius. The trees are permanently marked and re-assessed during subsequent surveys. Trees that are removed are replaced by newly sampled trees. The level II sampling areas are of size 0.25 ha, predominantly in monoculture stands representing typical forest landscapes with the main species spruce, beech, fir, pine or oak. Only stands with trees of age 60 years or older were selected and all trees are permanently marked. Here we combine the level I (88 % of trees) and II (12% of trees) of the years 1991 to 2020 data for our analysis and this results in around 800 sampling grid points. The number of observations range between 1915 and 5320 for the different species.

Percentage of crown defoliation is recorded in both surveys and is the variable we are interested in modelling. This is recorded by eye for individual trees in 5% intervals. Defoliation is a good measure of tree vitality and health and has been heavily affected by climate change. Defoliation values in this survey range from 0% to 100% (see Figure 1). The main species (Norway spruce, beech, fir, pine and

oak) differ in their root system. These differences mean that the tree species are affected differently by climate change. In addition coniferous trees carry needles from the previous 7 years and this implies different levels of autocorrelation in time for coniferous and deciduous trees. So it makes sense to run separate model by species. Each observation records the average defoliation for the survey location by species.

As 24 trees are surveyed at each location, we model the yearly mean defoliation of the respective species of trees at each survey location. This is appropriate because we are interested in mean defoliation, the forest is heavily managed and at any location trees will be of the same age. There were more than 70 predictor variables available which have been selected due to a possible relationship to defoliation. Predictor variables include climate variables, deposition of pollutants (e.g. deposition of nitatrate), soil characteristics (e.g. coarse soil content), topography (e.g. topographic position and wetness index) , site characteristics (e.g. geology, slope direction) and stand charactersitics (e.g. mean age).

The average defoliation rate over the whole period was 0.25 (Figure 1).

## 3. Spatio-temporal model for defoliation

We model mean defoliation $\bar{y}_{it}$ of trees by species per location $i$ and year $t$ using an (generalised) additive mixed model (GAMM) (Eickenscheidt et al., 2019; Augustin et al., 2009).

$$\begin{aligned}\mathbb{E}(\bar{y}_{it}) &= f_1(\texttt{north}_i, \texttt{east}_i, \texttt{year}_t) + f_2(\texttt{stand age}_{it}) + \sum\nolimits_k f_k(x_{ik}) + \sum\nolimits_l f_l(z_{il}(t)) \\ &= \mathbf{X}_{it}\boldsymbol{\beta} \\ \bar{y}_{it} &= \mathbb{E}(\bar{y}_{it}) + \epsilon_{it}\end{aligned}$$

with $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Lambda}_i)$ where the covariance matrix $\boldsymbol{\Lambda}_i$ is within location $i$ based on an AR1 process. The number of trees $\texttt{nobs}_{it}$, on which the mean defoliation $\bar{y}_{it}$ is based on, is used as a weight. The non-linear function $f_1(.)$ is a three dimensional tensor product smooth interaction with separate penalties for space and time using a thin-plate spline basis for space and a cubic regression spline basis for time. The tensor product set up results in separate smoothing parameters for space and time making it scale invariant (see Wood (2017) section 5.6 for details). The $x_{ik}$ are stationary predictor variables and the $z_{il}(t)$ are time varying predictor variables. The smooth functions $f_k$ and $f_l$ for these are set up using cubic regression spline bases except for factor variables were we use a basis with a ridge penalty (i.e. the identity matrix). The ridge penalty is equivalent to an assumption that the coefficients are i.i.d. normal random effects. To simplify notation we let $\mathbf{X}_{it}$ be the combined model matrix row for location $i$ at time $t$, covariates for linear and random effects, basis functions evaluated at observations $it$ for covariates with smooth effects. The parameter vector $\boldsymbol{\beta}$ contains all coefficients for $\mathbf{X}_{it}$.

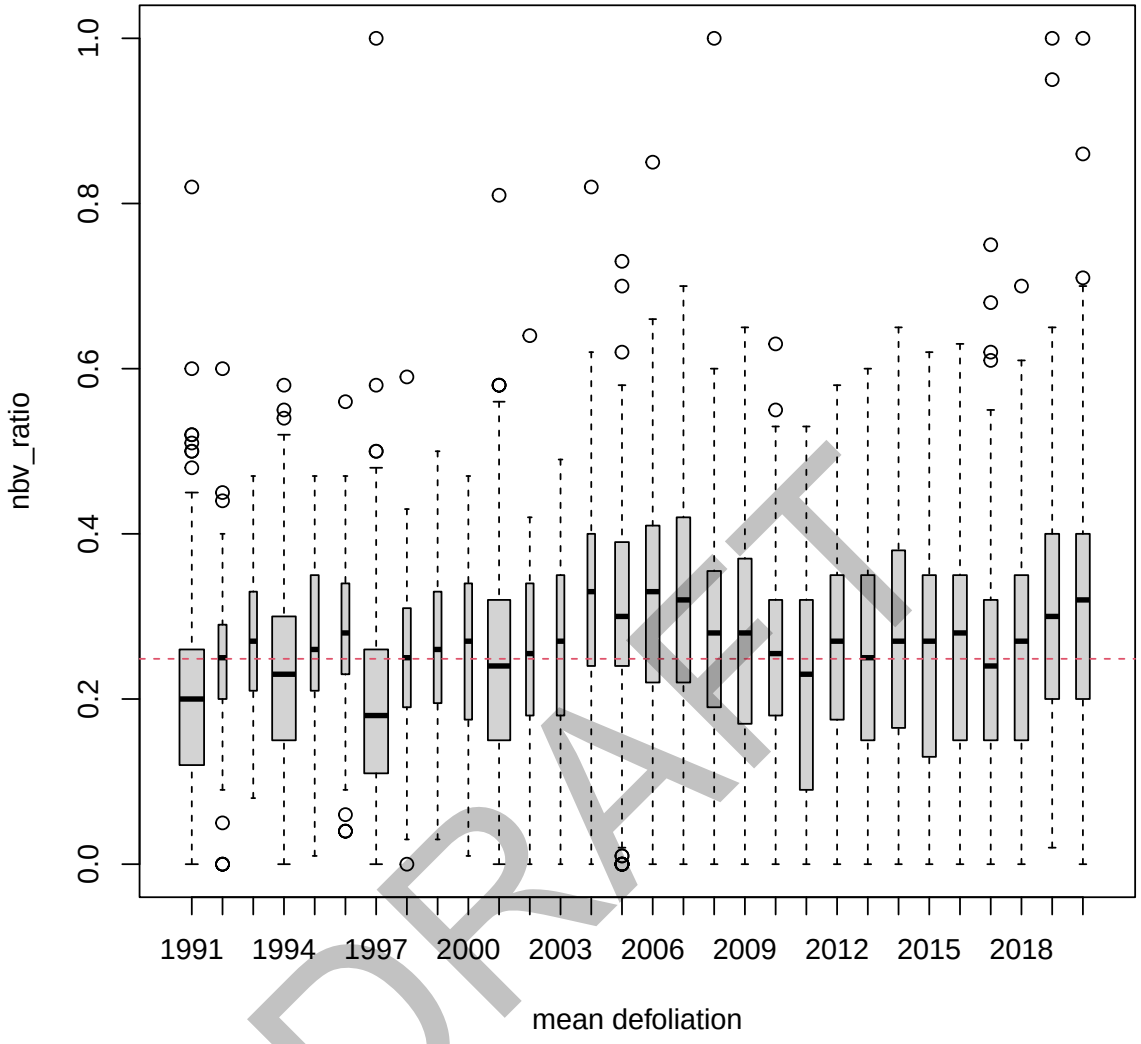
Note that we use a normal distribution with the identity link here. The response variable is an average of proportions and due to the central limit theorem we can expect the normal distribution to fit. It would also be possible to use a logit link in order to bound the fitted values between 0 and 1, but because we are mainly interested in the selected variables we are using the identity link for simplicity. If we were analysing the individual proportions rather than their mean, the quasibinomial family or beta distribution would be a possible choice. These options are all implemented in the `bam` and `gam` functions of the R package `mgcv` (Wood, 2017). To control smoothness and due to the random effect the penalized version of the log likelihood $l = \log L$, is optimised to find $\boldsymbol{\beta}$ given the penalty parameters $\lambda_j$, so

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} - l(\boldsymbol{\beta}) + \frac{1}{2}\sum\nolimits_j \lambda_j \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\beta} \tag{1}$$

The $\mathbf{S}_j$ is the penality matrix for smooth terms or random effect $j$ and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_j\boldsymbol{\beta}$ measures for each smooth term how rough the smooth function is. For example in the case of a cubic regression spline it is the integrated squared second derivative of the smooth function.

For estimating the smoothing parameters the fact that the model can be seen as a Bayesian model is exploited. This means the choice of smooth with a given type of penalty can be translated into a prior

DRAFT




DRAFT

Figure 1: Boxplot of yearly mean defoliation of Norway Spruce over time. Shown are the yearly mean defoliation values averaged over all spruce trees observed at a sampling point. The red dashed line shows the overall mean defoliation. The box widths are proportional to the square-roots of the number of observations in the years (633, 69, 59, 584, 67, 65, 600, 64, 63, 59, 525, 60, 61, 60, 199, 195, 193, 192, 189, 182, 179, 184, 188, 183, 185, 199, 197, 194, 203, 197).

$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}_\lambda^-)$ with $\mathbf{S}_\lambda = \sum_j \lambda_j \mathbf{S}_j$. Note that $\mathbf{S}_\lambda$ contains penalties relating to the smooth effects and the random effects. The $\mathbf{S}_j$ for the random effect $\mathbf{b}$ is $\mathbf{I}$ and the respective $\lambda_j$ is $\sigma^{-2}$. It follows that $\hat{\boldsymbol{\beta}}$ in (1) is the maximum a posteriori estimate. $\mathbf{S}_\lambda^-$ is an appropriate pseudo inverse. In this context smooths can be seen as latent Gaussian random fields and hence in terms of estimation we don't need to differentiate between random effects and smooths (see e.g. Kimeldorf and Wahba (1971); Silverman (1985)). Also for $n \to \infty$ and the number of parameters $p = o(n^{1/3})$ the posterior is (see e.g. Wood et al. (2017) section 6.10 for details)

$$\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\lambda} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\mathcal{T}}\mathbf{X} + \mathbf{S}_\lambda)^{-1}) \tag{2}$$

Using the above we can estimate $\boldsymbol{\lambda}$ to maximize the marginal likelihood

$$f(\mathbf{y}|\boldsymbol{\lambda}) = f(\mathbf{y}, \hat{\boldsymbol{\beta}}|\boldsymbol{\lambda})/f(\hat{\boldsymbol{\beta}}|\mathbf{y}, \boldsymbol{\lambda}) = L(\hat{\boldsymbol{\beta}})f(\hat{\boldsymbol{\beta}}|\boldsymbol{\lambda})/f(\hat{\boldsymbol{\beta}}|\mathbf{y}, \boldsymbol{\lambda})$$

This method can be parallelised, and made very efficient by marginal discretization of covariates (or exploiting natural discretization), thereby mitigating the $O(np^2)$ cost (Wood, 2011; Wood et al., 2017). Due to the large number of parameters and large sample sizes we use this efficient estimation implemented in the function `bam` in the R package `mgcv` in R (Wood, 2017).

## Model selection

As the 80 predictors contain variables which are measuring similar characteristics, we initially exclude redundant variables in discussion with experts. Then in a second step if pairs of variables have a strong association and are proxies of each other we ideally take the variable with less missing values and/or the variable which is more useful in terms of identifying causes of tree damage and predictor value ranges for critical conditions. The minimum model includes the space-time smooth and AR1 process for temporal correlation as it is likeley that this will lead to smaller bias in estimates of effects of variables of interest (Dupont et al., 2022) because (a) the model without the space-time smooth (and AR1) is misspecified and hence inference is not going to be correct and (b) important predictors we don't have, e.g. deposition, will be explained (at least partially) by the space-time smooth.

Because conventional backward selection would be rather slow for the number of predictors we use integrated backward selection with extra penalties (Marra and Wood, 2011). This method adds a penalty to the coefficients of each smooth penalizing the null space (the null space is the linear part in a one dimensional smooth) of its smooth so that they can in principle be penalized out of the model. This is done by replacing the zero eigenvalues corresponding to the linear part by ones in the eigendecomposition of the penalty matrix $\mathbf{S}_j$ and this makes the penalty corresponding to the linear part similar to a ridge penalty. Then we exclude the predictors which are below a threshold in terms of effective degrees of freedom (edf). An edf below 1 means that a penality has been also applied to the null space, i.e. linear, part of the smooth. This works for any smooth (also two-dimensional smooth for interactions).

To eliminate predictors with negligible effects resampling methods (bootstrap) are used. Variable selection is carried out repeatedly with resampled data yielding selection inclusion frequencies. Final model selection is carried out with a subset of predictors with selection inclusion frequencies $> 30\%$. The advantage of this approach is that we reduce the number of predictors and deal with correlated predictors in an automatic fashion. See Augustin et al. (2005); Holländer et al. (2006); Meinshausen and Bühlmann (2010).

DRAFT

## 4. Results

Preliminary results show that the variable selection has resulted in reducing the set of 70+ predictors, yielding (mostly) plausible sets of predictors for the different species. Shown in Table 1 are predictors selected in any of the models for spruce, beech, pine and oak. In the talk we will show more results mainly backing up the hypothesis that indeed the effects of climate change such as increased temperatures and dry conditions are contributing to forest damage.

## References

Augustin, N., Musio, M., von Wilpert, K., Kublin, E., Wood, S., and Schumacher, M. (2009). Modelling spatio-temporal forest health monitoring data. *Journal of the American Statistical Society*, 104(487):899–911.

Augustin, N., Sauerbrei, W., and Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, 5(2):95–118.

Table 1: Selected variables by species. SPEI stands for standard precipitation evaporation index for different time periods.

| variable | group | beech | spruce | oak | pine |
|---|---|---|---|---|---|
| age | | x | x | x | x |
| space-time smooth | | x | x | x | x |
| topogr. wetness index 25 | topography | x | x | x | x |
| topogr. position index 500 | topography | | x | | |
| humidity | soil | | | | x |
| organic carbon | soil | x | | | |
| humus | soil | | x | | |
| clay content | soil | | x | | |
| ph | soil | x | | | x |
| soil type | soil | | | | x |
| mean temp during beg. period (lag 1) | climate | | | x | |
| climatic water balance veg. period (lag 1) | climate | x | | | |
| hours sunshine in veg. period | climate | | | | x |
| SPEI 3 May | climate | x | | | |
| SPEI 3 August | climate | x | | | |
| SPEI 24 August | climate | | | x | |
| length of vegetation period | climate | x | x | | |
| N | | 4295 | 5320 | 2107 | 1915 |
| $R^2$ (age) (%) | | 23 | 50 | 31 | 3 |
| $R^2$ (age + selected variables -/+ space-time smooth) (%) | | 42/51 | 58/62 | 39/47 | 19/25 |

DRAFT

Damman, I., Herrman, T., Körver, F., Schröck, H., and Ziegler, C. (2001). *Dauerbeobachtungsflächen Waldschäden im Level II-Programm - Methoden und Ergebnisse der Kronenansprache seit 1983*. Bund-Länder-Arbeitsgruppe Level II / Arbeitskreis Krone. BMVEL, Bonn.

de Vries, W., Vel, E., Reinds, G., Deelstra, H., Klap, J., Leeters, E., Hendriks, C., Kerkvoorden, M., Landmann, G., Herkendell, J., Haussmann, T., and Erisman, J. (2003). *Intensive monitoring of forest ecosystems in Europe: 1. Objectives, set-up and evaluation strategy*, volume 174(1).

Dupont, E., Wood, S. N., and Augustin, N. H. (2022). Spatial+: a novel approach to spatial confounding (with discussion). *Biometrics*.

Eichhorn, J., Roskams, P., Potočìč, N., Timmermann, V., Ferretti, Mues, V., Szepesi, A., Durrant, D., Seletković, I., H-W.Schröck, Nevalainen, S., Bussotti, F., Garcia, P., and Wulff, S., editors (2017). *ICP Forests manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests.* Thünen Institute of Forest Ecosystems, Eberswalde,Germany.

Eickenscheidt, N., Augustin, N. H., and Wellbrock, N. (2019). Spatio-temporal modelling of forest monitoring data: Modelling german tree defoliation data collected between 1989 and 2015 for trend estimation and survey grid examination using gamms. *iForest Biogeosciences and Forestry*, 12:338–348.

Holländer, N., Augustin, N., and Sauerbrei, W. (2006). Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine*, 45(01):44–50.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Silverman, B. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, 47:1–52.

Wood, S. (2017). *Generalized Additive Models. An Introduction with R. Second Edition.* Chapman & Hall/CRC, Boca Raton.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata: modeling the uk black smoke network daily data. *Journal of the American Statistical Association*, 112(519):1199–1210.

DRAFT

# Inequalities in international students mobility

Kristijan Breznik[a], Giancarlo Ragozini[b], and Marialuisa Restaino[c]

[a]International School for Social and Business Studies, Celje, Slovenia;
kristijan.breznik@mfdps.si
[b]Department of Political Science, University of Naples Federico II, Napoli, Italy;
giragoz@unina.it
[c]Department of Economics and Statistics, University of Salerno, Fisciano (SA), Italy;
mlrestaino@unisa.it

**Abstract**

Since the international flow of students across countries has increased over the years, it is crucial to analyze the flows to understand the main characteristics of student mobility trajectories involved in both Erasmus and Erasmus+ programs. In this manuscript, we use data from the European Union Open Data Portal and focus on gender inequalities in STEM (Science, Technology, Engineering, and Mathematics) scientific fields. Specifically, using a network analysis approach, we aim to determine whether any differences exist in international students' mobility according to gender and between STEM and non-STEM fields. In general, the results revealed that male STEM students are more inclined towards Erasmus mobility than female students. Specifically, we identified Scandinavian countries as the most significant importers of STEM Erasmus mobility students.

***Keywords:*** student mobility, weighted network, gender gap, stem and no-stem, inequalities

DRAFT

## 1. Introduction

Internationalization is becoming an increasingly important topic for all higher education institutions, and the European Commission is providing new projects and programs to promote and encourage the internationalization of citizens, organizations, and universities.

In fact, the international flow of university students across countries has been increasing all around the world for several years. A general idea about the magnitude of incoming and outgoing students by country is possible to get from the website of the Institute of Statistics of UNESCO. [1]

Many European initiatives for learning and teaching mobility have been promoted to enhance personal development and increase cooperation between educational institutions. The most well-known program is the European Region Action Scheme for the Mobility of University Students (Erasmus), which aims to encourage and support academic mobility for students, professors, and academic staff in higher education within EU countries.

The number of participants in Key Action 1, *Learning Mobility of Individuals*, has increased from 3,000 participants in 1987 to 272,497 in 2013–2014, with approximately 954,000 individual mobility contracts concluded in 2020 under the Erasmus+ program 2014–2020 [2].

Exploring inequalities in international student mobility can be a complex and multi-faceted task that requires careful consideration of various factors such as socioeconomic background, nationality, gender,

[1]For details see http://uis.unesco.org/en/uis-studentfow.
[2]https://data.europa.eu/doi/10.2766/36418

language proficiency, and cultural capital. Over the years, several studies have been conducted to study international students' mobility, drawing trajectories and analyzing the influence of some factors on the decision to study abroad.

An interesting field of research deals with analyzing and capturing the structural features and patterns of student mobility across countries. Since international student mobility can be seen as a network of exchanges between higher education institutions (where the nodes are the universities/countries involved in the exchanges, and the links are the number of students going abroad for an exchange visit), the concept of networks can be easily applied in this context.

Several papers that used the social network analysis approach to understand the students' flows between countries have been published over the years. These papers aimed at i) detecting the presence of attractive countries by nodal centrality and authority scores (5); ii) identifying which countries can be called feeder and storer actors, i.e. good importers and good exporters (2; 12; 14); iii) revealing the presence of a core-periphery structure through block–modeling approach and clustering relational data (3; 14); iii) exploring the topology of the student mobility network by taking into account the exponential degree distribution and well–known network configurations (10); iv) identifying dense groups in the giant component using community detection algorithms (16); iv) assessing the determinants of student mobility patterns to test for the presence of homophily effects (17); and v) capturing the structural features and the patterns of student mobility across different countries by analyzing the factors pulling and pushing students to complete their higher education abroad (2).

Moreover, the gender gap in international students' mobility has been investigated by (8; 4) for emphasizing the presence of a denser network of connections involving females or engineering backgrounds, while the differences in the flows by disability, gender, and fields of study classified in stem and no-stem courses have been analyzed by (9).

In this manuscript, we focused on gender inequalities in STEM (Science, Technology, Engineering, and Mathematics) scientific fields. We postulate the following research question: "What are the gender-based inequalities in international students' mobility in STEM fields on the country level?"

## 2. Methodology

Methods of descriptive statistical analysis are supported using network analysis techniques. In this sense, we can define in a natural way a directed network with nodes representing countries in the Erasmus mobility network and directed edges by determined by the number of student mobilities among countries in the STEM fields.

Mathematically, weighted directed graph $\mathcal{G}$ is represented as $\mathcal{G}(\mathcal{C}, \mathcal{E}, \mathcal{W})$, where $\mathcal{C}$ represents the set of countries involved in the Erasmus program, $\mathcal{E} \subseteq \mathcal{C} \times \mathcal{C}$ is the set of edges given by the presence of students within STEM fields moving from one country to another, and $\mathcal{W}$ is the set of weights.

For any edge, $(c_i, c_j)$ in $\mathcal{E}$, the weight $w_{ij}$ is the number of students within STEM fields moving from country $c_i$ to the country $c_j$, where $c_i$ and $c_j$ are elements of $\mathcal{C}$. Therefore, the number of students involved in each exchange defines the edge weights in $\mathcal{W}$.

The population, and consequently the number of students, in countries involved in the Erasmus program, varies greatly, from the smallest country, Liechtenstein, to the largest, Germany. For this reason, the normalization of student mobilities between two countries (edge weights) was considered. Not only sending country as considered in previous studies on similar topic, e.g. (17), but also receiving country impact the number of students mobilities between two countries. Therefore, each weight was divided by a square root of a product of enrolled students among adjacent countries (for each country a mean value of enrolled students was calculated in the analyzed period). This is actually normalisation by geometric mean of number of students in adjacent countries and similar approach was used by (3; 5).

In order to identify the most important exporters and importers of students in STEM fields, hubs and authorities algorithm (11) was applied on two normalized networks, female and male student mobility networks of STEM fields.

DRAFT

## 3. Data

Data on international student flows among Erasmus and Erasmus+ countries are freely accessible and downloadable from the official EC website on Erasmus-Statistics.[3] The data cover the academic years from 2007–08 to 2013–14 and from 2014–2015 to 2018–2019, for Erasmus and for Erasmus+, respectively.

Given that some changes occur in the student mobility scheme between Erasmus and Erasmus+ programs, the datasets have been compared by checking the variables, in order to make them comparable. Since the datasets for Erasmus+ also contain information for all mobility participants (students and staff: study exchanges and work placements for students, and teaching assignments and staff training), we discard the data related to traineeships and staff mobility, focusing the analysis on the student mobility for studies abroad.

Table 1 shows the trend in student mobility for studies (SMS) in both programs, according to gender and stem courses. It is particularly evident that there is an upward trend in SMS for both programs. Moreover, the proportion of females and stem courses is constant over the years.

Table 1: Distribution of student mobility in Erasmus and Erasmus+ programs

| Erasmus | | | | |
|---|---|---|---|---|
| Academic Year | Total no. | SMS | % Female | % Stem |
| 2007–2008 | 182,697 | 162,694 | 61.15 | 12.55 |
| 2008–2009 | 198,523 | 168,193 | 60.63 | 13.91 |
| 2009–2010 | 213,266 | 177,705 | 60.86 | 10.80 |
| 2010–2011 | 231,408 | 190,495 | 60.86 | 11.00 |
| 2011–2012 | 252,827 | 204,744 | 60.61 | 12.79 |
| 2012–2013 | 268,143 | 212,522 | 60.64 | 13.12 |
| 2013–2014 | 272,497 | 212,208 | 60.22 | 13.11 |
| Erasmus + | | | | |
| Academic Year | Total no. | SMS | % Female | % Stem |
| 2014–2015 | 299,319 | 221,583 | 59.37 | 12.24 |
| 2015–2016 | 300,018 | 215,828 | 60.22 | 12.77 |
| 2016–2017 | 325,755 | 236,892 | 60.08 | 13.03 |
| 2017–2018 | 340,100 | 244,320 | 59.89 | 13.44 |
| 2018–2019 | 351,682 | 248,165 | 60.43 | 13.85 |
| 2019–2020 | 328,739 | 250,890 | 60.26 | 14.16 |

DRAFT

From Table 2, we can observe that the mobility ratio for non-STEM and STEM students differs between male and female groups. Every fifth male student on Erasmus mobility is from a STEM field, which is more than twice the proportion compared to the female group.

Table 2: Contingency table of Erasmus student mobility regarding gender and (non)STEM fields

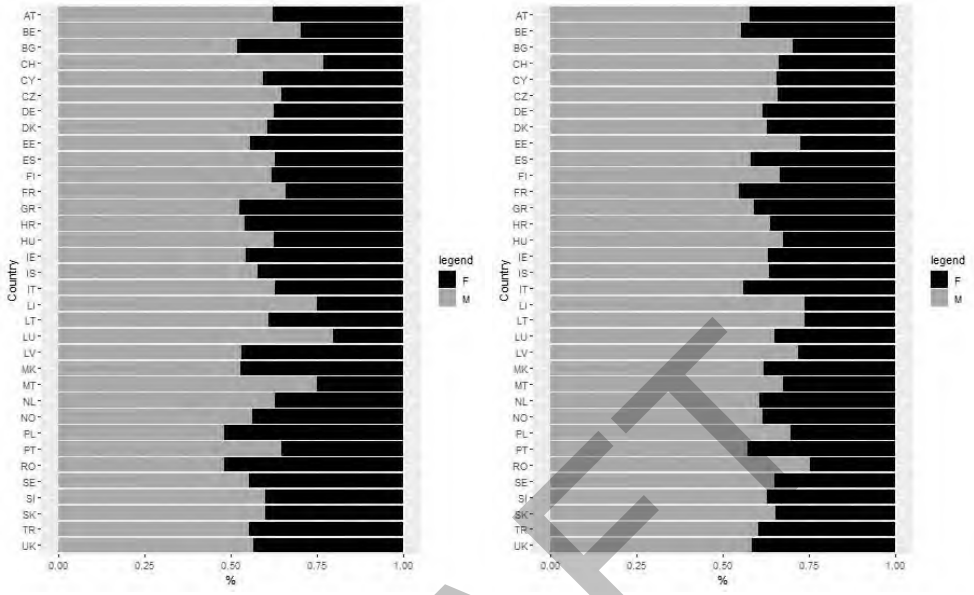
| | STEM | | |
|---|---|---|---|
| Gender | Yes | No | Total |
| Female | 129,303 (08.21%) | 1,445,381 (91.79%) | 1,574,684 (100.00%) |
| Male | 206,074 (20.08%) | 820,444 (79.92%) | 1,026,518 (100.00%) |
| Total | 335,377 (12.89%) | 2,265,825 (87.11%) | 2,601,202 (100.00%) |

Before merging the datasets, a preliminary analysis is performed to check whether the labels differ from one dataset to another and convert all codes of the field of education (ISCED6 1997–2011–2013) into the more recent classification.

[3] For details see https://data.europa.eu/euodp/en/data/publisher/eac.

## 4. Main results and discussion

The distribution of Erasmus and Eramsus+ flow by gender is shown in Figure 1 for outgoing (a) and incoming (b) students, where only the STEM courses are considered. For almost all countries, the percentage of males is higher than that of females. In a few countries, the percentage of men and women is substantially equal.




(a) Outgoing students (b) Incoming students

Figure 1: The distribution of Erasmus and Erasmus+ students (outgoing and incoming) for STEM courses and between males and females for all countries

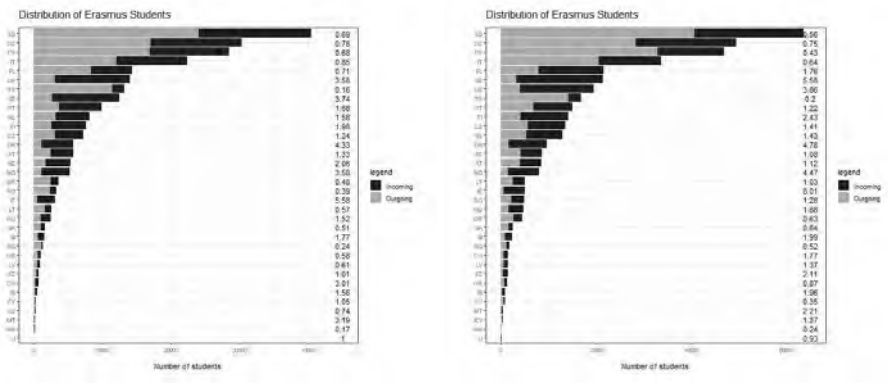
Figure 2 shows the distribution of Erasmus and Erasmus+ students by country over the twelve academic years for STEM courses, distinguished by gender. Spain, Germany, France, and Italy are very attractive for STEM being the top destinations for both incoming and outgoing students. Spain in particular has the highest number of students in terms of both incoming and outgoing exchanges. Germany and France rank second and third for incoming and outgoing students. Italy is in fourth place for both incoming and outgoing students. This is true for both females (Figure 2 a) and males (Figure 2 b).

In addition, Figure 2 reports the ratio of incoming to outgoing students for females (a) and males (b), which has a similar role to the coverage ratio used in trade networks to analyze the trade balance of foreign countries. Values greater than one indicate the attractiveness of countries with more incoming compared to outgoing students. The Scandinavian countries (Sweden, Norway, and Denmark) present the highest ratio values, while the United Kingdom appears to be the most attractive. Spain, France, Germany, and Italy seem to be slightly more export-oriented.

We applied the hubs and authorities algorithm on normalized networks to ensure comparability among countries in terms of their size and the number of students. When comparing hub (authority) scores between countries, the further away they are from the coordinate origin, the stronger their hub (authority) score tends to be. Moreover, countries located above the dashed line demonstrate a higher hub (authority) value for male STEM students, while those located below the dashed line exhibit a higher hub (authority) value for female STEM students.

In Figure 3 a) hub scores of female STEM students is compared with their male counterparts. The most active exporters of STEM students are Spain, France, Germany, and Italy which coincides with results of whole Erasmus student mobility (5). The highest difference in favor of male STEM students is observed for France, while Spain is slightly favoring female STEM students. Among other, not-so-active exporters, the results of Poland and Turkey are also in the favor of female STEM students.

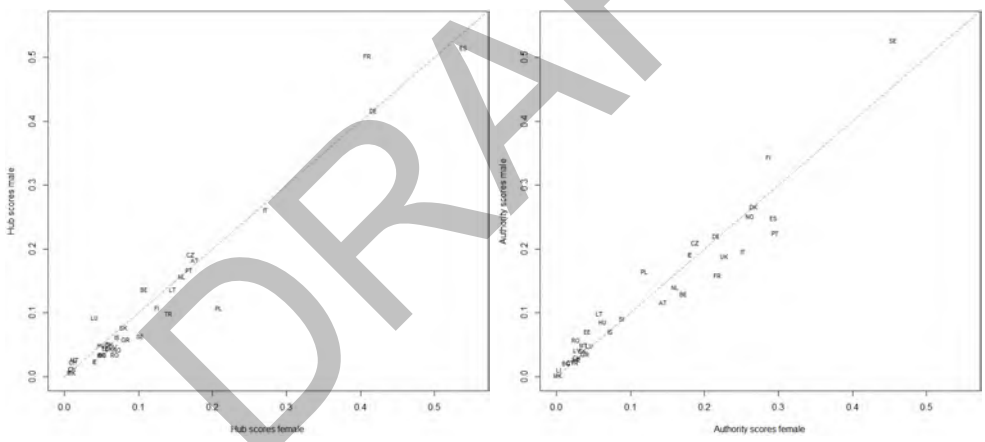
DRAFT




(a) Female students in STEM courses (b) Male students in STEM courses

Figure 2: The distribution of Erasmus and Erasmus+ students (incoming and outgoing) for all countries. The ratio between incoming and outgoing students is shown at the end of each bar.

Completely different results are obtained for authority scores in Figure 3 b), measuring the import of STEM students. The most active countries are from Scandinavia, with Sweden dominating, followed by Finland, Denmark, and Norway. The latter two are on the same level as Spain and Portugal. On one hand, male STEM students are more commonly visiting Sweden and Finland, on the other hand, female STEM students find Spain and Portugal more popular.




(a) Hub scores (b) Authority scores

Figure 3: Hubs and Authorities results compared between female and male Erasmus students per countries

## 5. Conclusion

Differences in Erasmus student mobility among STEM students with respect to gender have been observed. In relative terms, male STEM students are more inclined towards Erasmus mobility than their female counterparts. Results of descriptive statistics were further supported by hubs and authorities algorithm which takes into account also the relative importance of countries in terms of sending and receiving Erasmus students.

Countries with well-established traditions in Erasmus student mobility also happen to be the largest exporters of STEM students. However, Scandinavian countries are more attractive to STEM mobility students as they import a relatively larger number of such students. Sweden and Finland tend to import

DRAFT

more male STEM students, while Mediterranean countries such as Spain, Portugal, and Italy tend to be more attractive to female STEM students. With regard to exporting, France is dominated by male Erasmus STEM students.

In the future, we can compare the distributions of female and male students across different STEM fields. An obvious extension would be to analyze student mobility at the institutional level and identify the most active educational institutions in terms of student mobility. Additionally, identifying factors that contribute to gender-based inequalities would be a challenging task.

## References


[1] Amendola, A., & Restaino, M.: An evaluation study on students' international mobility experience. Quality & Quantity, 51(2), 525–544 (2017)

[2] Barnett, G.A., Ke Jiang, M.L., Park, H.W.: The flow of international students from a macro perspective: a network analysis. Compare: A Journal of Comparative and International Education, 46(4), 533–559 (2016).

[3] Breznik, K., Ragozini, G.: Exploring the Italian Erasmus agreements by a network analysis perspective. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 837–838 (2015)

[4] Breznik, K.: Institutional network of engineering students in the Erasmus programme. Global journal of engineering education, 19(1), 36–41 (2017)

[5] Breznik, K., & Skrbinjek, V.: Erasmus student mobility flows. European Journal of Education, 55(1), 105–117 (2020)

[6] Bryła, P.: International student mobility and subsequent migration: the case of Poland. *Studies in Higher Education*, *44(8)*, 1386–1399 (2019)

[7] Dabasi-Halász, Z., Kiss, J., Manafi, I., Marinescu, D. E., Lipták, K., Roman, M., & Lorenzo-Rodriguez, J.: International youth mobility in Eastern and Western Europe - the case of the Erasmus+ programme. Migration Letters, 16(1), 61–72 (2019)

[8] De Benedictis, L., Leoni, S.: Gender bias in the Erasmus network of universities. Applied Network Science, 5(1), 1–25. https://doi.org/10.1007/s41109-020-00297-9 (2020)

[9] De Benedictis, L., & Leoni, S.: Inclusive universities: evidence from the Erasmus program. Applied Network Science, 6(1), 1–21 (2021)

[10] Derzsi, A., Derzsy, N., Káptalan, E., & Néda, Z.: Topology of the Erasmus student mobility network. Physica A: Statistical Mechanics and its Applications, 390(13), 2601–2610 (2011)

[11] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46(5)*, 604–632. https://doi.org/10.1145/324133.324140

[12] Kondakci, Y., Bedenlier, S., Zawacki-Richter, O.:Social network analysis of international student mobility: uncovering the rise of regional hubs. Higher Education, 75(3), 517–535 (2018)

[13] Perez-Encinas, A., Rodriguez-Pomeda, J., & de Wit, H.: Factors influencing student mobility: a comparative European study. Studies in Higher Education, 46(12), 2528–2541 (2021)

[14] Restaino, M., Vitale, M. P., & Primerano, I.: Analysing international student mobility flows in higher education: a comparative study on European countries. Social Indicators Research, 149(3), 947–965 (2020)

[15] Roy, A., Newman, A., Ellenberger, T., & Pyman, A.: Outcomes of international student mobility programs: A systematic review and agenda for future research. Studies in Higher Education, 44(9), 1630–1644 (2019)

[16] Savić, M., Ivanović, M., Putnik, Z., Tütüncü, K., Budimac, Z., Smrikarova, S., & Smrikarov, A.: Analysis of ERASMUS staff and student mobility network within a big European project. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 613–618 (2017)

[17] Vögtle, Eva M., & Windzio, M.: Networks of international student mobility: enlargement and consolidation of the European transnational education space?. *Higher Education*, *72*, 723–741 (2016)


DRAFT

# A new measure of discrimination in machine learning algorithms

Roberta Pappadà[a] and Francesco Pauli[a]

[a]Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'B. de Finetti', Università degli Studi di Trieste (Italy); rpappada@units.it,francesco.pauli@deams.units.it

**Abstract**

Machine learning algorithms are routinely used for business decisions that may directly affect individuals in various contexts, such as credit scoring, employment, and criminal justice. When such algorithms are used in the decision process, their behavior concerning discrimination depends on the information it is given, and discrimination may occur unconsciously or explicitly based on sensitive attributes (e.g., gender and ethnicity). Statistical tools and methods are then required to handle such potential biases. We propose a sequential procedure based on Coarsened Exact Matching (CEM), which allows us to obtain a discrimination measure for all units. Different simulation scenarios are considered to check the proposed method's performance. Finally, we present an application to data of historical decisions and compare the proposed methodology with existing proposals.

***Keywords:*** discrimination, machine learning, Coarsened Exact Matching, sensible attribute

DRAFT

## 1. Introduction

Nowadays, automated decision-making occurs in many different contexts, representing a huge open scientific challenge due to the acknowledged implications on individuals. Examples of contexts in which such algorithms are currently used include credit scoring for loan decisions, screening applicants for hiring decisions, and criminal sentencing in the judicial system. This makes automated decision procedures relevant from a social, ethical, and legal point of view, opening a large debate on the inherent risks related to the lack of transparency, accountability, and fairness. Even when the procedure is not fully automatic, the suggestion coming from an algorithm is rarely ignored, being considered objective, and the human taking the decision may not fully understand the underlying algorithmic process (see, e.g., (2) and the references therein).

Automated decision procedures are based on machine learning algorithms that use individual information to build predictive models yielding an outcome or decision (e.g., whether to hire the individual for a job or to give them a loan). Such algorithms exploit a learning phase based on historical data, which may contain "neutral" attributes but also a sensitive attribute $S$ (e.g., ethnicity, gender, religion). In particular, if the sensitive attribute is available to the algorithm (i.e., it is included in the learning data and can be used for predictions), it may discriminate either because the data it is taught contain irrational prejudice or because the sensitive attribute is associated to an unobserved attribute that is relevant for the prediction of $Y$, the outcome of interest.

Our focus is on fairness assessment, which requires the evaluation of the equality of treatment of individuals (rate of a positive outcome) unconditionally or conditionally on some relevant characteristics different from $S$ which are deemed lawful to use to discriminate (see (3; 1)). This, in turn, may require different strategies to guarantee the protection of individuals against (direct or indirect) discrimination

based on membership to a category, group, or minority, without regard to individual features (8). Not including $S$ in the algorithm is insufficient to prevent discrimination. Indeed, assuming $S$ is not explicitly considered, and that discrimination is not intentional, an algorithm may be unfair if the data it learns upon contain discrimination.

Depending on the level at which action is taken, strategies for discrimination discovery and prevention aim to (i) modify the learning data (ii) integrate the learning algorithm with a non-discrimination objective, (iii) tamper with the final algorithm after it has been fitted, and (iv) change the final predictions.

We focus on data pre-processing techniques, i.e., statistical methods that aim to establish whether and to what extent the available data are discriminatory. If discrimination is detected, data are modified to make them discrimination-free (6). In this framework, the authors in (7) propose a method of discrimination discovery that allows labeling each tuple in a dataset as discriminated or not, based on a measure of the difference of treatment among its neighbors belonging to a protected-by-law group and its neighbors not belonging to it. The methodology in (7) is based on the $k$ nearest neighbors classification method ($k$-NN), which uses a suitable distance to search for neighbors of a given tuple.

Here, we propose a sequential procedure based on Coarsened Exact Matching (CEM), a technique from causal inference that aims to estimate the causal effect of a treatment on an outcome (5; 4). In this context, the units labeled as treated are individuals belonging to a protected category, and (conditional) discrimination is detected by comparing a protected unit's outcome with the outcome of unprotected units exhibiting similar values on the observed variables different from $S$. Discrimination prevention is tackled by assigning each unit a discrimination score, which can be used to derive a discrimination-free dataset in the pre-processing step.

This paper is organized as follows. In Sect. 2, we review the framework and present the proposed CEM-based discrimination measure. The sequential procedure based on CEM is applied in Sect. 3. to discrimination detection, highlighting the comparison with the approach proposed in (7). Finally, Sect. 4. summarizes the contributions of the paper.

DRAFT

## 2. The discrimination measure

This section presents our proposal for detecting discrimination in the data. Let $Y$ represent the dichotomous decision of a machine learning algorithm on individuals, with $Y \in \{0, 1\}$ and $Y = 1$ being the desirable outcome. Given the individual attributes $\mathbf{X} = (X_1, \ldots, X_d)$, equal treatment conditional on $\mathbf{X}$ requires

$$P(Y = 1|S = 1, \mathbf{X} = \mathbf{x}) = P(Y = 1|S = 0, \mathbf{X} = \mathbf{x}),$$

where $S = 1$ is "belonging to a protected category", i.e., possessing the sensible attribute (s.a.). Hence, the kind of discrimination we refer to occurs when the collective possessing the s.a. experiences the desired outcome less frequently than those not possessing the s.a. ($S = 0$) but similar in other respects. In Fig. 1-(a) two continuous neutral attributes $x_1, x_2$ are represented on a plane, and units are distinguished based on $S$ and $Y$. To measure the potential discrimination against unit $i$, possessing the s.a., it is reasonable to compare it to its neighbors. For example, the authors in (7) propose using the difference in the rate of unfavorable outcomes in units with $S = 1$ (black triangles) and units with $S = 0$ (black circles), i.e., non-sensitive ones, among the $k$ nearest neighbors of $i$ (see Fig 1-(b)).

Our approach to measuring the potential discrimination against unit $i$ (having the s.a.) is based on comparing the outcome $y_i$ with the outcome of the units matched by the CEM technique. The CEM algorithm matches a treated unit to those untreated units that share the same values on the available (continuous and categorical) attributes, possibly after coarsening the continuous ones. Moreover, the CEM algorithm discards those units that can not be matched.

Assume that for unit $i$ we have $S = 1$. Let $\bar{y}_i^{(S=0)}$ be the relative frequency of positive outcomes among units matched with $i$ and not possessing the s.a., that is, $\bar{y}_i^{(S=0)} = \#\{i|y_i = 1, s_i = 0, i \in C_i\}/\#\{i|s_i = 0, i \in C_i\}$, where $\#A$ is the cardinality of a set A. We define

$$D_i = y_i - \bar{y}_i^{(S=0)} \tag{1}$$

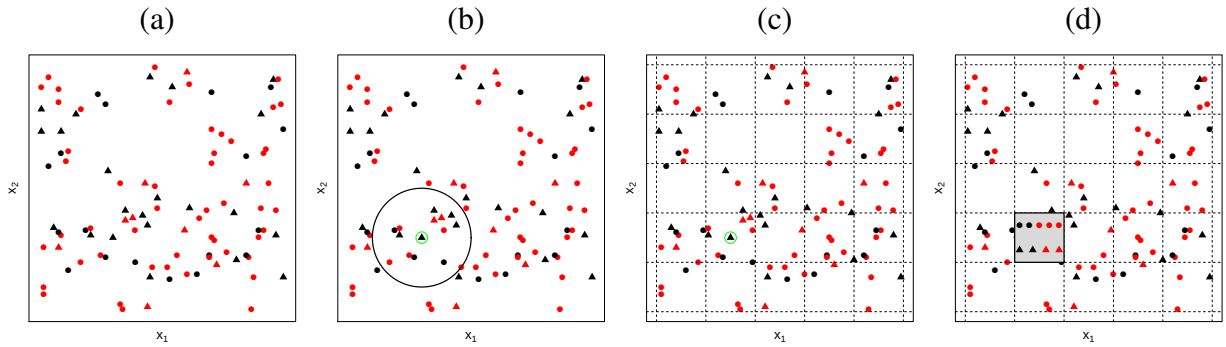



Figure 1: Two continuous neutral attributes $x_1$ and $x_2$ represented on a plane. (a) Units are marked with red circles ($S = 0$, $Y = 1$), red triangles ($S = 1$, $Y = 1$), black circles ($S = 0$, $Y = 0$), black triangles ($S = 1$, $Y = 0$). (b) A single unit (green circle) with $Y = 0$ and $S = 1$ and its $k$ nearest neighbors. (c) A single unit (green circle) with $Y = 0$ and $S = 1$, and (d) units matched by CEM for computing the discrimination measure as in Eq. (1).

as a measure of discrimination for unit $i$, which takes negative values when the unit $i$ is discriminated against and positive values when the unit $i$ is favored ($D_i \in [-1, 1]$). However, not all units are matched in a standard implementation of the CEM algorithm, making it unsuitable for our purpose.

A possible strategy to exploit the CEM technique is to apply it sequentially as follows.

**Step 0** Unit $i$ is compared to all other units and $D_i^{(0)}$ is computed as

$$D_i^{(0)} = y_i - \bar{y} = y_i - \hat{P}\{Y = 1|S = 0\} = y_i - \frac{\#\{i|y_i = 1, s_i = 0\}}{\#\{i|s_i = 0\}},$$

that is, no matching is performed.

**Step 1** Variable $X_1$ is considered in matching, and let $C_i^{(1)}$ be the set of units matched with unit $i$ based on $X_1$. Then, if $C_i^{(1)} \cap \bar{\mathcal{S}} \neq \emptyset$,

$$D_i^{(1)} = y_i - \bar{y}_i = y_i - \frac{\#\{i|y_i = 1, s_i = 0, i \in C_i^{(1)}\}}{\#\{i|s_i = 0, i \in C_i^{(1)}\}},$$

otherwise, $D_i^{(1)} = D_i^{(0)}$ (STOP).

**Step 2** Variable $X_1$ and $X_2$ are considered in matching, and let $C_i^{(2)}$ be the set of matched units based on $X_1, X_2$. Then, if a non-empty subset $C_i^{(2)}$ of similar units based on $X_1, X_2$ has been found and $C_i^{(2)} \cap \bar{\mathcal{S}} \neq \emptyset$, set $D_i^{(2)} = y_i - \bar{y}_i$; otherwise, $D_i^{(2)} = D_i^{(1)}$.

The procedure is continued until either a STOP is reached or all variables have been used.

A measure $D_i$ is hence obtained for each unit $i$ in the dataset. Note that, unlike the method based on $k$-NN, the sequential CEM procedure does not rely on a distance measure to identify similar units. Since the final discrimination score $D_i$ depends on the order in which the variables $X_1, \ldots, X_K$ are used, we repeat the sequential CEM procedure over $M$ random permutations $i_1, \ldots, i_K$ of $1, \ldots, K$. Finally, the discrimination scores are computed as $\tilde{D}_i = (1/M) \sum_{i_1,\ldots,i_K} D_i^{(i_1,\ldots,i_K)}$, where $D_i^{(i_1,\ldots,i_K)} = D_i^{(K)}$.

## 3. Testing CEM-based measure on COMPAS score

In this section, we explore the performance of the proposed measure of discrimination and compare it with the alternative method proposed in (7). The data we use consist of historical decision records concerning inmates' risk of criminal recidivism. Specifically, we use data available from the public

DRAFT

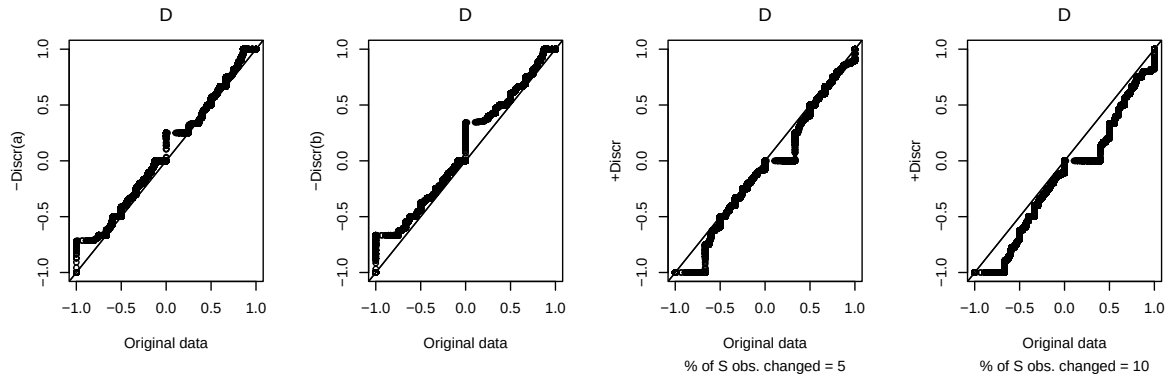



Figure 2: Q-Q plots for discrimination scores $D$ for the discrimination-free data and the data with added discrimination against the original COMPAS data. Discrimination is removed by modifying the outcome (-Discr(a)) or by modifying the sensitive attribute (-Discr(b)); discrimination is added by modifying the outcome with $v_1 = v_2 = 5\%$ and $v_1 = v_2 = 10\%$, as described in Sect. 3.

database of Parolees Under-Community-Supervision for the State of New York, which is based on the COMPAS score, a predictive model employed to assign a specific level of supervision (on a scale from one–highest–to four–lowest). Recently, the journalists of *propublica.org* have argued that the COMPAS score has a strong ethnic bias, showing that "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend"[1].

To test the CEM-based discrimination measure on COMPAS score, we analyzed 40,704 observations related to 2018 and 2019. The data contain information on individuals' ethnicity, where non-white people represent 61.8% of units. We assume $S = 1$ "being non-white", and the desired outcome ($Y = 1$) is the fourth level of supervision, requiring two face-to-face contacts every four months. The other variables are the region of supervision, the county of residence, the gender, the age, and the inmate's crime type. The dataset is split into learning (32,488 units) and test (8,216 units) subsamples for the analysis.

To test the sensitivity of the proposed measure to the presence of potential bias in historical records, we simulate discrimination-free scenarios by following two alternative approaches:

(a) the outcome $Y$ is modified according to the classification tree estimate of the probability of a positive outcome based on all variables but the $S$;
(b) the value of $S$ is modified for all units via a random permutation on the sensitive attribute.

Moreover, we modify the original data by introducing discrimination for a subset of units. Specifically, we modify the dataset (i) by changing the outcome from positive to negative ($Y = 1 \to Y = 0$) for a fraction $v_1$ of units with $S = 1$ and $Y = 1$, and (ii) by changing the outcome from negative to positive ($Y = 0 \to Y = 1$) for a fraction $v_2$ of units with $S = 0$ and $Y = 0$. For instance, if $(v_1, v_2) = (2.5, 0)$, then the outcome is changed to negative for 2.5% of units in the protected category, thus introducing discrimination against such units.

To assess weather the proposed measure is able to detect the removed or added discrimination in the simulated scenarios, we compare the discrimination scores returned by the sequential CEM algorithm computed for the modified data against the original COMPAS data. Fig. 2 shows the Q-Q plots for the four scenarios arising from methods (a) and (b) used for removing discrimination as described above, and two different settings for introducing discrimination by modifying the outcome for some units. The different behavior of $D$ when discrimination is removed or added is clearly spotted for negative values, which indicate discrimination against the protected category.

We compare our approach with the measure based on $k$-NN in terms of Correct Prediction Ratio (CPR), that is, the fraction of units for which the added discrimination is correctly detected. Fig. 3 shows the resulting ratios for different combinations of $(v_1, v_2)$ and different thresholds for discrimination detection corresponding to five different quantiles of $\tilde{D}$.

[1] http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

DRAFT

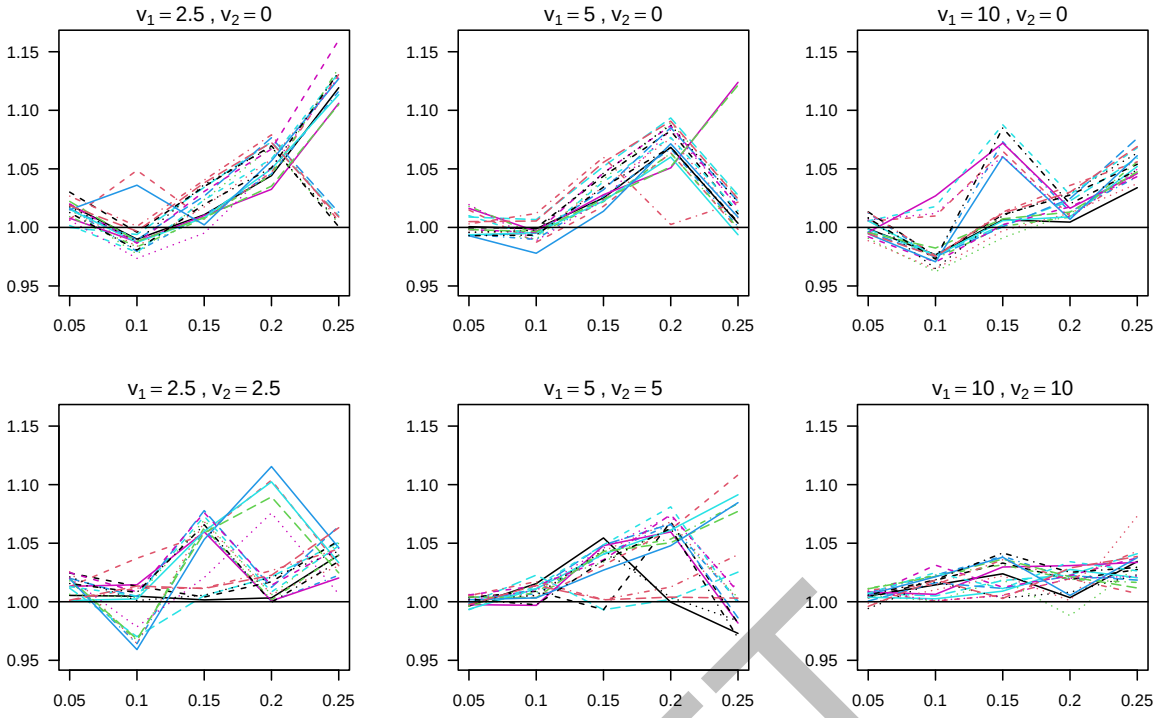



Figure 3: Correct prediction ratio (CPR) of Iterated CEM versus $k$-NN for COMPAS data: the $y$-axis displays the ratios CPR-CEM/CPR-KNN for $B = 20$ simulated scenarios and different fractions of units for which discrimination against ($v_1$) or in favor ($v_2$) is introduced. The $x$-axis displays the thresholds for discrimination detection corresponding to quantiles of $\tilde{D}$ quantiles.

The proposed measure shows an overall satisfactory performance compared to the competing method proposed by (7). Indeed, with few exceptions at low quantiles, the CEM-based method seems to outperform the $k$-NN approach.

## 4. Conclusions

We propose a sequential procedure based on the CEM technique that returns a discrimination score for each unit (at worst, it is based on the rate of a positive outcome for the whole non-sensitive sample).

Unlike existing proposals, the approach based on sequential CEM has the advantage to match the units with respect to the coarsened variables, rather than relying on an arbitrary distance measure. Hence, the proposed methodology easily allows for nominal variables. Moreover, by defining appropriate coarsened versions of the continuous variables (and possibly appropriate recoding of the nominal ones), it enforces the concept of similarity among the units.

The application to COMPAS score data shows that the CEM-based discrimination measure is suitable for detecting and measuring discrimination against a protected group, providing an overall satisfactory performance in terms of prediction accuracy.

## References

[1] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 259-268. ACM (2015)

[2] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of

DRAFT

methods for explaining black box models. ACM Computing Surveys (CSUR). **51**(5), 1–42 (2018).
[3] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29 (NIPS 2016), pp. 3315-3323 (2016)
[4] Iacus, S., King, G., Porro, G.: CEM: Software for Coarsened Exact Matching. Journal of Statistical Software **30**(13), 1–27 (2009)
[5] Iacus, S.M., King, G., Porro, G.: Multivariate matching methods that are monotonic imbalance bounding. Journal of the American Statistical Association **106**(493), 345–361 (2011)
[6] Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33(1), 1-33 (2012)
[7] Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 502-510. ACM (2011)
[8] Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowledge Engineering Review **29**(5), 582–638 (2014).

DRAFT

# Challenges on Ethics, and Privacy in AI Applications to Fintech

Catarina Silva, Joana Matos Dias, and Bernardete Ribeiro

University of Coimbra
catarina@dei.uc.pt, joana@fe.uc.pt, bribeiro@dei.uc.pt

**Abstract**

AI in Fintech presents several ethical and privacy challenges. One of the main concerns is the potential for algorithmic bias, where machine learning models may unintentionally perpetuate existing discrimination, such as by considering factors like race, gender, or age. Another issue is the difficulty in interpreting the decision-making process of these models, which can create accountability issues and distrust in financial institutions. Additionally, there are concerns about data privacy and security, as large amounts of sensitive financial data are being used to train and validate these models. In this paper we analyse what has been the dynamics of the research community in addressing these concerns. A preliminary bibliometric analysis is done, considering publications within the last 10 years. It is possible to conclude that many of these aspects are still not considered by most publications, and no significant researchers' networks exist that could leverage research in this area.

***Keywords:*** Ethics, Privacy, AI, Fintech.

DRAFT

## 1. Introduction

Nowadays, the issues of explainability and accountability of artificial intelligence algorithms are at the forefront of discussions. These concerns are particularly critical in the financial sector, where decisions based on AI outputs can have significant impacts on the lives of individuals and society as a whole. Numerous questions have been raised regarding this topic that merit the attention of the scientific community, namely (see Figure 1):

- **Governance**: Who should be accountable for the development and maintenance of AI models used for decision making? Who is responsible for evaluating and ensuring the quality of these models?
- **Policy development**: Who is responsible for creating policies and guidelines for the use of AI? How can the perspectives of different stakeholders be incorporated? Who are these stakeholders?
- **Cultural acceptance**: What steps are necessary to foster trust in AI models? How can concerns about societal impacts, such as job displacement, be addressed? How can inertia towards technological and cultural shifts be overcome?
- **Risk assessment**: How can risks associated with AI model building and use be measured, defined, and mitigated? How can we prevent fraud, manipulation, and cybercrime? How should national and international legislation address the risks and consequences of biased or misapplied AI models?

Figure 1: AI in Fintech Challenges

- **Model selection**: What criteria should be used to select AI models for different situations, given inherent conflicts between criteria such as accuracy and explainability?
- **Process and technology**: What platforms and technologies should be used for AI models? What process adjustments are necessary to fully leverage the benefits of AI?

## 2. Preliminary bibliometric analysis

To evaluate the progress in research on the aforementioned topics, we propose conducting a bibliometric analysis, which will be detailed in the following sections.

### 2.1 Bibliometric search

We considered all journal and conference papers published from 2017 to the end of February 2023, indexed in Scopus. Additionally, the search considered the following constraints:

- TITLE (finance OR fintech)
  AND TITLE ((artificial AND intelligence ) OR ( machine AND learning ))
  AND TITLE (ethics OR privacy OR explainability OR transparency))

- KEY ( finance OR fintech )
  AND KEY ( ( artificial AND intelligence ) OR ( machine AND learning ) )
  AND KEY ( ethics OR privacy OR explainability OR transparency) ) )

- LIMIT-TO ( DOCTYPE , ”ar” ) OR LIMIT-TO ( DOCTYPE , ”cp” ) )

This means that we have selected all publications that have, either in the title or in the keywords, the words ”finance” or ”fintech” and refer to AI or ML approaches. The title or keywords must explicitly consider one of these subjects: ethics, privacy, explainability or transparency.

The search considered the following query string:

```
( (TITLE ( finance OR fintech ) AND TITLE ( ( artificial AND
intelligence ) OR ( machine AND learning ) ) AND TITLE ( ethics OR
privacy OR explainability OR transparency) ) OR (KEY ( finance OR
fintech ) AND KEY ( ( artificial AND intelligence ) OR ( machine
AND learning ) ) AND KEY ( ethics OR privacy OR explainability OR
```

DRAFT

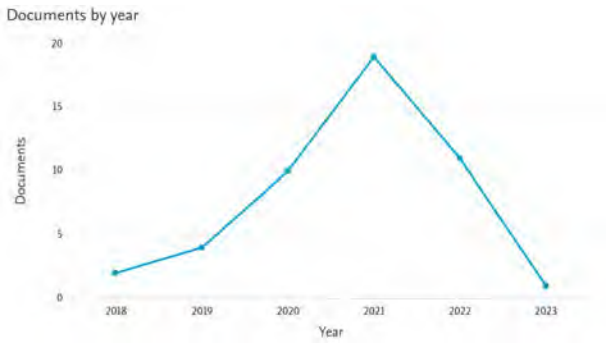



Figure 2: Documents per year (Source: Scopus)

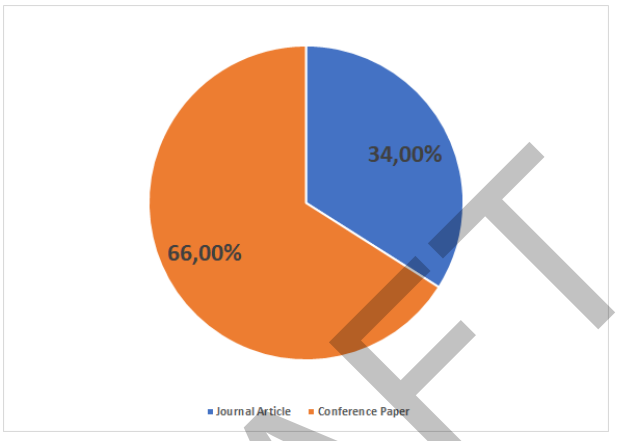



Figure 3: Publication Media (Source: Scopus)

DRAFT

```
transparency) ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO (
DOCTYPE , "cp" ) ) AND ( LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO (
PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR
, 2020 ) OR LIMIT-TO ( PUBYEAR , 2021 ) OR LIMIT-TO ( PUBYEAR , 2022
) OR LIMIT-TO ( PUBYEAR , 2023 ) )
```

Only 47 publications[1] were found that comply with this query, distributed per year as depicted in Figure 2. If the search is extended to also include the abstract, then this number increases to 264 publications. This is an interesting observation on its own: it can lead us to conclude that the ethical questions are not the central part or concern in many of these works, but they are somehow being taken into account in the research made. In this preliminary analysis, only the small set of 47 manuscripts will be considered. The enlarged dataset will be studied in future works. We have read the abstract of all the 47 papers, to assure that each one should be considered for further analysis. All of them were considered, even though some of them considered fintech alongside other AI applications. It is interesting to note that the year where more publications can be found is 2021. Most manuscripts have been published in conference proceedings, as depicted in Figure 3, which would be expected in a relatively recent area of research.

## 2.2 Authors network

There are a total of 146 authors associated with these manuscripts. On average, each author has participated in one manuscript only. China is the country with more contributions (15 in total). Looking at the

[1]The reference list of the publications considered can be found in https://drive.google.com/file/d/1q9dnzoR6otsiLrBuTZNHlDHz1VT_eqgv/view?usp=share_link

Figure 4: Authors' network (Source: Scopus)

set of authors, it can be seen that there are still no networking taking place: the coauthors of each article are not connected to other authors. So it seems that each research team is still working on its own, and no cooperation is in place in terms of published research. This can be visualized in Figure 4, when it is clear that there are no really connections among others associated with different manuscripts.

## 2.3 Keywords network

Figure 5 depicts the keywords network, where different colours represent different keyword clusters. This image considers keywords that were present at least 3 times (from the set of 578 different keywords associated with this publication set). There are four clusters that represent the most frequently used keywords. The red cluster considers manuscripts that are not addressing any particular problem, but rather consider in more general terms the use of AI and ML in finance, expliciting the ethic dimension of these works. The yellow cluster considers works dedicated to financial products (like cryptocurrencies), being also concerned with data security and protection. The green cluster is composed by keywords associated with financial services and the corresponding risk assessment and management. The blue cluster is linked with model development (learning algorithms, privacy by design). It is interesting to realize that, although ethics and data protection are presented (due to the query done to retrieve the chosen manuscripts), there are important keywords that are missing, related with legal regulation, explainability, accountability, for instance.

DRAFT

# 3. Conclusions

Although an increased concern with ethical questions is clearly emerging in the scientific community considering the application of AI in critical contexts, like finance, the number of manuscripts that consider, in its core, fintech and ethics, explainability, accountability are still scarce.

The recent proposals by the European Union, namely the Digital Services Act that aims at defining a common set of rules, obligations, and accountability across EU to provide all types of digital services, while ensuring a high level of protection to all users [2]. In this proposal, financial applications are considered high-risk applications subject to strict obligations before they can be put on the market[3]:

- adequate risk assessment and mitigation systems;
- high quality of the datasets feeding the system to minimise risks and discriminatory outcomes;

[2] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-enviro en

[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

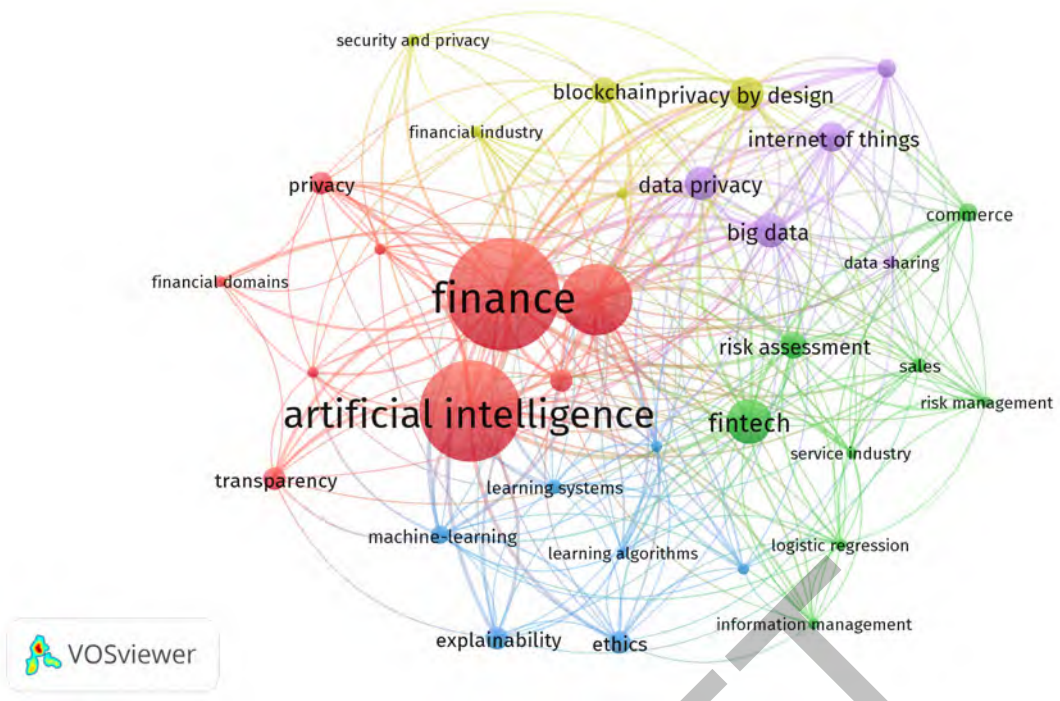



Figure 5: Publication Media (Source: Scopus)

- logging of activity to ensure traceability of results; detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance;
- clear and adequate information to the user;
- appropriate human oversight measures to minimise risk;
- high level of robustness, security and accuracy.

In this work we can conclude that it seems that there is a literature gap in what concerns the connection between law and AI, since no manuscripts are clearly concerned with algorithm development and application evolving alongside specific legislation creation and adoption. AI developers and researchers should be working together with regulators.

## 4. Acknowledgements

This work has been partly supported by Fundação para a Ciência e a Tecnologia (FCT) under project grants UIDB/00308/2020, UIDB/00326/2020. This work has been partly funded by the FCT – Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

## 5. References

For the list of all the publications considered in this work, please refer to https://drive.google.com/file/d/1q9dnzoR6otsiLrBuTZNHlDHz1VT_eqgv/view?usp=share_link.

DRAFT

# Uncertainty & fairness metrics

Anna Gottard[a]

[a]Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti"
Florence Center of Data Science - University of Florence; anna.gottard@unifi.it

**Abstract**

Nowadays, many decisions are made taking as suggestion predictive models based on observed data. Even if the learning process is fair and not malicious, predictive models may systematically discriminate against certain groups of people. The resulting decisions will be unfair. Fairness-aware statistical machine learning investigates how to create discrimination-free predictive models. We discuss the various source of unfairness as solutions to unfairness cannot be separated from its source. We provide a way to distinguish the lack of fairness due to the data generating process/sample from that induced by the predictive algorithm. In addition, we provide some simple strategies to evaluate uncertainty in fairness metrics for the large sample case. An illustrative example on synthetic data from a specific data generating process provides insights into the CART algorithm behaviour.

***Keywords:*** Fairness, Predictive modeling, Statistical machine learning, Uncertainty

DRAFT

## 1. Introduction

Decisions that affect individuals increasingly rely on statistical machine learning procedures. Algorithms support criminal justice decisions, credit banking, hiring practices, and personalized medicine. Fairness in machine learning has become a critical issue as the growing use of automated decision-making procedures has highlighted the potential for bias and discrimination. The definition of fairness is not univocal and depends on context. See, among many, (11), (12), (13) and (15) for detailed reviews on fairness.

Fairness is generally defined as the absence of discrimination in the outcome of an automated decision-making system. This includes discrimination against individuals or groups based on sensitive or protected characteristics, such as race, gender, or age, as well as discrimination against individuals or groups due to their socioeconomic status, or other factors.

We can distinguish three main situations that result in a lack of fairness.

1. *Algorithm bias*, when the algorithm induces bias about sensitive attributes,
2. *Sampling bias*, when the data set is biased as a result of an unfair selection mechanism,
3. *Population bias*, when Nature is unfair.

While these situations may occur simultaneously, each recalls a specific definition of fairness and requires unique solutions. It is crucial to recognize these possible sources of unfairness and take steps to mitigate them when building and deploying statistical machine learning algorithms.

The first source of unfairness considered is algorithmic bias. This type of bias arises when a model is developed to make predictions based on certain features or variables that are not directly relevant and result in discrimination against certain groups or individuals. For instance, if a model is designed to make predictions based on an individual's race or gender, it could lead to the unjust treatment of particular groups.

The second source of unfairness arises when the data used to train the model is not representative of the population for which the model is intended. This can occur due to selection effects, such as when data are predominantly drawn from a particular demographic group or race, leading to biased and unfair treatment of under-represented groups or individuals. Machine learning algorithms assume that data are an *iid* sample representative of the population and are thus unsuitable for handling selection effect. Additionally, a dataset can be discriminative because it is drawn from a population affected by population bias. This brings us to the third source of bias.

The third situation is particularly intriguing, because it involves unfairness and discrimination in the population itself. If Nature is unfair, every random sample drawn from the population of interest will be unfair. Addressing fairness issues in this type of situation requires the algorithm to move away from the true data generating process and produce predictions for a better world, intentionally biased to counter the effects of Nature's unfairness. This can be achieved by adjusting the dataset or modifying the loss function of the predictive algorithm. Some authors treat fairness issues as a problem of imbalanced data, as discussed in (7).

Consider, for instance, tree-based regression and classification algorithms, a popular class of predictive models known for their simplicity and ability to handle both nonlinear and interaction effects. The most widely used algorithm is the Classification and Regression Tree (CART) algorithm (4). Despite not having the highest predictive performance, CART is valued for its transparency and ease of interpretation. However, fairness can be problematic in classification trees that use greedy search. Not only can CART amplify bias present in the dataset or population, but it can also introduce unfairness in predictions. According to (9), tree-based algorithms that use greedy search tend to rely on background variables, typically including sensitive attributes, while ignoring some variables that directly affect the response. This issue is propagated to ensemble methods based on CART.

The purpose of this paper is two-fold. The first objective is to contribute to the ongoing effort of quantifying uncertainty in fairness metrics. We propose a straightforward method for assessing the level of uncertainty in fairness metrics. Prior research studies, such as (10) and (6), have observed that fairness metrics tend to exhibit inconsistency when applied to different train-test splits of the data. We argue that these metrics should be viewed as estimates of the *true* metric value, and that sample variability can be accounted for using simple statistical procedures. At this aim, confidence intervals can be computed for the true level of fairness measured with a given metric. The importance of confidence intervals for fairness metrics has been also highlighted by (3). The second objective is to provide a procedure to identify the source of unfairness by comparing simple intervals of fairness metrics that can be easily computed on subsets of the observed sample. In particular, we will examine the metric of statistical parity, which measures the difference in proportion of positive outcomes between different groups. By comparing the confidence intervals of this metric across training and test data, we can gain insights into the factors that contribute to unfairness. The approach presented in this paper can be easily implemented in a wide range of contexts, making it a valuable tool for researchers and practitioners alike.

The remainder of this paper is organized as follows. Section 2. discusses statistical parity as a fairness metric and its estimator. Section 3. provides two confidence intervals for evaluating the uncertainty of such a metric as well as a statistical procedure for comparing unfairness. A synthetic data example is included. Finally, concluding remarks are provided in Section 4.

## 2. A measures of fairness: statistical parity

Consider a set of explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)$, and a response $Y$ that, for simplicity, we assume to be binary, $\mathcal{Y} = \{0, 1\}$. Suppose we are given a random sample $\mathcal{D} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ of units independently drawn from a distribution $F_{XY}$ having domain $\mathcal{X} \times \mathcal{Y}$.

A *classifier* $h$ can be formally defined as a mapping function $h : \mathcal{X} \to \mathcal{Y}$, that assigns a class label $y \in \mathcal{Y}$ to a given set of values $\boldsymbol{x} \in \mathcal{X}$. The mapping function $h(\boldsymbol{X})$ is learned by minimizing a loss function that is typically a function of the classification risk, or *error rate*, $R(Y, h(\boldsymbol{X})) := \mathbb{P}(Y \neq h(\boldsymbol{X}))$. In the binary case, it can be proved that the classifier minimizing the classification risk depends

DRAFT

on the conditional expectation

$$m(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}),$$

where a unit is assigned to the class 1 whenever $m(\boldsymbol{x}) > \frac{1}{2}$. As a consequence, a good classifier is the one providing an accurate approximation of $m(\boldsymbol{x})$, say $\widehat{m(\boldsymbol{x})}$, the estimate of the conditional probability $\pi_{Y|X}$ and $h(\boldsymbol{X})$ assigns $\widehat{Y} = 1$ whenever $\widehat{m(\boldsymbol{x})} > \frac{1}{2}$. The algorithm CART also adopts this kind of classifier as it utilizes a *majority vote* prediction rule. A classifier is typically trained on a subset of data called *training data*, while the rest of the data, called *test data*, are used to assess prediction accuracy.

Let $S$ represent the group attribute to be protected, such as, for instance, gender or race. It can or cannot be included within the set of explanatory variables $\boldsymbol{X}$, as a subject-specific choice. Notice that if one is interested in more than one sensitive attribute, $S$ can be defined on the cross-product of the sensitive attributes. For simplicity, we assume $S$ to be binary.

Several fairness measures have been proposed in the literature due to the increased interest in fair machine learning. For brevity, we focus here on one specific metric, statistical parity, but the consideration provided can be easily extended to other measures, such as those proposed by (12).

**Definition 1.** ***Statistical parity*** *is a group fairness metric that requires the distribution of positive predictions to be equal across the different levels of the sensitive variable $S$, taking values in $\mathcal{S}$. It can be defined as*

$$\mathbb{P}\Big(Y = 1 \mid S = s\Big) \;=\; \mathbb{P}\Big(Y = 1\Big) \qquad \forall s \in \mathcal{S}. \tag{1}$$

*to measure unfairness in the sample or the population, and*

$$\mathbb{P}\Big(\widehat{Y} = 1 \mid S = s\Big) \;=\; \mathbb{P}\Big(\widehat{Y} = 1\Big) \qquad \forall s \in \mathcal{S}. \tag{2}$$

*to measure unfairness in the prediction.*

The metric (1) may serve as an initial method for detecting unfairness in data, whereas the metric (2) can be used to assess fairness in the predicted values. When $Y$ is binary, statistical parity, also known as demographic parity, implies that $Y \perp\!\!\!\perp S$ for (1) or $\widehat{Y} \perp\!\!\!\perp S$ with equation (2). When $Y$ is multinomial, statistical parity is achieved if $\mathbb{P}(Y = y \mid S = s) = \mathbb{P}(Y = y)$ or $\mathbb{P}(\widehat{Y} = y \mid S = s\Big) = \mathbb{P}(\widehat{Y} = y)$, for each $y \in \mathcal{Y}$ and $s \in \mathcal{S}$. Hereafter, we assume that both Y and S are binary variables, with $\mathcal{Y} = \{0, 1\}$ and $\mathcal{S} = \{s, \overline{s}\}$.

Statistical parity considers the conditional distribution of $Y$ given $S$ once marginalized over $\boldsymbol{X}$. As a consequence, it is subject to the Simpson paradox (14). It can happen that, even if (1) and (2) are satisfied, there might exist some discrimination for specific sub-populations, conditional on $\boldsymbol{X}$. On the other hand, if also $Y \perp\!\!\!\perp S \mid \boldsymbol{x}$ for each $\boldsymbol{x} \in \mathcal{X}$, then there is no context-specific unfairness that cancels out when marginalizing over $\boldsymbol{X}$. Other fairness metrics are conditional on $X$ and, therefore, do not have this issue. However, they require further assumptions on the dependence of $Y$ on $\boldsymbol{X}$.

A metric for violation of statistical parity is therefore

$$\Delta = \mathbb{P}\Big(Y = 1 \mid S = s\Big) \;-\; \mathbb{P}\Big(Y = 1 \mid S = \overline{s}\Big).$$

for the population, and

$$\Delta_P = \mathbb{P}\Big(\widehat{Y} = 1 \mid S = s\Big) \;-\; \mathbb{P}\Big(\widehat{Y} = 1 \mid S = \overline{s}\Big)$$

for the prediction. These quantities are actually unknown and the values computed from observed data are only estimates of these metrics. Training data can be used to compute a consistent estimate $\widehat{\Delta}$ of the population's lack of statistical parity, as it is a simply the difference of conditional probabilities. An *honest* estimator $\widehat{\Delta}_p$ can be obtained by the sample proportions computed from the test data set, used neither for learning nor tuning the parameter of the classifier,

$$\widehat{\Delta}_P = \frac{1}{n_{\text{test}|s}} \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{\widehat{Y}_i = 1\} \cdot \mathbb{I}\{S_i = s\} \;-\; \frac{1}{n_{\text{test}} - n_{\text{test}|s}} \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{\widehat{Y}_i = 1\} \cdot \mathbb{I}\{S_i = \overline{s}\}.$$

DRAFT

where $n_{\text{test}}$ is the dimension of the test set, with $n_{\text{train}} = n - n_{\text{test}}$, and with

$$n_{\text{test}|s} = \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{S_i = s\}.$$

Large values of $\widehat{\Delta}$ can be attributed to both sample and population unfairness. Unfortunately, no metrics or test statistics on the observed data can disentangle a lack of fairness due to population bias from sampling bias. Large values of $\widehat{\Delta}_P$ when $\widehat{\Delta}$ is close to zero may suggest algorithm bias. As $\widehat{\Delta}$ and $\widehat{\Delta}_P$ are simply an estimate of the *true* statistical parity measures, it is essential to consider the uncertainty in the estimate of such metrics.

## 3. Uncertainty evaluation

To check the source of unfairness and to evaluate their uncertainty, one can think of several statistical tests and confidence intervals for $\widehat{\Delta}$ and $\widehat{\Delta}_P$. Let us denote with $n_{\text{test1}|s} = \sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{\widehat{Y}_i = 1\} \cdot \mathbb{I}\{S_i = s\}$ and with $n_{\text{train1}|s} = \sum_{i=1}^{n_{\text{train}}} \mathbb{I}\{Y_i = 1\} \cdot \mathbb{I}\{S_i = s\}$. In addition, we denote $n_{\text{test0}|s}$ and $n_{\text{train0}|s}$ the observed frequency of $\widehat{Y} = 0$ in the test set and $Y = 0$ in the training set. A similar notation is adopted for the subsets with $S = \overline{s}$. In the case of large sample sizes, one can define the confidence intervals for the metrics of interest as follows.

$$\text{CI for } \Delta: \quad \left(\frac{n_{\text{train1}|s}}{n_{\text{train}|s}} - \frac{n_{\text{train1}|\overline{s}}}{n_{\text{train}|\overline{s}}}\right) \pm z_{\alpha/2}\sqrt{\frac{n_{\text{train1}|s} \cdot n_{\text{train0}|s}}{n^3_{\text{train}|s}} + \frac{n_{\text{train1}|\overline{s}} \cdot n_{\text{train0}|\overline{s}}}{n^3_{\text{train}|\overline{s}}}} \tag{3}$$

$$\text{CI for } \Delta_P: \quad \left(\frac{n_{\text{test1}|s}}{n_{\text{test}|s}} - \frac{n_{\text{test1}|\overline{s}}}{n_{\text{train}|\overline{s}}}\right) \pm z_{\alpha/2}\sqrt{\frac{n_{\text{test1}|s} \cdot n_{\text{test0}|s}}{n^3_{\text{test}|s}} + \frac{n_{\text{test1}|\overline{s}} \cdot n_{\text{test0}|\overline{s}}}{n^3_{\text{test}|\overline{s}}}} \tag{4}$$

These are Wald-type confidence intervals. The Agresti and Caffo (1) confidence intervals can be obtained by adding one success and one failure to each one of the groups defined by $S$. Confidence intervals (3) and (4) that include zero suggest that we cannot exclude statistical parity in the sampled population and the prediction, respectively. An interval consisting in only negative values suggest discrimination toward the subgroup with $S = s$, whereas positive values suggest discrimination toward the subgroup with $S = \overline{s}$. A further possible evaluation of uncertainty in fairness concerns the comparison of the two observed metrics, for instance, to evaluate a fairness-aware algorithm. Assuming a null hypothesis $H_0\, \Delta = \Delta_P$, we could be interested to an alternative $H_1 : \Delta > \Delta_P$ to verify if the fairness-aware algorithm is effective. Conversely, we can chose an alternative such as $H_1 : \Delta < \Delta_P$ to check if an algorithm is inducing lack of fairness.

A test statistic that can be used at the aim concerns the difference of differences of proportions (8). The test statistic has the form

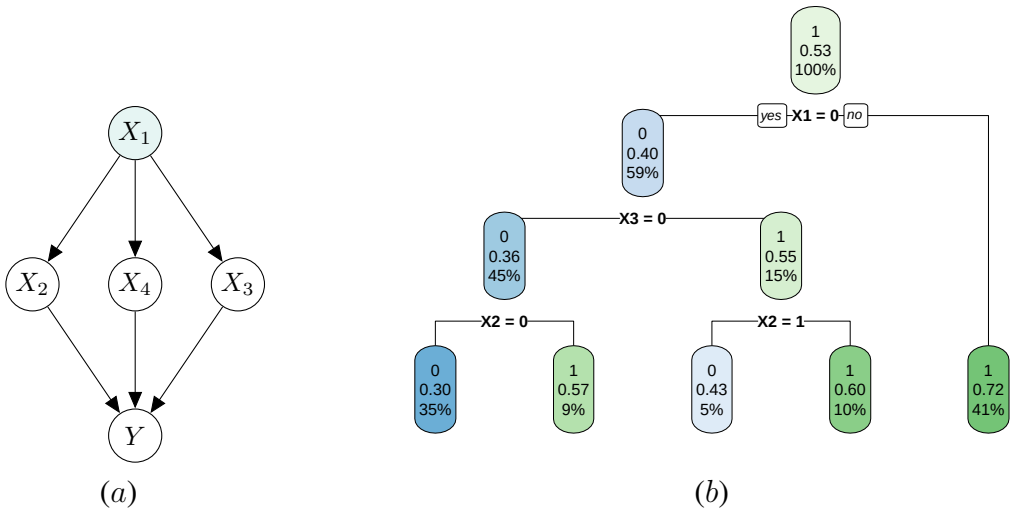
$$T = \left(\frac{n_{\text{train1}|s}}{n_{\text{train}|s}} - \frac{n_{\text{train1}|\overline{s}}}{n_{\text{train}|\overline{s}}}\right) - \left(\frac{n_{\text{test1}|s}}{n_{\text{test}|s}} - \frac{n_{\text{test1}|\overline{s}}}{n_{\text{train}|\overline{s}}}\right).$$

Under the null, it is normally distributed with mean zero and variance equal to

$$\sigma^2_T = \frac{n_{\text{train1}|s} \cdot n_{\text{train0}|s}}{n^3_{\text{train}|s}} + \frac{n_{\text{train1}|\overline{s}} \cdot n_{\text{train0}|\overline{s}}}{n^3_{\text{train}|\overline{s}}} + \frac{n_{\text{test1}|s} \cdot n_{\text{test0}|s}}{n^3_{\text{test}|s}} + \frac{n_{\text{test1}|\overline{s}} \cdot n_{\text{test0}|\overline{s}}}{n^3_{\text{train}|\overline{s}}}$$

Such variance is appropriate as, under the assumption of *iid* data, train data and test data are independent. When the overall sample dimension $n$ is small, the random splitting of the data set can make the results unstable and alternative solutions should be preferred.

DRAFT




$(a)$ $(b)$

Figure 1: Data generating process $(a)$ and tree learn by CART $(b)$ for the synthetic data example

## 3.1 Example on synthetic data

Consider the data generating process shown in Figure 1$(a)$, where $X_1$ is a binary sensitive attribute, with distribution $X_1 \sim \text{Ber}(\pi_1 = 0.5)$. The three predictors $X_j$, $j = 2, 3, 4$ are also binary and have conditional distributions $X_j \mid X_1 = s \sim \text{Ber}(\pi_s)$ with $\pi_1 \approxeq 0.69$ and $\pi_0 \approxeq 0.27$. We assume that the three predictors $X_j$, $j = 2, 3, 4$ with a logistic regression parametrization with intercept $-1$ and the same coefficient, $0.8$. According to Figure 1$(a)$, $Y \perp\!\!\!\perp X_1 \mid X_2, X_3, X_4$ but $Y \not\perp\!\!\!\perp X_1$. By marginalizing over $X_2, X_3, X_4$, the conditional distribution of $Y \mid X_1 = s$ is $\text{Ber}(\pi_{Y|s})$, with $\pi_{Y|1} \approxeq 0.65$ and $\pi_{Y|0} \approxeq 0.42$. The true value of the statistical parity is therefore $0.23$, indicating mild unfairness.

We generated a data set of size 300 from this data generating process and split the data into equal-sized training and test sets. We fitted a classification tree on the training data using the `rpart` package in R with default settings and computed the predicted values for the test data.

Regarding fairness in the population, we estimated $\widehat{\pi}_{Y|1} = 0.721$ and $\widehat{\pi}_{Y|0} = 0.404$, resulting in $\widehat{\Delta}_P = 0.317$, which is slightly overestimating the true value. The $95\%$ confidence interval for $\Delta_P$ is $(0.165; 0.469)$, not including zero, correctly suggesting lack of fairness.

Regarding the fairness of the algorithm, we know from (9) that this data generating process tends to make CART over-utilize the background variables. Figure 1$(b)$ shows that the first variable selected by the greedy algorithm for slitting is exactly the sensitive variable $X_1$. The estimate of $\widehat{\Delta}$ is indeed $0.563$ with a $95\%$ confidence interval $(0.448; 0.679)$, and the test statistic $T$ yields a $p$-value $0.0056$, indicating that the algorithm is significantly increasing bias.

A common approach is to make the algorithm blind to the sensitive variable and exclude $X_1$ as a predictor. In this case, the estimate of $\widehat{\Delta}$ is reduced to $0.393$ and the confidence interval to $(0.257; 0.529)$. This time the test $T$ cannot reject the null hypothesis, as the resulting $p$-value is around $0.2$.

# 4. Concluding remarks

As machine learning becomes increasingly integrated into human life, it is important to recognize that these algorithms have the potential to be both groundbreaking and problematic. They can inadvertently perpetuate or alleviate bias toward certain protected attributes, such as race, religion, and gender. As a result, it is crucial for researchers and practitioners to carefully evaluate the fairness of their data and models. In this paper, we emphasize that the fairness measure computed from the data is actually an estimate of the unknown *true* fairness metric. Therefore, it is essential to assess uncertainty the uncertainty of this estimate. These measures are easy to compute and could therefore be included in fairness-aware algorithms to relax the trade-off between different metrics of fairness simultaneously and

DRAFT

accuracy. Discussion of this trade-off can be found, for instance, in (2) and (5).

Furthermore, we propose a method for differentiating between discrimination resulting from data or Nature and discrimination induced by the algorithm, also taking into account sample variability.

**Acknowledgments** I wish to express my sincere gratitude to Tamàs Rudas for providing invaluable insights during our discussions on the topic. Additionally, I would like to extend a special thank you to Sabrina Giordano for her unwavering support and encouragement.

## References

[1] Agresti, A., Caffo, B.: Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. AM STAT, **54(4)**, 280-288 (2000)

[2] Berk, R. A., Kuchibhotla, A. K., Tchetgen, E. T.: Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. ArXiv preprint arXiv:2111.09211 (2021)

[3] Besse, P., del Barrio, E., Gordaliza, P., Loubes, J. M.: Confidence intervals for testing disparate impact in fair learning. ArXiv preprint arXiv:1807.06362 (2018)

[4] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C.J.: Classification and regression trees. CRC Press (1984)

[5] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797-806 (2017)

[6] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 329-338 (2019)

[7] Dablain, D., Krawczyk, B., Chawla, N.: Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. ArXiv preprint arXiv:2207.06084 (2022)

[8] Goodman, L. A. (1961). Modifications of the Dorn-Stouffer-Tibbitts Method for" Testing the Significance of Comparisons in Sociological Data". AM J SOCIOL, 66(4), 355-363.

[9] Gottard, A., Vannucci, G., Marchetti, G.M.: A note on the interpretation of tree-based regression models. BIOM. J. **62.6**, 1564-1573 (2020)

[10] Ji, D., Smyth, P., Steyvers, M.: Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. ADV NEUR IN, **33**, 18600-18612 (2020)

[11] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: Choices, assumptions, and definitions, ANNU REV STAT APPL, **8**, 141-163 (2021)

[12] Pedreschi, D., Ruggieri, S., Turini, F.: A study of top-k measures for discrimination discovery. In Proceedings of the 27th Annual ACM Symposium on Applied Computing, 126-131 (2012).

[13] Pessach, D., Shmueli, E.: A review on fairness in machine learning, ACM COMPUT SURV, **55.3** 1-44 (2022)

[14] Simpson, E. H.: The interpretation of interaction in contingency tables. JRSSB, **13(2)**, 238-241 (1951)

[15] Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. ArXiv preprint arXiv:1511.00148 (2015)

DRAFT

# Analysis of University Grades: An IRT Model for Responses and Response Times with Censoring

Michela Battauz[a]

[a]Department of Economics and Statistics, University of Udine; michela.battauz@uniud.it

**Abstract**

In this paper, we proposed an Item Response Theory (IRT) model to analyze the grades obtained by the university students and the time needed to pass the exams. The model includes two latent variables which an be interpreted as ability and speed. In fact, in Italian universities, the students have a great flexibility in the choice of the order in which to give the exams and they have several occasions to give them during the academic career. However, in our dataset, the grade is available only for the students who pass the exam. Furthermore, several students drop out of university before completing their studies. It is then fundamental to account for these two sources of censoring when modelling these data. An application to a real dataset illustrates the model.

*In questo articolo, proponiamo un modello Item Response Theory (IRT) per analizzare i voti ottenuti dagli studenti universitari e il tempo necessario per superarli. Il modello include due variabili latenti che possono essere interpretate come abilità e velocità. Infatti, nelle università italiane, gli studenti hanno molta flessibilità nella scelta dell'ordine con cui dare gli esami e ci sono diverse occasioni (appelli) per darli durante la carriera accademica. Tuttavia, nei nostri dati, il voto è disponibile solo per gli studenti che superano l'esame. Inoltre, diversi studenti abbandonano l'università prima di completare i loro studi. Risulta, quindi, fondamentale tener conto di queste due fonti di censura nella modellazione di questi dati. Un'applicazione a dei dati reali illustra il modello.*

***Keywords:*** academic careers, academic performance, censoring, IRT, response time.

DRAFT

## 1. Introduction

In the Italian university system, the students have a great flexibility in the choice of the order in which to give the exams and they have several occasions to give them during the academic career. Furthermore, they can drop out of university before completing their studies, hence introducing missing values in the data composed of the grades. For these reasons, the analysis of the grades should take into account the censoring process and the time plays an important role in it. The availability of the only positive grades in our dataset is another cause of censoring that should be taken into proper account. In this paper, we use a graded response model (Samejima, 1969) for the ordinal grades, though other choices are possible. Following van der Linden (2006), the response times are modelled using a lognormal distribution, jointly with the grades. Our proposal shares many similarities with Guo, Xu, Ying and Zhang (2022), who dealt with computer-based assessment with time limits. The main differences of

our work are the ordinal responses, an additional source of censoring due to the lack of availability of the grades for non-passed exams, and the censoring time being a random variable rather than a known constant. A different approach to the analysis of the university grades is given by Bacci, Bartolucci, Grilli and Rampichini (2017), who do not consider the time needed to pass the exams, but rather indicator variables of the choice of the students to attempt the exams.

In the following sections, after introducing the model and the relative likelihood-based methods, an application will be presented and some concluding remarks will be given.

## 2. Models and methods

Let $X_{ij}$ be an ordinal grade obtained by student $i$ in exam $j$. According to the graded response model, the probability of observing a response equal to $x$ is given by

$$P(X_{ij} = x|\theta_i) = P(X_{ij} \geq x|\theta_i) - P(X_{ij} \geq x+1|\theta_i) \,, \tag{1}$$

with

$$P(X_{ij} \geq x|\theta_i) = \frac{e^{\alpha_j\theta_i+\beta_{jx}}}{1+e^{\alpha_j\theta_i+\beta_{jx}}} \,, \tag{2}$$

where $\alpha_j$ is a slope parameter, $\beta_{jx}$, $x = 1, \ldots, m_j$, are category-specific intercept parameters, $m_j$ is the number of response categories, and $\theta_i$ is a latent variable which represents the ability of the student.

The time at which exam $j$ is passed by student $i$ is denoted by $T_{ij}$. We assume a log-normal distribution for the response times

$$\log T_{ij}|\tau_i \sim N\left(\gamma_j - \tau_i, \frac{1}{\lambda_j^2}\right) \,, \tag{3}$$

where $\gamma_j$ and $\lambda_j^2$ are the parameters that determine the average time required to pass exam $j$ and its variability, while $\tau_i$ is a latent variable which represents the speed of the student.

The latent variables are assumed to follow a bivariate normal distribution

$$\begin{pmatrix}\theta_i \\ \tau_i\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho\sigma_\tau \\ \rho\sigma_\tau & \sigma_\tau^2\end{pmatrix}\right) . \tag{4}$$

Both the grade and the time are observed only if the grade is positive, so that the exam is passed, and the student does not drop out of university before succeeding in the exam. Let $C_i$ be the time at which the student graduates or drops out and $R_{ij}$ an indicator variable equal to 1 if grade and time are observed. Hence, $R_{ij} = 1$ if $T_{ij} \leq C_i$ and $X_{ij} \geq x_{min}$, where $x_{min}$ is the threshold for passing the exam. It can be shown that the joint probability of $R_{ij}, X_{ij}, T_{ij}$, given the latent variables, is then given by:

$$f(r_{ij}, x_{ij}, t_{ij}|\theta_i, \tau_i) = P(x_{ij}|\theta_i)^{r_{ij}} f(t_{ij}|\tau_i)^{r_{ij}} \left\{1 - P(T_{ij} \leq c_i|\tau_i)P(X_{ij} \geq x_{min}|\theta_i)\right\}^{1-r_{ij}} , \tag{5}$$

The censoring time $C_i$ is a random variable in this context. However, it is sensitive to assume that its distribution does not depend on the parameters that determine the distribution of the grades and the times to pass the exams. This assumption, which is usually referred to as noninformative censoring in the literature (Kalbfleisch and Prentice, 2002), allows to drop the distribution of $C_i$ from the likelihood function, which can be expressed as follows:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \rho, \sigma_\tau^2) = \prod_{i=1}^{n} \int \int \prod_{j \in P_i} f(r_{ij}, x_{ij}, t_{ij}|\theta_i, \tau_i) f(\theta_i, \tau_i) d\theta_i d\tau_i, \tag{6}$$

where $n$ is the number of students, $P_i$ is the set of courses in the programme of student $i$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)^\top$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_J^\top)^\top$, with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_j})^\top$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\top$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^\top$. As frequently happens in IRT modelling, the integrals do not have a closed form solution and they are approximated by numerical methods, such as Gaussian quadrature. The maximization of the log-likelihood function provides estimates of the parameters.

DRAFT

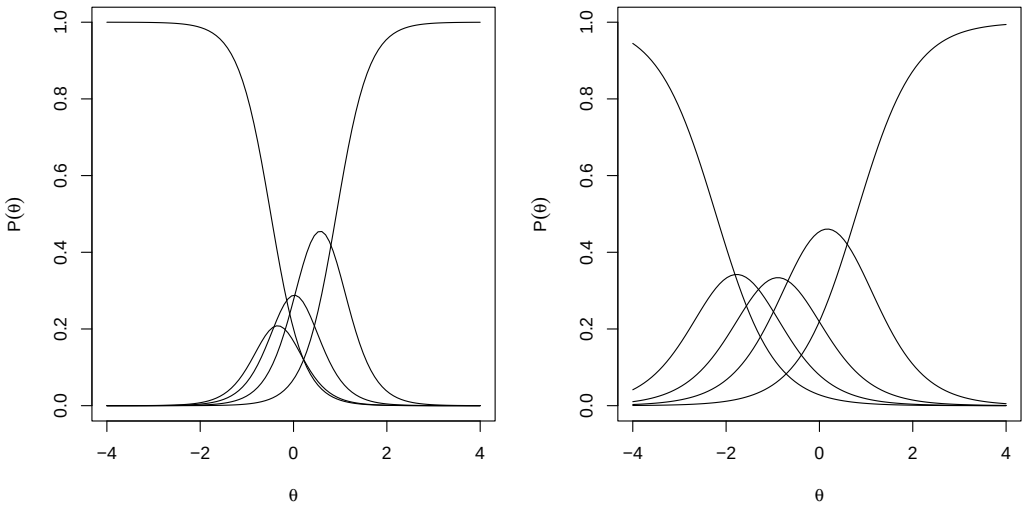



Figure 1: Probability curves for one exam. On the left panel the curves are obtained without correction for censoring, on the right panel the curves are corrected for censoring. The first category refers to non-passed exams, the others refer to the grades grouped in 4 classes.

## 3. Application

The model proposed in this paper was used to analyze the grades obtained by university students in Economics and Business from the University of Udine. These two bachelor degrees share many courses, especially at the first and second year, thus permitting to analyze them together. The sample is composed of a cohort of 324 students who enrolled in 2017. 128 of them dropped out of university or decided to change bachelor course, 180 attained the degree, while 16 are still enrolled in 2022, last year of data available. For this analysis we considered only courses with at least 50 students that have passed the exam, resulting in a total of 22 courses. For each student, we observed only the grades in passed exams which are in the programme of the student and are among the 22 selected ones. Hence, the number of observed exams per student is variable and ranges from 4 to 20, the median is 17.5, while the mean is 15.1. The grades, which range from 18 to 30, have been grouped in 4 classes to reduce the number of parameters of the model. The time is considered equal to the days between the date when the exam is passed and the 1st of October 2017, conventionally assumed as the beginning of the career.

All analyses were performed in R. In order to understand the effect of ignoring the censoring process, we estimated two separate models for the grades and for the times that do not account for censoring. We applied a graded response model to the grades, substituting the missing values with a unique value below the threshold for passing the exam. For this purpose we used the mirt package. Doing so, we incorrectly assume that the students that have an exam in their programme with a missing value for the grade, failed the exam, while it might also be that they dropped out or changed course before even attempting. The analysis of the times that ignores the censoring process was performed by fitting a random effects model to the logarithms of the times with random intercepts for the students, and where the independent variables are dummy variables that identify the exam. The variability of the level-1 errors is allowed to vary across subjects. In this model, the missing values are simply omitted. This model was fitted using the package nlme.

Figure 1 shows the probability curves for one exam taken as example ignoring the censoring process (on the left) and applying our proposal (on the right). The curve that presents the most important change is the one that refers to grades below the threshold for passing. Assuming that the students that don't have the grade registered failed the exam, erroneously results in higher probability of failing the exam for students located at lower levels. This effect was observed for the other exams as well, though the figures are not shown in the paper for space reasons.

Figure 2 shows a comparison of the predicted values of the latent variables ignoring the censoring process and accounting for it by applying the proposal of this paper. In the first case, the predicted values

DRAFT

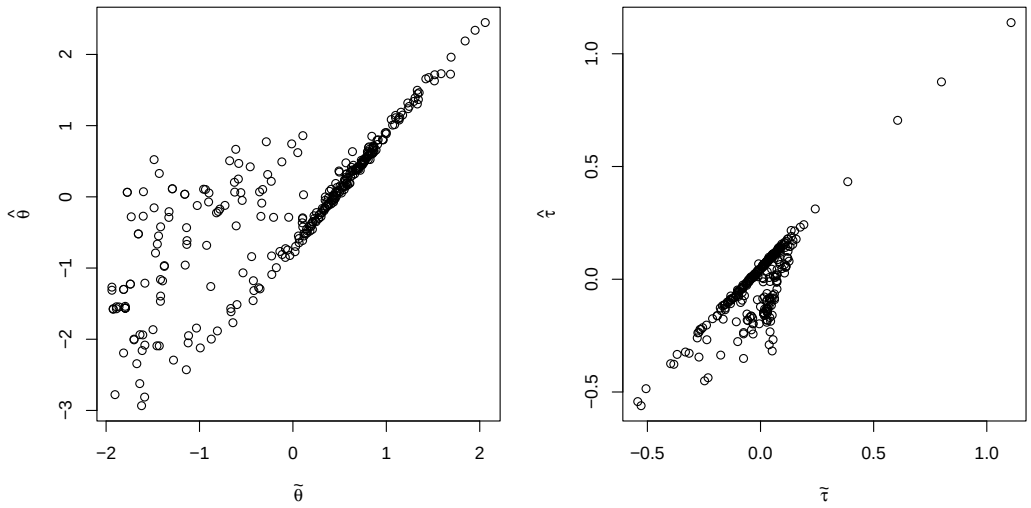



Figure 2: Predicted values of the latent variables obtained without correcting for censoring (horizontal axis) and correcting for censoring (vertical axis).

DRAFT

are denoted by $\tilde{\theta}$ and $\tilde{\tau}$, while in the second case they are denoted by $\hat{\theta}$ and $\hat{\tau}$. It is possible to observe that the predicted values for some students are unchanged since they lay on a line. Such line is not the bisector because of the unidentifiability of the scale of the latent variables, with $\theta$ assumed to have zero mean and variance equal to one. The students with predicted ability levels that do not lay on the line, when accounting for the censoring process, present higher values. This is easily explainable since these are students whose missing values were considered as negative grades. These are anyway students on the lower range of ability, since the high achievers tend to not have missing values as they have passes all the exams in their programme. Observing, instead, the predicted speeds, some students present lower values when accounting for the censoring process. These are students that passed a few exams before dropping out or changed course. Hence, just omitting the non-passed exams leads to assign them higher values than their actual speed.

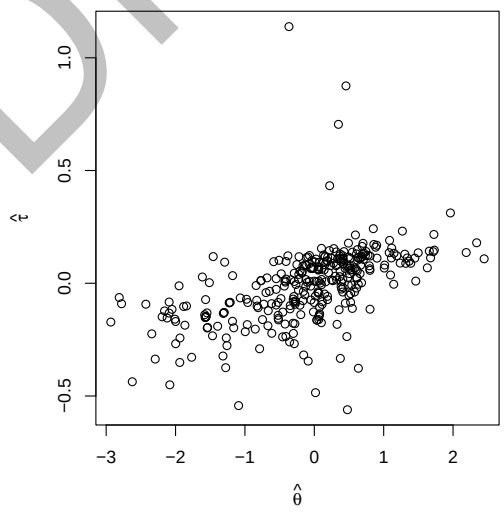



Figure 3: Scatter plot of the predicted values of the latent variables.

Figure 3 shows a scatter plot of the predicted abilities against the predicted speeds. It is apparent a positive correlation between the two. In fact, the correlation index $\rho$ is estimated at about 0.25. A few students present outlier values, with average values of ability and very large values of speed. These students graduated at the second year, hence anticipating the attainment of the degree which normally

requires three years. This suggests a trade-off between attaining high grades and speed at the individual level, though in this population the students who perform well in terms of grades tend to be the fastest too. The standard deviation of $\tau$, which is estimated at about 0.2, is considerably smaller than that of $\theta$, fixed at 1.

## 4. Conclusions

The analysis of the students' performance can provide important information about the careers of the students, including dropping out of university. In this paper we have proposed a model to analyze the students' grades and the time needed to pass the exams taking into account two sources of censoring that are present in our data. Specifically, these are the availability of the only positive grades and dropping out before attempting the exams. Our analyses showed that it is fundamental to take in proper account the censoring process in the estimation of the parameters and the prediction of the latent variables.

In future research, we aim at expanding our model by considering multidimensional ability and speed latent variables. In fact, we suspect the presence of at least two dimensions in these data, which could capture different abilities required by the courses. Another extension of the model that we will pursue is the use of regularization methods for the IRT parameters in order to allow for the estimation of a large number of parameters and thus avoid grouping of the grades. Finally, we will conduct an extensive simulation study to better understand the properties of the estimators and the effect of ignoring the censoring process.

## References


[1] Bacci, S., Bartolucci, F., Grilli, L., Rampichini, C.: Evaluation of student performance through a multidimensional finite mixture IRT model. Multivariate Behav. Res. (2017) doi: 10.1080/00273171.2017.1361803

[2] Kalbfleisch, J. D., Prentice, R. L.: The Statistical Analysis of Failure Time Data. John Wiley & Sons, Inc., Hoboken (2002)

[3] Guo, J., Xu, X., Ying, Z., Zhang, S.: Modeling not-reached items in timed tests: A response time censoring approach. Psychometrika (2022) doi: 10.1007/s11336-021-09810-0

[4] Lawless, J. F.: Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc., Hoboken (2003)

[5] Samejima, F.: Estimation of Ability Using a Response Pattern of Graded Scores. Psychometrika Monograph, 17. Psychometric Society, Richmond, VA (1969)

[6] van der Linden, W. J.: A lognormal model for response times on test items. J. Educ. Behav. Stat. (2006) doi: 10.3102/10769986031002181


DRAFT

# Predicting high schools' students performances with registry's data: a machine learning approach

Lidia Rossi [a], Marta Cannistrà [a], and Tommaso Agasisti [a]

[a]Department of Management, Economics and Industrial Engineering, Politecnico di Milano - Via Lambruschini 4/B, 20156, Milano, Italy
lidia.rossi@polimi.it, marta.cannistra@polimi.it, tommaso.agasisti@polimi.it

**Abstract**

This work aims to predict student performance for a high school in Italy using Machine Learning algorithms, particularly the innovative Ordinal Random Forests. The study intended to find the optimal moment to obtain accurate predictions of students' final marks as early as possible during the academic year to support timely interventions from schools. The models used data from electronic registers, which were updated weekly, and showed that predictive models after four months of schooling could correctly predict the final marks of a high percentage of students. The study's findings have implications for research and practice as school managers can use these models to promote improvements.

***Keywords:*** Early Warning Systems; Learning Analytics; Students' performance prediction; Machine Learning

DRAFT

## 1. Introduction

The use of Early Warning Systems (EWS) is becoming increasingly prevalent in education, as they enable the development of personalized interventions aimed at improving student outcomes. To create these systems, it is necessary to predict student performance as soon as possible (1)(2). This prediction is a crucial area of research in Learning Analytics. Early and accurately predicting academic performance can aid in the planning and implementation of personalized and timely remedial interventions for students, which can improve their academic achievement and prevent failure (3). The prediction of students' performance is challenging due to various factors that influence it, such as psychological traits, socio-demographic characteristics and educational experiences. These factors are often complex and nonlinearly correlated, making it difficult to accurately model their interrelationships (4).

This study addresses the primary research question of how accurately and early machine learning models can predict students' final marks.

## 2. Data and methodology

This study aims at predicting students' performance ($Y$) using both time-invariant ($Z$) and time-varying ($X_t$) predictors. Data from the electronic register merged with administrative data of students in an Italian high school are used to predict students' average of all final marks and

the final marks in three specific subjects: Mathematics, Italian, and English. More specifically, administrative database provide information regarding the personal characteristics of students such as demographic information (gender, citizenship, age) and indicators about minor or major certified disabilities; information about classes - such as the total number of teachers and the type of their contract or the school track; and information of students' academic performance obtained in the previous year such as whether the student was promoted in June or he/she had some specific restorative exams in September or he/she was rejected or transferred, the average of marks in the previous academic year (on a scale between 1 and 10) and the final mark obtained in Mathematics, Italian and English. The second source of information is the electronic registry database updated weekly. These data allow for the construction and updating of summary variables regarding marks in Mathematics, Italian and English, the number of delays, absences, and merit, diligence and disciplinary notes. This typology of data makes the work innovative and relevant in the field of Early Warning System. It allows to predict students' performance in different time points in students' academic careers, deepening the mechanisms behind it. Object of this work is to find the optimal moment for final marks' prediction, balancing between earliness and accuracy. Data of 438 students enrolled in a private high school in Milan during the year 2018/2019 are analysed.

Machine Learning approaches have been adopted in this context to formulate accurate predictions and deal with highly complex data. In particular, the Random Forest algorithm for the classification of ordinal output is performed (5) (6). Iteratively, using updated data every week, the models are constructed and validated computing, with 10-fold cross validation, two performance indices: Youden's J statistic and Mean Square Error.

Due to the limited amount of available observations, a Monte Carlo simulation is developed to test the results over a larger sample.

DRAFT

## 3. Results

The evolution over the year of performance indices are shown in Figure 1 and Figure 2.

At the beginning of the year the Youden's J statistic is equal to 0.27 for the prediction of Mathematics grade, it increase during the year to reach the value of 0.43 in the last week of school; for the same models MSE is equal respectively to 1.18 and 0.54. For what concern final Italian mark, at the beginning of the year Youden's J statistic is equal to 0.33 while MSE to 0.73, at the end of the year the first one increases of 0.29 and the second one decreases of 0.42. The Youden's J statistic in the model that predict English final mark pass from 0.39 to 0.60 while the MSE from 0.71 to 0.29. Finally, for what concern the average of all marks, the indices at the beginning of the year are respectively equal to 0.48 and 0.39 while at the end of the year 0.59 and 0.26.

## 4. Conclusions

The innovative aspects of this paper rely on the model adopted (Ordinal Random Forest) and on the data used. Indeed, the model combines weekly data from the school register, including intermediate evaluations, disciplinary notes, and delays, with administrative data to predict final outcomes. Findings show that, as expected, going forward with the weeks, the predictions' performance improve. Interestingly, mathematics' marks are the hardest to predict, probably because of the fluctuations over the academic year.

The models, and related results, can be used as input to decide the optimal moment to intervene, balancing between early and accurate predictions. Also, the EWS can be used to create

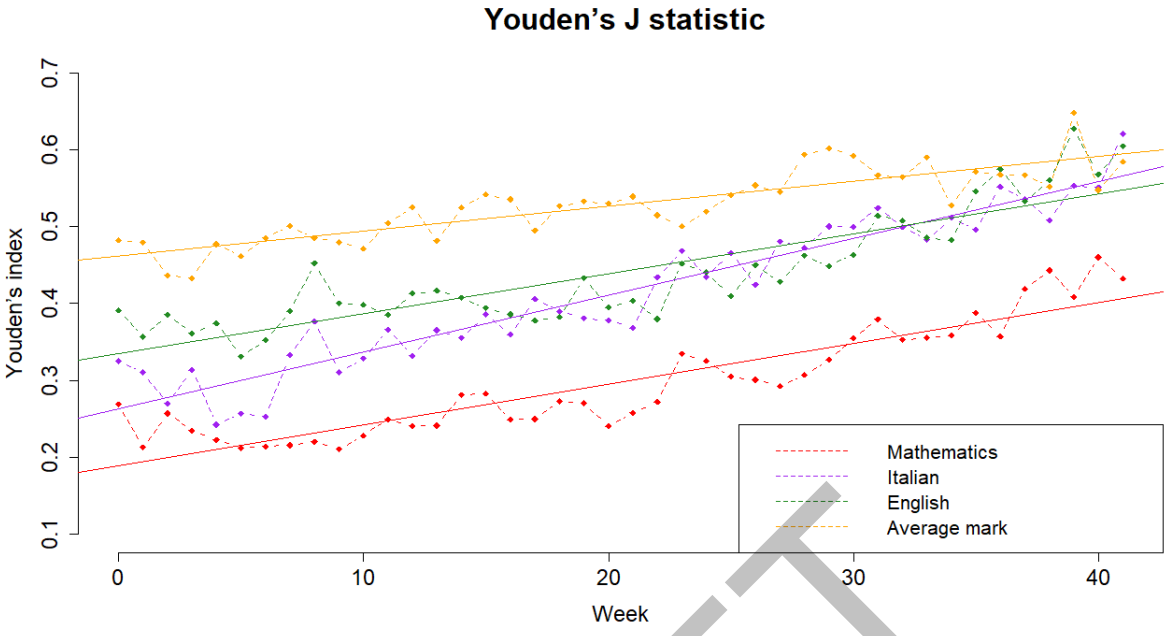



Figure 1: Evolution over time of Youden's J statistic applying Random Forest on the original dataset

DRAFT

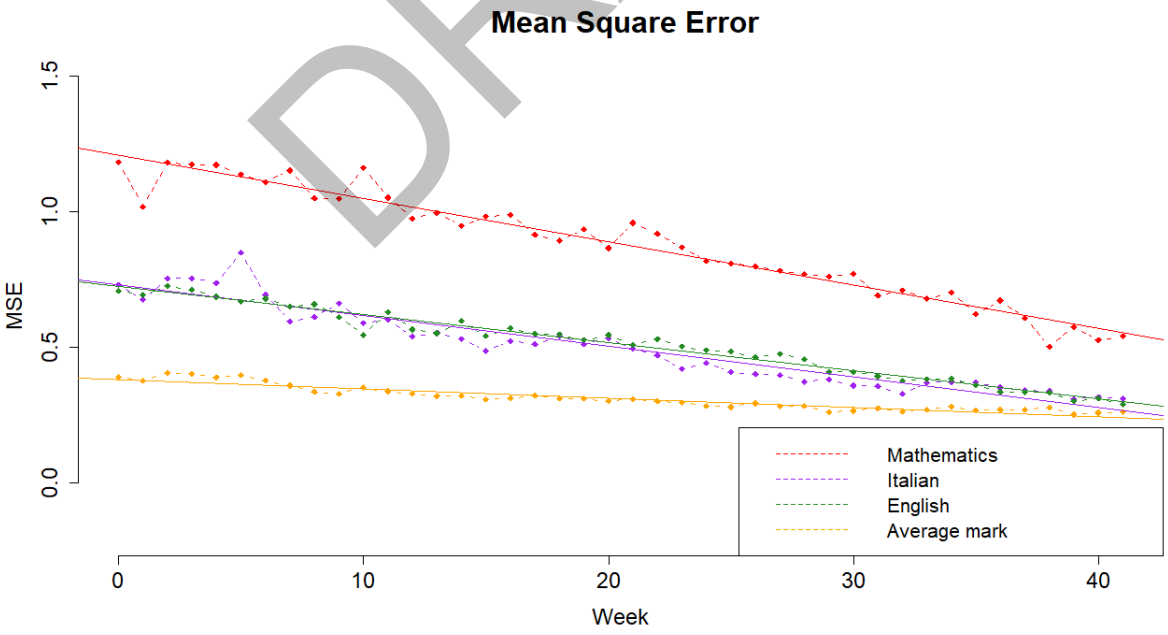



Figure 2: Evolution over time of Mean Square Error applying Random Forest on the original dataset

informative dashboards for teachers and principals, but caution must be taken to ensure that it complements, rather than replaces, teachers' professional judgment, to follow data security and privacy protocols, and to ensure accurate and regular data collection.

Further research is needed to validate the EWS's ability to predict results in standardized tests.

## References


[1] Carl, B., Richardson, J. T., Cheng, E., Kim, H., & Meyer, R. H. (2013). Theory and application of early warning systems for high school and beyond. Journal of Education for Students Placed at Risk (JESPAR), 18(1), 29-49.

[2] Davis, M., Herzog, L., & Legters, N. (2013). Organizing schools to address early warning indicators (EWIs): Common practices and challenges. Journal of Education for Students Placed at Risk (JESPAR), 18(1), 84-100.

[3] Sousa, E. B. D., Alexandre, B., Ferreira Mello, R., Pontual Falcão, T., Vesin, B., & Gašević, D. (2021). Applications of learning analytics in high schools: a systematic literature review. Frontiers in Artificial Intelligence, 4, 737891.

[4] Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 39(10), 9468-9476.

[5] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[6] Hornung, R. (2020). Ordinal forests. Journal of Classification, 37, 4-17.


DRAFT

# Using response times to identify cheaters in CAT: A simulation study

Luca Bungaro[a], Bernard P. Veldkamp[b], Mariagiulia Matteucci[a]

[a] Department of Statistical Sciences, University of Bologna, Italy; luca.bungaro2@unibo.it, m.matteucci@unibo.it

[b] Faculty of Behavioral, Managerial and Social Sciences, University of Twente, The Netherlands; b.p.veldkamp@utwente.nl

**Abstract**

In this paper, we introduce a method that can identify those who are cheating during a test, to modify the test itself, the ability and the speed estimation methods, while the test is still ongoing. The model is based on the idea of using response times to estimate e a new person fit statistic that can be updated as the test is carried out, to mark as outliers those who overcome a certain threshold. At that point, the items' selection algorithm will choose only those items that have been previously indicated as more secure, decreasing the probability that the cheater already knows the correct answer. In this way, as shown by a simulation study, we have considerably improved the estimates of cheaters' ability, while not affecting those of non-cheaters.

***Keywords***: computerized adaptive testing, response times, aberrant behaviours, simulation.

DRAFT

## 1. Introduction

Due to advances in information technology, many standardized tests rely on computer-based testing (CBT) because of its operational advantages. In fact, CBT offers the opportunity to administer large-scale test at frequent time intervals (also for online classes), which is referred to as continuous testing. In addition, CBT enables testing organizations to record scores more easily and to provide feedback and test results immediately after the test administration.

Another advantage of CBT is that it offers the possibility of collecting response time (RT) information on items. RTs provide information not only about candidates' ability ($\theta_i$) and response behaviour, but also about item and test characteristics. This information can be used for testing operation, such as item calibration and test design, but also for more. For instance, in computerized adaptive testing (CAT), a type of CBT, the difficulty level of items is adapted to the response pattern of the candidate. In fact, items are sequentially selected in real time from a large item pool with hundreds of items, according to the candidate's current performance. In this context, RTs can also be used to improve the estimation of interim ability ($\theta_{i_m}$) and the adaptive item selection. Lastly, RTs could be used for the detection of cheating.

In fact, although continuous testing provides test takers with considerable flexibility and convenience, it also raises serious security concerns. Frequently administered items are at great risk for becoming compromised, thereby undermining the integrity of the test. Individuals who take the test earlier (the sources) could share the items orally or online, which would benefit subsequent test takers (beneficiaries), compromising the validity and fairness of the test.

To counter such a security issue in CAT, much psychometric research has been focused on preventive measures involving some kind of item exposure control while still maintaining the efficiency or accuracy of ability estimation as much as possible. The Sympson-Hetter (SH) method (Sympson and Hetter,

1985) is one of the widely used applications of this strategy. However, even the most successful exposure controls cannot entirely prevent cheating since most items will necessarily be administered multiple times. Therefore, there is a great need for diagnostic measures to also spot anomalous behaviours of test takers. The general strategy is to detect an aberrant pattern of responses or RTs across all items that have been administered to the test-taker. There is extensive literature on the use of person misfit statistics, for example the person fit statistic $l_i^t$ (Marianti et al., 2014, Fox and Marianti, 2017).

However, most of the literature focuses on methods that are viable only after the test has ended, making the resulting intervention less effective (for example, replacing items which are identified as compromised) and sometimes impractical or time consuming (for example, to check for cheating, alternative forms of testing, such as oral testing, may be used).

In this article, we would like to propose and evaluate, through a simulation study, a method based on RT to identify cheating behaviours and intervene while the test is taking place.

## 2. Method

In this section, we present a method for the identification of test takers who, during a CAT, are presumed to have come into possession of the correct answers to some questions even before having started the test (pre-knowledge). They may have got this information in many ways: they found the questions and the related solutions online, or by word of mouth of those who had already done the test, or even managed to hack the database with the questions. The aim of the method is to detect aberrant behaviours of the test takers, so it tends to focus more on test takers' response patterns, rather than on identifying compromised items.

Like other methods already proposed in the literature (Marianti et al. 2014; Fox and Marianti, 2017), the identification of cheaters takes place through the analysis of RTs: those who tend to have RTs that differ greatly from the expected time for each item, may be possible cheaters.

The peculiarity of this approach, compared to the existing literature, is that such identification takes place while the test is carried out and not after. The idea is to define a person misfit statistic that can be calculated in real time during the test, and that is updated as the test taker answers to the items that are administered. In this way, the misfit statistic can be integrated within the item selection algorithm to try to compensate for the supposed cheating.

First, we define a misfit statistic starting from the current literature. We start from the person misfit statistic based on RT defined by Marianti et al. (2014) and then modified by Fox and Marianti (2017):

$$l_i^t = \sum_{k=1}^{K} \left( \frac{lnRT_{ik} - \mu_{ik}}{\sigma_k} \right)^2, \quad (1)$$

where $RT_{ik}$ is the response time of subject $i$ to item $k$, $\mu_{ik}$ indicates the expected time that a test taker $i$, with an estimated speed $\zeta_i$, takes to respond to item $k$ and $\sigma_k$ is the standard deviation of the logarithm of the RT for item $k$, for $i = 1, \dots, n$ and $k = 1, \dots, K$ (where $K$ is the total test length). In fact, in the literature, a widely used choice is to assume a log normal distribution for the RT (van der Linden, 2006):

$$RT_{ik} = \lambda_k - \varphi_k \zeta_i + \varepsilon_{ik} = \mu_{ik} + \varepsilon_{ik}, \varepsilon_{ik} \sim N(0, \sigma_{\varepsilon_k}^2), \quad (2)$$

where $\zeta_i$ is the speed parameter, representing the constant working speed of that test-taker $i$, $\lambda_k$ is the time-intensity parameter of item $k$, representing the population-average time (on a logarithmic scale) needed to complete item $k$, $\varphi_k$ is the time-discrimination parameter of item $k$, representing the sensitivity of the item for different speed levels of the test takers. Lastly, $\varepsilon_{ik}$ is an additional error term that can model variations in RTs that cannot be explained only by the structural mean term, such as when test-takers operate with different speed values, take small pauses during the test, or change their time management.

From Equation (1), the statistic $l_i^t$ is defined as a sum of differences that includes all the test items, and these differences are calculated with respect to the estimate of the response speed $\zeta_i$, that is derived at the end of the test using the Gibbs sampling algorithm (Fox et al., 2021), which is very difficult to implement during the test, because it is very time-consuming, thus undermine one of the main advantages of CAT, where the next item has to be selected almost instantly. For this reason, $l_i^t$ can only be defined after the test has been completed.

DRAFT

Our proposal of real-time computable statistics provides for the replacement of $\mu_{ik}$ and $\sigma_k$ with parameters that can be calculated in each step *m* of the test (where each step *m* consists of both the phase of interim ability estimation and that of item selection). Under the hypothesis of log normal distribution of RTs, such parameters are the *expected response time* (Fan et al., 2012, Veldkamp, 2016) and the *reciprocal time-discrimination* (van der Linden, 2006):

$$E\left[RT_{ik}\middle|\hat{\zeta}_{MLE_m i}\right] = e^{(\lambda_k - \hat{\zeta}_{MLE_m i} + \frac{1}{2\,\varphi_k{}^2})}\ ;\ \hat{\zeta}_{MLE_m i} = \frac{\sum_{k\in R_m}[\,\varphi_k{}^2(\lambda_k - lnRT_{ik})]}{\sum_{k\in R_m}[\,\varphi_k{}^2]}$$

$$\sigma_k = \frac{1}{\varphi_k}. \tag{3}$$

The new statistic, called *interim* person fit statistic, is defined as follows:

$$l_{i_m}^t = \sum_{k=1}^{m}\left(\frac{lnRT_{ik} - ln\ \left[RT_{ik}\middle|\hat{\zeta}_{MLE_m i}\right]}{\frac{1}{\varphi_k}}\right)^2. \tag{4}$$

The statistic in Equation (4), as well as in Equation (1), follows a $\chi^2$ distribution with *m* degrees of freedom, as it represents the standardized error of normally distributed logarithms of RT. This means that if a significance level of $\alpha$ is chosen, the threshold value *C* can be easily found from the $\chi^2$ (*m*) distribution. If the threshold *C* is exceeded, the test taker is identified as a cheater.

As proposed by Veldkamp (2016), after a test taker has been flagged as a cheater, the item selection algorithm, through an approach called the shadow test approach (STA; van der Linden, 2010), will start to administer the next items from a *more secure* bank of items (a bank of items that has a very low exposure rate), in order to reduce the probability that the cheater has pre-knowledge of the answers to those items.

## 3. Simulation study

Our proposal was tested by a simulation study. This section presents the characteristics of the simulations and the main results.

### 3.1. Simulation design

In the simulation, we compare the method we propose, i.e., a constrained method to identify cheaters during the test, with a method without constraints. We compare the goodness of the ability estimates for both the methods by considering bias and RMSE. In addition, as for the constrained method, we keep track of the times a student has been correctly identified as a cheater or as not a cheater to check for the classification accuracy of the approach.

Both methods were tested in different simulated scenarios, in which some characteristics are kept fixed and other characteristics are manipulated.

- $n = 100$.
- Fixed length CAT with $K = 35$.
- For each student we simulated $\theta_i$ and $\zeta_i$ from a bivariate normal distribution (van der Linden et al. 2007, van der Linden et al. 2016, Fox et al. 2021) with mean equal to zero and an assumed negative correlation (-0.5), so, on average, more proficient students tend to respond more slowly, not to risk mistakes, while less proficient students tend to underestimate the test and respond quickly. Of course, this does not exclude the possibility that more proficient students respond faster because they understand the item.
- 20% of the students are *cheaters*. $\theta_i$ and $\zeta_i$ of *cheaters* follow the same distribution as the other students. A *cheater* is assumed to always respond correctly to items on which she/he has pre-knowledge, regardless of her/his ability and the item difficulty. A *cheater* responds faster than average to items on which she/he has pre- knowledge.

DRAFT

- RTs that *cheaters* take to respond to items on which they have pre-knowledge depends on the scenario and can be equal to ¼ or ½ of the time that it would take in case they were not cheaters.
- The number of items in the main database on which the *cheaters* have pre-knowledge depends on the scenario and can be 50%, 75% or 100% of the total. It is also assumed that *cheaters* have no item pre-knowledge for the *more secure* database.
- The main database consists of 170 items taken from the *Credential Form* database (available in the *R* package `LNIRT`; Fox et al., 2021), whose psychometric characteristics ($\alpha$, $\beta$, $\varphi$ and $\lambda$) have been estimated using the *R* package `LNIRT`. The *more secure* database is mirrored to the main one, therefore it contains items with the same psychometric characteristics.

The simulations were performed on *R studio* using the packages: `LNIRT` (Fox et al., 2021), `ShadowCAT` (Karel-Kroeze, 2017) and `catIrt` (Steven Nydick, 2022). Each scenario (2 speed multiplier and 3 item pre-knowledge proportion, for a total of 6 scenarios) has been simulated 100 times (keeping fixed person and item parameters) and the results reported in the next paragraph are the average of the simulation results.

### 3.2. Results

Table 1 shows the bias and RMSE of the ability estimates for the two methods (unconstrained and constrained) for different levels of pre-knowledge (p-k) of cheaters, by keeping the speed multiplier equal to 4.

Table 1: Ability estimates for different levels of pre-knowledge (speed multiplier equal to 4).

| Methods | RMSE | BIAS | BIAS CHEATERS | BIAS NON-CHEATERS | ACCURACY CHEATERS[a] |
|---|---|---|---|---|---|
| No constraints (p-k 50%) | 0.288 | 0.196 | 0.861 | 0.030 | 0 |
| No constraints (p-k 75%) | 0.968 | 0.398 | 1.869 | 0.030 | 0 |
| No constraints (p-k 100%) | 7.238 | 1.212 | 5.940 | 0.030 | 0 |
| Constraints (p-k 50%) | 0.095 | 0.050 | 0.128 | 0.030 | 0.940 |
| Constraints (p-k 75%) | 0.094 | 0.055 | 0.153 | 0.030 | 0.974 |
| Constraints (p-k 100%) | 1.847 | 0.338 | 1.571 | 0.030 | 0.790 |

[a] Indicates the ratio of times the model correctly identified a cheater to the total number of cheaters.

As can be seen from Table 1, the use of the constrained method results in an improvement of the estimates compared to the non-constrained model (lower bias and RMSE), for all the three pre-knowledge levels. Also, from the separate bias analysis for cheaters and non-cheaters (columns 4 and 5 of Table 1), it is possible to notice that the bias of non-cheaters is the same for both methods.

However, as the level of pre-knowledge of cheaters increases, the bias grows, reaching high levels for 100% of pre-knowledge. This is also explained by the analysis of classification accuracy, in fact for cheaters, the constrained method goes from an accuracy of 0.94 for a 50% pre-knowledge, to an accuracy of 0.79 when the degree of pre-knowledge is 100.

Finally, the sensitivity level of the proposed method was also tested, decreasing the speed multiplier from 4 to 2. Lower speed multiplier results in a slight decrease in classification accuracy and a corresponding increase in bias and RMSE for low pre-knowledge levels, while it increases considerably for high pre-knowledge levels (clearly, the method without constraints is not affected by the different speed multiplier of cheaters).

Table 2: Ability estimates for different levels of pre-knowledge (speed multiplier equal to 2).

| Methods | RMSE | BIAS | BIAS CHEATERS | ACCURACY CHEATERS[a] |
|---|---|---|---|---|
| Constraints (p-k 50%) | 0.099 | 0.06 | 0.177 | 0.869 |
| Constraints (p-k 75%) | 0.159 | 0.087 | 0.312 | 0.922 |
| Constraints (p-k 100%) | 7.238 | 1.212 | 5.940 | 0 |

[a] Indicates the ratio of times the model correctly identified a cheater to the total number of cheaters.

DRAFT

## 4. Concluding remarks

In conclusion, the method proposed for the identification of cheaters during the test, with the introduction of a new *interim* person fit statistic $l_{i_m}^t$, fully meets expectations: it manages to considerably improve the estimates of cheaters' ability, while not going to affect those of non-cheaters. In fact, as confirmed also by an analysis carried out on the degree of classification accuracy, the constrained method has been able to identify non-cheaters as such 100% of the time.

Although, this improvement tends to decrease, even considerably, both when the percentage of cheaters' pre-knowledge increases and when their speed multiplier decreases. This decrease in classification accuracy is due to an important limitation of this method, which this preliminary analysis has shown, namely the fact that statistic (4) reacts well to a change in speed, however, if a student keeps her/his speed more or less constant throughout the test (and a cheater who knows all the answers will have a very high speed and always constant), then the statistic has more difficulty to identify a cheater. Surely this is an important aspect to take into account for the improvement of the method.

A possible idea to improve this limitation could be to administer, during the initial stages of the test, items from the more secure database to those who have a high interim speed estimate (3). In this way, cheaters who will face a question of which they do not know in advance the answer will tend to slow down and the statistic will be able to better identify them. This would also help to better distinguish those who have item pre-knowledge from those who make *fast guessing* (not knowing the answer, they choose it very quickly at random).

Future developments will test this as well as analysing the issue of item exposure and fast guessing in more complex scenarios than those shown in this preliminary study.

DRAFT

## References


[1] Fan, Z., Wang, C., Chang, H.-H., Douglas, J.: Utilizing response time distributions for item selection in CAT. *J. Educ. Behav. Stat.* **37**(5), 655--670 (2012)

[1] Fox, J.-P., Marianti, S.: Person-fit statistics for joint models for accuracy and speed. J. Educ. Meas. **54**(2), 243--262 (2017)

[2] Fox, J.-P., Klotzke, K., Simsek, A.S.: LNIRT: An R package for joint modeling of response accuracy and times. *arXiv preprint arXiv:2106.10144* (2021)

[3] Kroeze, K.: R Package `ShadowCAT` https://github.com/Karel-Kroeze/ShadowCAT (2017)

[4] Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., Tijmstra, J.: Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* **39**(6), 426--451 (2014)

[5] Nydick, S.: R Package `catIrt`, https://github.com/swnydick/catIrt (2022)

[6] Sympson, J. B., Hetter, R.D.: Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (1985)

[7] van der Linden, W.J.: A lognormal model for response times on test items. *J. Educ. Behav. Stat.* **31**(2), 181--204 (2006)

[8] van der Linden, W.J.: A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* **72**(3), 287--308 (2007)

[9] van der Linden, W.J., Glas, C.A.W. (eds.): Elements of Adaptive Testing. Vol. 10. Springer, New York (2010)

[10] Veldkamp, B.P.: On the issue of item selection in computerized adaptive testing with response times. J. Educ. Meas. **53**(2), 212--228 (2016)

# A geostatistical investigation of the ammonia-livestock relationship in the Po Valley, Italy

Paolo Maranzano[a,b], Kelly McConville[c], Philipp Otto[d], and Felicetta Carillo[e]

[a]Dept. of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126, Milano, Italy; paolo.maranzano@unimib.it
[b]Fondazione Eni Enrico Mattei (FEEM), Corso Magenta, 63, Milano, 20123, Milano, Italy
[c]Dept. of Statistics, Faculty of Arts and Science (FAS), Harvard University, Oxford Street, 1, 02138, Cambridge, MA, USA; kmcconville@g.harvard.edu
[d]Leibniz University Hannover, Institute of Cartography and Geoinformatics, Appelstrasse 9a, Hannover, 30167, Lower Saxony, Germany; philipp.Otto@ikg.uni-hannover.de
[e]Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria - Centro di ricerca Politiche e Bioeconomia (CREA-PB), Milano, Italy; felicetta.carillo@crea.gov.it

DRAFT

**Abstract**

We explore the relationship between ammonia ($NH_3$) emissions and manure processed in the Lombardy region (Italy) at the sub-regional level (i.e., the agrarian subregions). We propose a two-step spatio-temporal statistical analysis. In the first step, we use several spatio-temporal specifications of area-level Small Area (SAE) Models to obtain credible estimates of the quantity of animal manure processed for each agrarian subregion from 2016 to 2020 using data provided by the Italian Farm Accountancy Data Network annual survey. In the second step, we perform an exploratory data analysis studying the empirical relationship between $NH_3$ emissions and manure processed. Preliminary results reveal a strong positive but non-linear relationship between processed manure and ammonia emissions generated by livestock farms.

***Keywords:*** Small Area Models; Spatio-temporal statistics; Manure and livestock; Lombardy; Ammonia emissions

## 1. Introduction

Ammonia ($NH_3$) is a crucial contributor to air pollution levels, as it can become particulate matter after combining with other pollutant materials from various sources. Reducing ammonia emission limits both the potential for air pollution and the wider environmental impact associated with nitrogen pollution. Around 75% of European ammonia emissions come from livestock production. Emissions occur at all stages of manure management: from buildings housing livestock; during manure storage; following manure application to land; and from urine deposited by livestock on pastures during grazing (3).

Specifically, the Lombardy Region of Italy is characterized by high intensity of production and concentration of crops and livestock, mainly in the plains area in the South (1). In this framework, the difficulty of the relationship between agricultural activities and the environment is evident. This situation is also exacerbated by the fact that, in the last decades, the imbalance between animal breeding activities and land used for animal feed production increased. In principle, it is possible to avoid large

part of ammonia emissions from agricultural activities. This is possible if sufficient investments are made in suitable technologies or infrastructures. However, such investments can be very expensive. The burden can be particularly onerous for the many farms that do not have access to external financial support and for which long-term investments are very challenging. However, some interventions require only minor management changes, and although they may make only modest gains, any reduction in emission rates contributes to an overall net benefit. Therefore, even for farms where investment is not possible, minor improvements in existing management can still help achieve emission reduction targets.

Motivated by important production and policy implications, this article investigates the relationship between ammonia emissions and manure processing in the Lombardy region, Italy, using different data sources, including the Italian Farm accountancy data network (FADN). From a methodological point of view, we propose a two-step spatio-temporal statistical analysis. Each step is associated with a specific goal. In the first step, we aim at obtaining a reliable estimate of the manure processed at the sub-regional level by implementing spatio-temporal Small Area Estimation (SAE) techniques (2). In the second step we explore the empirical relationship between manure and $NH_3$ emissions for sub-regions using a data-driven statistical analysis.

The research is part of the Agrimonia (1) project (https://agrimonia.net/), an international research group that aims at quantifying the influence of agriculture on local air quality for Lombardy region using statistical learning and data science methodologies.

## 2. Data and pre-processing

Data used in this paper are collected from several institutional sources and comprise survey sampling data, census (administrative) information, and satellite measurements.

The farm-level annual data from 2016 to 2020 are provided by Italian Farm Accountancy Data Network (FADN). The FADN is an annual sample survey established by the European Economic Commission in 1965, with EEC Regulation 79/56 and updated with EC Reg. 1217/2009 and subsequent amendments. It has been carried out in Italy since 1968 with a similar approach in all Member States of the European Union and represents the only harmonized source of micro-level data on the evolution of incomes and production systems, and on the economic-structural dynamics of farms. Italian FADN provides yearly information on the economic (balance sheet and the income statement), productive, and structural characteristics of Italian farms. At the national level, the annual sample counts about 11000 farms, while the Lombardy sample includes around 600 farms corresponding to a grand total of about 34 thousand farm companies.

According to the Italian FADN, Lombardy is partitioned into the so-called *agrarian sub-regions*, that is, internally homogeneous areas with similar agricultural characteristics (e.g., Oltre Po Pavese for wine production or Lomellina for rice production). Officially, Lombardy is partitioned into 77 areas. However, some of them have very few or even none farms in the regional sample.Therefore, we performed a spatial aggregation of the poorly represented territories, eventually obtaining 66 areas. In Figure 1, we represent the spatial distribution of the farms across the sub-regions. The figure shows that farms are heterogeneously distributed throughout Lombardy. Indeed, farms are mainly concentrated in the southeastern plain, where the region's large intensive livestock farms are located, while the alpine areas in the North are poorly covered. This fact is consistent with the morphological conformation of the region, with the sparsely inhabited mountain range of the Alps and little agricultural space at North, and the intensively agriculture-oriented Po Valley at South.

For each farm, among the other information, FADN provides annual measurements of the total amount (quintals) of manure produced by swine and bovines and the total amount of manure processed (i.e., spread across the agricultural land). Farm-level data are then aggregated at the area-level by employing the sampling weights representative of the overall regional population[1]. Specifically, farm-level quantities of manure produced and processed are aggregated to area-level measurements through the

[1]The regional samples are built using a stratification approach based on two criteria: economic size (production volume) and techno-economic orientation (https://rica.crea.gov.it/APP/documentazione/?tag=ote). Also, only farms with at least 8000â¬$of standard output production are considered.$

DRAFT

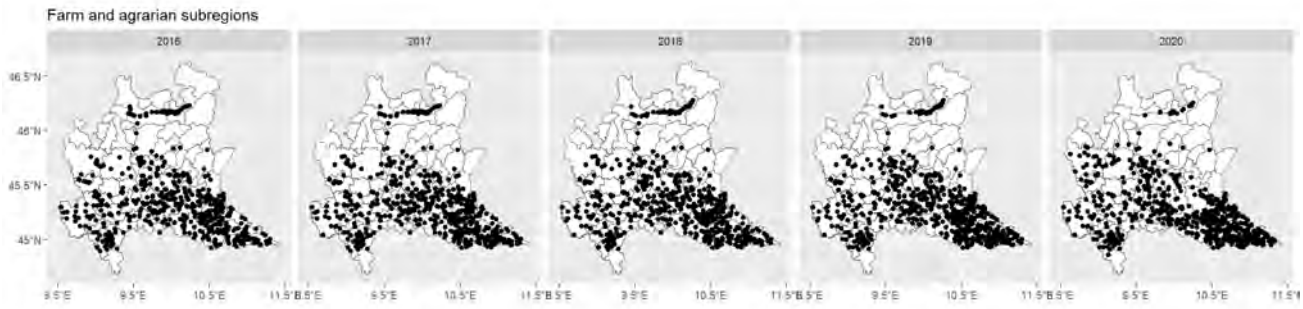



Figure 1: Agrarian sub-regions of Lombardy and spatial distribution of sampled farms from 2016 to 2020.

Hájek direct estimator of the mean (see Section 2.8.2 of (2)), denoted as $\hat{\bar{y}}$.

Further, we consider census information on livestock from the Banca Dati Nazionale Veterinaria (BDN). Data are collected at the municipal level for the whole period and then aggregated at the sub-regional level as the total number of heads per area. Data on $NH_3$ emissions ($kg/hm^2$) from the livestock sector management are collected from the CAMS-anthropogenic emissions dataset. Starting from the total annual emissions on $0.1° \times 0.1°$ grid, the total emissions of each agricultural subregion are given by the sum of the cells in the area. The amount of manure spread in the soil depends on the agricultural land available and the quality of the terrain. Therefore, we consider soil use information gathered from the Lombardy Region Agriculture Information System (SIARL). In particular, we extracted the total agricultural land surface ($hm^2$) and the total surface of soil classified as high-quality agricultural land ($hm^2$). Finally, we consider the average elevation of the sub-regions provided by ISTAT. Full data descriptions on BDN livestock, CAMS emissions data and SIARL data are available in (1).

The final dataset is structured in a longitudinal format with yearly data from 2016 to 2020 ($T = 5$) and $N = 66$ cross-sectional units, i.e., the agrarian sub-regions.

DRAFT

## 3. Empirical strategy

The proposed two-step analysis approach combines several statistical methodologies for modeling spatiotemporal data. In the first step, the goal is to obtain reliable estimates of the quantity of animal manure processed for each agrarian subregion from 2016 to 2020. To do so, we employ several spatio-temporal specifications of area-level SAE models (2). The rationale behind SAE models is to refine a direct estimate of the population mean or the population total through linear mixed models capable of leveraging exogenous information (covariates) of phenomena correlated with the response variable. Specifically, we take advantage of auxiliary information on animals and soil to improve direct estimates of thousand tonnes of manure processed for each agrarian region, while also taking into account the spatio-temporal structure of the data.

Let $d$ be a generic agrarian sub-region and $t$ a generic year from 2016 to 2020. Also, let $\hat{\bar{y}}_{dt}$ be the direct estimate of the average processed manure (Hájek direct estimator) at time $t$ at the agrarian subregion $d$ obtained aggregating the survey data from the FADN. The SAE models can be expressed as a linear mixed model including a linear fixed effects (regression) component and a random effect component as follows:

$$\hat{\bar{y}}_{dt} = X_{dt}\beta + u_{dt} + e_{dt} \tag{1}$$

where $X_{dt}$ is the design matrix of area-level covariates for year $t$, which includes the total number of cattle and pigs, the total agricultural land, the amount of land with highly-rated agricultural value, and the average elevation of the agrarian subregion $d$ at time $t$. The term $u_{dt}$ is a model-specific random effect that accounts for purely random, spatial, or spatio-temporal effects.

- Simple Fay-Herriot model with area-level covariates (hereafter FH), i.e., yearly-specific LMM where sub-regions are linked from independent random effects;
- Spatial FH model with Queen contiguity structure for the neighbors and area-level covariates (hereafter SFH), i.e., a yearly-specific LMM where sub-regions are linked from spatially-correlated random effects;
- Spatio-temporal FH model with independent domain-time random effects and area-level covariates (hereafter STFH), i.e., LMM where sub-regions are linked from spatially-correlated random effects in addition to an independent time random effect;
- Spatio-temporal FH model with autoregressive AR(1) domain-time random effects and area-level covariates (hereafter STFH-AR1), i.e., LMM where sub-regions are linked from spatially-correlated random effects plus an AR(1)-like temporal random effect.

In model FH, $u_{dt}$ is a collection of independent Gaussian distributed random variables; in model SFH they are a collection of spatially correlated following a simultaneously autoregressive (SAR) process with Queen proximity matrix; in STFH, in addition to the spatial effect, we include a temporal random effect using a sequence of mutually uncorrelated Gaussian random variables and uncorrelated in space; in model STFH-AR1, the previously stated temporal random effects follow an AR(1) temporal structure. Eventually, $e_{dt}$ is the sampling error term.

The second step consists of an exploratory data analysis (EDA) of the empirical relationship between $NH_3$ emissions and manure processed. We propose to employ several spatio-temporal linear models, that is Linear Models (LMs), LMMs, and Generalized Additive Models (GAMs) (4). In particular, LMMs and GAMs are implemented to accommodate non-linear relationships and to exploit the spatiotemporal dynamics of the observations. Let $\hat{M}_{dt}$ be EBLUP estimate (2) of the total annual amount of manure processed for region $d$ at time $t$ computed from the SAE models in the first step. Further, let $E_{dt}$ be the $NH_3$ emissions for region $d$ and year $t$. Then, the ammonia-manure relationship can be specified as follows:

$$E_{dt} = f(\hat{M}_{dt}) + \varepsilon_{dt} \tag{2}$$

where $f(\hat{M}_{dt})$ is a smooth function of the estimated region-and-year-specific manure processed, and $\varepsilon_{dt}$ is an error term. Depending on the specification of the model, the smooth function $f(.)$ can be set with a linear form (for LMs) or with a non-linear form, e.g., with an expansion of splines bases in GAMs. Eventually, the error term $\varepsilon_{dt}$ is defined as a i.i.d. term.

DRAFT

## 4. Results

For the sake of brevity, here we report only the essential results relevant to the understanding of the phenomenon of interest. Extensive analyses and full results are available at the following GitHub webpage https://github.com/PaoloMaranzano/PM_KMC_PO_FC_ManureSAE.

The main findings can be summarized as follows. First, the use of SAE specifications that account for spatio-temporal effects strongly improves the fit to the data. In fact, both AIC and BIC identify STFH, which includes independent spatial and temporal effects, as the best model. Second, we consider the estimated quantities of manure and the degree of territorial heterogeneity. As previously stated, farms are mainly concentrated in the southern area of Lombardy, that is the Po Valley. Comparing the annual estimates reported in Figure 2, it can be observed that the direct estimator (upper panel), which does not use auxiliary information, shows poor territorial heterogeneity, except for some areas with higher density. On the contrary, the STFH estimates (lower panel), which make use of covariates, such as the number of animals, and that explicitly shape the spatiotemporal dynamics of the data, highlight how the agricultural subregions in the southeast process much more manure than the rest of the region. Further analyses show

that the heterogeneity of the SAE models is strongly guided by the total number of animal heads, that is, by the actual source of manure. Eventually, the variability of the estimates (i.e., the coefficient of variability) computed on the direct estimates tends to be much higher than that of the SAE models when the number of sampled companies in the areas is small, while for large samples, the variability is very similar.

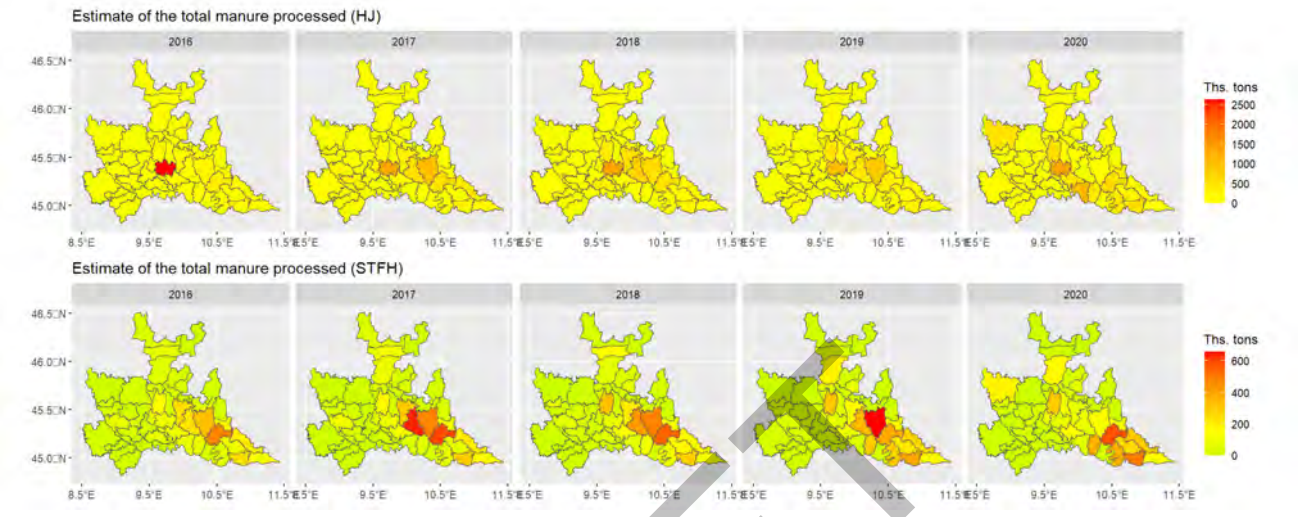



Figure 2: Upper panel: direct estimate (Hájek estimator) of the total manure processed by agrarian subregion. Lower panel: estimated total manure processed using the spatio-temporal Fay-Herriott model with independent temporal effects and spatial random effect (STFH model).

In Figure 3 we report the empirical relationship among the estimated quantity of manure processed from model STFH and the $NH_3$ emissions by livestock farms. The plot reveals a strongly positive but non-linear relationship between the two quantities, meaning that a high concentration of livestock farms and the following spread of manure is associated with a higher, but not proportional, emissions of $NH_3$ in the atmosphere. The non-linearity is well identified by the regression models reported in Table 1, in which the flexible regression models are able to detect the positive relationship while providing better fitting performances.

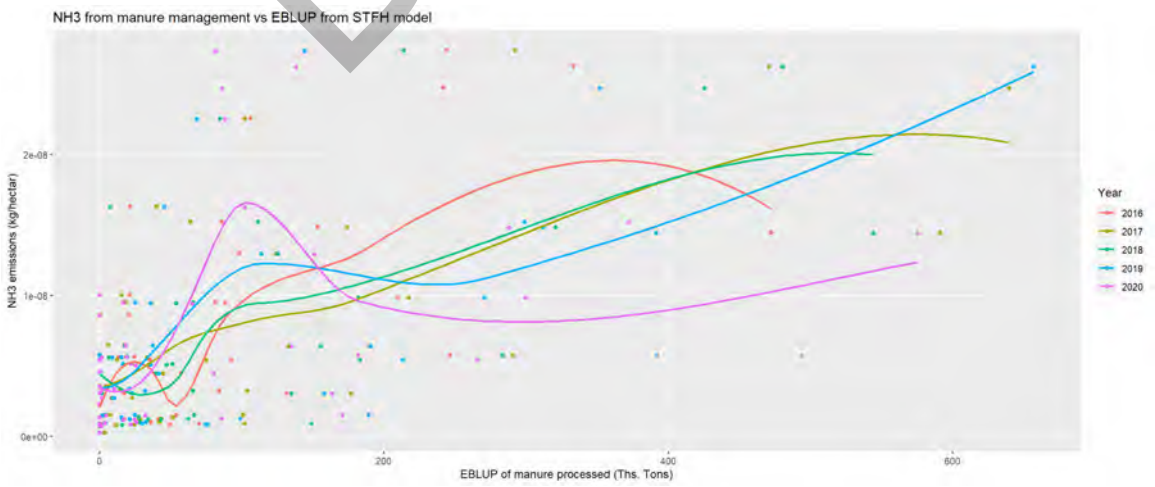



Figure 3: Empirical relationship between NH3 ammonia emissions from livestock sector (y-axis) and total manure estimated using the SAE models (x-axis)

DRAFT

| | LM | GLM (gamma) | GAM (norm) | GAM (gamma) | GAM (norm) | GAM (gamma) |
|---|---|---|---|---|---|---|
| (Intercept) | $0.000^{*}$ | $-153.385$ | $0.000$ | $-7.406$ | $0.000$ | $-12.307$ |
| | $(0.000)$ | $(201.350)$ | $(0.000)$ | $(31.445)$ | $(0.000)$ | $(30.522)$ |
| EBLUP manure (STFH-AR1) | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | $0.000^{***}$ | | |
| | $(0.000)$ | $(0.000)$ | $(0.000)$ | $(0.000)$ | | |
| Latitude | $-0.000^{*}$ | $2.524$ | | | | |
| | $(0.000)$ | $(3.998)$ | | | | |
| Longitude | $-0.000^{*}$ | $16.171$ | | | | |
| | $(0.000)$ | $(18.475)$ | | | | |
| Year | $-0.000$ | $0.009$ | $-0.000$ | $-0.006$ | $-0.000$ | $-0.003$ |
| | $(0.000)$ | $(0.043)$ | $(0.000)$ | $(0.016)$ | $(0.000)$ | $(0.015)$ |
| Latitude:Longitude | $0.000^{*}$ | $-0.354$ | | | | |
| | $(0.000)$ | $(0.408)$ | | | | |
| s(Latitude,Longitude) | | | $27.540^{***}$ | $27.062^{***}$ | $27.637^{***}$ | $27.096^{***}$ |
| | | | $(28.859)$ | $(28.756)$ | $(28.865)$ | $(28.750)$ |
| s(EBLUP manure) | | | | | $3.380^{***}$ | $2.693^{***}$ |
| | | | | | $(4.195)$ | $(3.354)$ |
| $R^2$ | 0.356 | 0.398 | 0.864 | 0.795 | 0.880 | 0.884 |
| AIC | $-8425.074$ | $-8622.193$ | $-8780.813$ | $-9021.406$ | $-8806.823$ | $-9030.910$ |
| BIC | $-8400.709$ | $-8597.828$ | $-8670.318$ | $-8912.395$ | $-8686.241$ | $-8914.096$ |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Estimated models: linear regression with Gaussian response (LM), generalized linear model with Gamma response (GLM), generalized additive model with Gaussian response (GAM - norm), and generalized additive model with Gamma response (GAM - gamma).

DRAFT

## 5. Conclusions and policy implications

Ammonia is released in large quantities by livestock production. Poor manure management practices, which are common issues worldwide, can lead to significant emissions of greenhouse gas methane, as well environmental degradation, negative health impacts, and the loss of valuable nutrients that could be added to soil. Suitable policies could facilitate changes in manure management practices at local levels. This study proposes a statistical approach that allows to provide a knowledge on the spatio-temporal relationships between livestock farms present in the sub-regional areas and the levels of ammonia, facilitating knowledge sharing and support for selective interventions of policy.

## References

[1] A. Fassó, J. Rodeschini, A. F. Moro, Q. Shaboviq, P. Maranzano, M. Cameletti, F. Finazzi, N. Golini, R. Ignaccolo, and P. Otto. Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data*, 10(1):143, 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02034-0. URL https://doi.org/10.1038/s41597-023-02034-0.

[2] D. Morales, M. D. Esteban, A. Pérez, and T. Hobza. A course on small area estimation and mixed models. *Methods, theory and applications in R*, 2021.

[3] J. Webb, H. Menzi, B. F. Pain, T. H. Misselbrook, U. Dämmgen, H. Hendriks, and H. Döhler. Managing ammonia emissions from livestock production in Europe. *Environmental Pollution*, 135(3): 399–406, 2005. ISSN 0269-7491. doi: https://doi.org/10.1016/j.envpol.2004.11.013. URL https://www.sciencedirect.com/science/article/pii/S0269749104004634.

[4] S. N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2006. ISBN 0429093152.

# Minimum contrast for point processes' first-order intensity estimation

Nicoletta D'Angelo[a] and Giada Adelfio[a]

[a]Department of Economics, Business and Statistics, University of Palermo;
nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it

**Abstract**

In this paper, we exploit some theoretical results, from which we know the expected value of the K-function weighted by the true first-order intensity function of a point pattern. This theoretical result can serve as an estimation method for obtaining the parameter estimates of a specific model, assumed for the data. The only requirement is the knowledge of the first-order intensity function expression, completely avoiding writing the likelihood, which is often complex to deal with in point process models. We illustrate the method through simulation studies for spatio-temporal point processes.

***Keywords:*** Second-order characteristics, Spatial statistics, Spatio-temporal point processes, Local models, Minimum contrast

DRAFT

## 1. Introduction

In this work, we introduce some preliminary results of a novel point processes' first-order intensity estimation procedure. We start from the theoretical result related to the expectation of the weighted $K$-function, using the true first-order intensity function. This theoretical result can serve as an estimation method for obtaining the parameters' estimates of a specific model assumed for the data. Indeed, parameters in spatial-temporal point process models are typically estimated by maximum likelihood method or some of its variants. Such a novel procedure may become crucial to avoid dealing with the complex likelihoods of some point process models, like the ETAS process, and their maximization. By further considering the local second-order characteristics, we can obtain the whole set of parameters assumed for the fitted model, one for each point of the analysed point pattern.

The structure of the paper is as follows. Section 2. recalls spatio-temporal $K$-function and its estimator. Section 3. introduces the proposed method. Section 4. shows some provisional simulation results. Section 5. outlines future works.

## 2. The spatio-temporal K-function and its estimator

We consider a spatio-temporal point process with no multiple points as a random countable subset $X$ of $\mathbb{R}^2 \times \mathbb{R}$, where a point $(\mathbf{u}, t) \in X$ corresponds to an event at $\mathbf{u} \in \mathbb{R}^2$ occurring at time $t \in \mathbb{R}$. A typical realisation of a spatio-temporal point process $X$ on $\mathbb{R}^2 \times \mathbb{R}$ is a finite set $\{(\mathbf{u}_i, t_i)\}_{i=1}^n$ of distinct points within a bounded spatio-temporal region $W \times T \subset \mathbb{R}^2 \times \mathbb{R}$, with area $|W| > 0$ and length $|T| > 0$, where $n \geq 0$ is not fixed in advance. In the sequel, $N(A)$ denotes the number of events of the process falling in a bounded region $A \subset W \times T$.

For a given event $(\mathbf{u}, t)$, the events that are close to $(\mathbf{u}, t)$ in both space and time, for each spatial distance $r$ and time lag $h$, are given by the corresponding spatio-temporal cylindrical neighbourhood of the event $(\mathbf{u}, t)$, which can be expressed by the Cartesian product as $b((\mathbf{u}, t), r, h) = \{(\mathbf{v}, s) : ||\mathbf{u} - \mathbf{v}|| \leq r, |t - s| \leq h\}, (\mathbf{u}, t), (\mathbf{v}, s) \in W \times T$, where $||\cdot||$ denotes the Euclidean distance in $\mathbb{R}^2$. Note that $b((\mathbf{u}, t), r, h)$ is a cylinder with centre (**u**, t), radius $r$, and height $2h$.

Product densities $\lambda^{(k)}$, $k \in \mathbb{N}$ and $k \geq 1$, arguably the main tools in the statistical analysis of point processes, may be defined through the so-called Campbell Theorem (2), which states that, given a spatio-temporal point process $X$, for any non-negative function $f$ on $(\mathbb{R}^2 \times \mathbb{R})^k$

$$\mathbb{E}\left[\sum_{\zeta_1,\ldots,\zeta_k \in X}^{\neq} f(\zeta_1, \ldots, \zeta_k)\right] = \int_{\mathbb{R}^2\times\mathbb{R}} \cdots \int_{\mathbb{R}^2\times\mathbb{R}} f(\zeta_1, \ldots, \zeta_k)\lambda^{(k)}(\zeta_1, \ldots, \zeta_k) \prod_{i=1}^{k} \mathrm{d}\zeta_i,$$

that constitutes an essential result in spatio-temporal point process theory. In particular, for $k = 1$ and $k = 2$, these functions are respectively called the *intensity function* $\lambda$ and the *(second-order) product density* $\lambda^{(2)}$.

(3) define the spatio-temporal inhomogeneous $K$-function and propose a non-parametric estimator for it.

**Definition 1.** *(3) A spatio-temporal point process is second-order intensity reweighted stationary and isotropic if its intensity function is bounded away from zero and its pair correlation function depends only on the spatio-temporal difference vector $(r, h)$, where $r = ||\boldsymbol{u} - \boldsymbol{v}||$ and $h = |t - s|$.*

**Definition 2.** *(3) For a second-order intensity reweighted stationary, isotropic spatio-temporal point process, the space-time inhomogeneous $K$-function is*

$$K(r, h) = 2\pi \int_0^r \int_0^h g(r', h')r'dr'dh'$$

*where $g(r, h) = \lambda^{(2)}(r, h)/(\lambda(\boldsymbol{u}, t)\lambda(\boldsymbol{v}, s))$, $r = ||\boldsymbol{u} - \boldsymbol{v}||$, $h = |t - s|$.*

The simplest expression of an estimator of the spatio-temporal $K$-function is given as

$$\hat{K}(r, h) = \frac{1}{|W||T|} \sum_{i=1}^{n} \sum_{j>i}^{n} \mathbf{1}(||\mathbf{u}_i - \mathbf{u}_j|| \leq r, |t_i - t_j| \leq h). \tag{1}$$

For a homogeneous Poisson process it holds $\mathbb{E}[\hat{K}(r, h)] = \pi r^2 h$, regardless of the intensity $\lambda$. The spatio-temporal $K$-function can be used as a measure of spatio-temporal clustering and interaction (3). Usually, the estimate $\hat{K}(r, h)$ is compared with the theoretical $\mathbb{E}[\hat{K}(r, h)] = \pi r^2 h$. Values $\hat{K}(r, h) > \pi r^2 h$ suggest clustering, while $\hat{K}(r, h) < \pi r^2 h$ suggests a regular pattern.

The spatio-temporal inhomogeneous version of the $K$-function in (1) is given by (3) as

$$\hat{K}_I(r, h) = \frac{|W||T|}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i} \frac{\mathbf{1}(||\mathbf{u}_i - \mathbf{u}_j|| \leq r, |t_i - t_j| \leq h)}{\hat{\lambda}(\mathbf{u}_i, t_i)\hat{\lambda}(\mathbf{u}_j, t_j)}, \tag{2}$$

where $\lambda(\cdot)$ is the first-order intensity at an arbitrary point. We know that $\mathbb{E}[\hat{K}_I(r, h)] = \pi r^2 h$, that is the same as the expectation of $\hat{K}(r, h)$ in (1), when the intensity used for the weighting is the true generator model. This is a crucial result that allows to use the weighted estimator $\hat{K}_I(r, h)$ as a diagnostic tool, for assessing the goodness-of-fit of spatio-temporal point processes with generic first-order intensity functions. Indeed, if the weighting intensity function is close to the true one $\lambda(\mathbf{u}, t)$, the expectation of $\hat{K}_I(r, h)$ should be close to $\mathbb{E}[\hat{K}(r, h)] = \pi r^2 h$ for the Poisson process. For instance, values $\hat{K}_I(r, h)$ greater than $\pi r^2 h$ indicate that the fitted model is not appropriate, since the distances computed among points exceed the Poisson theoretical ones.

DRAFT

Successively, (1) introduced local versions of both the homogeneous and inhomogeneous spatio-temporal $K$-functions, and used them as diagnostic tools, while also retaining for local information. Defining an estimator of the overall intensity by $\hat{\lambda} = n/(|W||T|)$, they propose the local version of (1) for the $i$-th event $(\mathbf{u}_i, t_i)$ as

$$\hat{K}^i(r,h) = \frac{1}{\hat{\lambda}^2|W||T|} \sum_{(\mathbf{u}_i,t_i)\neq(\mathbf{v},s)} \mathbf{1}(||\mathbf{u}_i - \mathbf{v}|| \leq r, |t_i - s| \leq h) \tag{3}$$

and the local version of (2) as

$$\hat{K}^i_I(r,h) = \frac{1}{|W||T|} \sum_{(\mathbf{u}_i,t_i)\neq(\mathbf{v},s)} \frac{\mathbf{1}(||\mathbf{u}_i - \mathbf{v}|| \leq r, |t_i - s| \leq h)}{\hat{\lambda}(\mathbf{u}_i, t_i)\hat{\lambda}(\mathbf{v}, s)}, \tag{4}$$

with $(\mathbf{v}, s)$ being the spatial and temporal coordinates of any other point. The authors proved that the inhomogeneous second-order statistics behave as the corresponding homogeneous ones, basically proving that the expectation of both (3) and (4) is equal to $\pi r^2 h$.

## 3. Proposal

Suppose that the intensity $\lambda(\mathbf{u}, t; \boldsymbol{\theta})$ (in brief: $\lambda(\boldsymbol{\theta})$) of the given point process model incorporates a vector of parameters $\boldsymbol{\theta} \in \Theta$. By minimizing

$$\mathcal{M}(\boldsymbol{\theta}) = \int_{h_0}^{h_{max}} \int_{r_0}^{r_{max}} \phi(r,h)\{(\hat{K}_I(r,h;\lambda(\boldsymbol{\theta})) - \pi r^2 h)\}^2 \mathrm{d}r\mathrm{d}h \tag{5}$$

with respect to $\boldsymbol{\theta}$, we obtain a vector of estimates $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{M}(\boldsymbol{\theta}).$$

Here $r_0$, $h_0$, $r_{max}$ and $h_{max}$ are the lower and upper space and time lag limits of the contrast criterion, and $\phi(r,h)$ is a weight that depends on the space-time distance.

For this purpose, we further suggest using a penalized objective function

$$\mathcal{M}^R_{tot}(\boldsymbol{\theta}) = \mathcal{M}(\boldsymbol{\theta}) + \mathcal{M}^R_{pen}(\boldsymbol{\theta})$$

by means of the radial penalization presented in (5)

$$\mathcal{M}^R_{pen}(\boldsymbol{\theta}) = \tau(||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_2 - R)^2,$$

where $||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_2 = \sqrt{\sum_{\boldsymbol{\theta}_j\in\Theta}(\theta_j - \hat{\theta}_j)^2}$, $\tau$ is the tuning parameter representing the penalization strength, and it is commonly selected as $1/R^2$ (5).

The penalty term $\mathcal{M}^R_{pen}(\boldsymbol{\theta})$ has its minimum at a sphere with radius $R$ centered around $\hat{\boldsymbol{\theta}}$. The final estimates are found as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{M}^R_{tot}(\boldsymbol{\theta}).$$

Our proposal poses the basis for a local extension. Suppose that the model incorporates a vector of parameters $\boldsymbol{\theta}$. Let $\hat{K}^i_I(r,h;\lambda(\boldsymbol{\theta}))$ denote the local estimators calculated from the data. For each point indexed by $i$ we consider

$$\mathcal{M}_{local}(\boldsymbol{\theta}_i) = \int_{h_0}^{h_{max}} \int_{r_0}^{r_{max}} \phi(r,h)\{(\hat{K}^i_I(r,h;\lambda(\boldsymbol{\theta})) - \pi r^2 h)\}^2 \mathrm{d}r\mathrm{d}h.$$

Then, we can obtain a vector of estimates $\hat{\boldsymbol{\theta}}_i$, one for each point $i$, as

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{M}_{local}(\boldsymbol{\theta}_i).$$

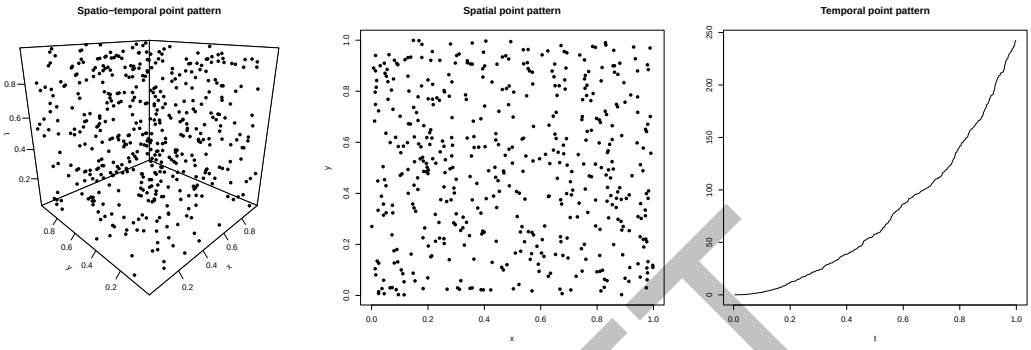
DRAFT

## 4. Simulations

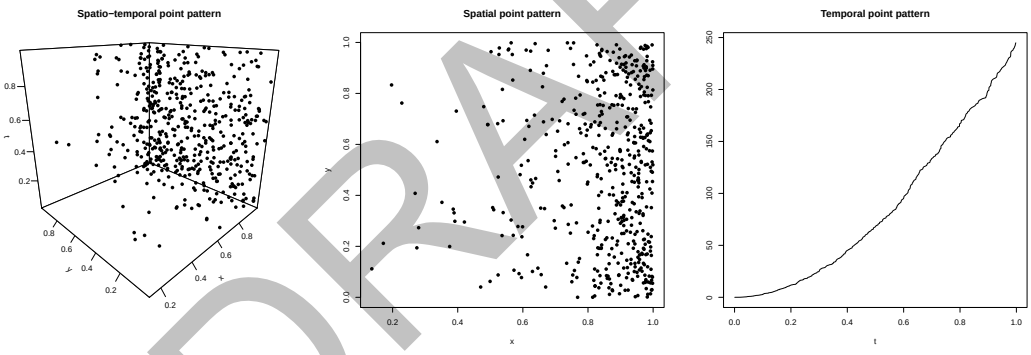
In this section, we report some provisional simulation results, for assessing the proposed estimation procedure. We simulate 1000 space-time point patterns in the unit cube with $n = 500$ number of points on average from the following point processes:

- homogeneous Poisson process with constant intensity $\lambda$;
- inhomogeneous Poisson process with intensity $\lambda(x, y, t) = \exp(\alpha + \beta x)$.

Examples of such simulated patterns are in Figure 1.




(a) Realization of a homogeneous point process with intensity $\lambda(x, y, t) = 500$




(b) Realization of an inhomogeneous point process with intensity $\lambda(x, y, t) = \exp(2 + 6x)$

Figure 1: Space and time locations of two examples of simulated homogeneous and inhomogeneous patterns.

Table 1 reports the $mean$, $RMSE$, and the mean of the standard errors ($mean(s.e.)$) of the estimated of the two considered processes, averaged over 1000 simulations. In particular, the spatial and temporal distances in the observed weighted $K$-functions are 15 values ranging from 0 to $r_{max}$ and $h_{max}$ in Equation (3), equal to 1/4 of the maximum (space or time) distances.

We notice that the mean of the intensity function for the homogeneous scenario appears systematically overestimated. This might be due to different approximations in the computation of the space-time $K$-function.

For the inhomogeneous point processes, we employ the penalized procedure, where the penalty enters like an additional data point which is used to "pull" in the parameter direction where the data provide the least information. The mean of the (unpenalized) estimates for the $\alpha$ and $\beta$ parameters are 4.3780 and 2.9476 (with 2.1560 and 2.3885 as standard errors, respectively). Adding the penalization, with radius $R = 2.5$ as in the purely spatial case, the estimated parameters are pulled towards the true values (see Table 1).

Further results, not shown for brevity, indicate that the proposed procedure works quite well also in the purely spatial case, especially when the parameter to estimate is unique. This scenario includes both

DRAFT

Table 1: Results over 1000 simulations.

| | | | MLE | | Min. Con. | | |
|---|---|---|---|---|---|---|---|
| Process | $\mathbb{E}[n]$ | True par | mean | RMSE | mean | RMSE | mean(s.e.) |
| Hom. Pois. | 500 | $\lambda = 500$ | 523.0372 | 517.3331 | 529.3286 | 39.1756 | 414.0448 |
| Inhom. Pois. | 500 | $\alpha = 2$ | 2.5129 | 2.4991 | 2.8496 | 2.3815 | 1.1852 |
| $R = 2.5$ | | $\beta = 6$ | 5.3527 | 2.4289 | 4.8887 | 2.3385 | 1.5173 |

a homogeneous specification and inhomogeneous one with only a slope (i.e. $\alpha = 0$). Moreover, we still encounter the same identifiability problem of the spatio-temporal context when dealing with multiple parameters. However, also in the purely spatial case, this problem can be overcome including a penalty in the objective function to minimize.

We note that the computational times of our proposal are longer than the MLE ones, increasing more than linearly with the number of points and the space and time lags considered in the employed $K$-function. Note instead, that the computational times are not influenced nor by $r_{max}$ or $h_{max}$.

## 5. Future work

The idea presented in this paper represents a work in progress. Therefore, much can be done in future. First, we want to run extended simulation studies to assess the performance of the proposed procedure in more complex settings, for instance, with Self-Exciting models such as the ETAS ones. A similar idea is being developed by (4), who showed that parameters in spatial-temporal point process models, alternatively to MLE, can be estimated consistently, under general conditions, by instead minimizing the Stoyan-Grabarnik (SG) statistic. Therefore, we wish to compare our proposal based on the $K$-function to both the Maximum Likelihood Estimation and (4)'s proposal. Moreover, we want to explore the local extension to fit local models to the data.

DRAFT

## Fundings

This work was supported by "FFR 2023 - Giada Adelfio", "FFR 2023 - Nicoletta D'Angelo", and by the PNRR project "Growing Resilient, INclusive and Sustainable - GRINS" Spoke 06: UNIPD "Low Carbon Policies".

## References

[1] Adelfio, G., Siino, M., Mateu, J., and Rodríguez-Cortés, F. J. (2020). Some properties of local weighted second-order statistics for spatio-temporal point processes. *Stochastic Environmental Research and Risk Assessment*, 34(1):149–168.

[2] Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

[3] Gabriel, E. and Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63(1):43–51.

[4] Kresin, C. and Schoenberg, F. (2022). Estimation of spatial-temporal point process models using a stoyan-grabarnik statistic. *METMA X*, page 13.

[5] Kreutz, C. (2018). An easy and efficient approach for testing identifiability. *Bioinformatics*, 34(11):1913–1921.

# Exploring the level of digitalization of the Italian museums through a multilevel ordered logit model

Claudia Cappello[a], Sabrina Maggio[a], and Sandra De Iaco[a]

[a]Department of Economic Sciences, University of Salento, Complesso Ecotekne, Lecce (Italy), claudia.cappello@unisalento.it, sabrina.maggio@unisalento.it, sandra.deiaco@unisalento.it,

**Abstract**

The Italian cultural heritage includes a wide range of museums, with different institutional features, types of collection, exhibition space and number of visitors. Museums contribute to enrich the Italian social and cultural background, since they are places in which education, research, exhibition and aesthetics harmoniously combine, in order to improve the visitor-experience. The use of digital technologies can be considered as a keystone of the attractiveness of the cultural sites. In this context, this paper aims to model the probability for the museums to adopt ever-increasing degrees of digitalization, by using a multilevel approach which takes into account the geographical locations of the investigated units. In particular, a multilevel multinomial ordered model will be implemented in order to evaluate how the level of digitalization of Italian museums is influenced by prominent factors which might stimulate to invest in new digital technologies.

***Keywords:*** multilevel ordered model; medium/low digitalization, high digitalization

DRAFT

## 1. Introduction

During the last decades, the digital transformation has involved museums, with the construction of more and more sophisticated websites, as well as the adoption of increasingly advanced digital technologies to improve the visitors experience ([6; 13]). Indeed, the proliferation of information technology is transforming all aspects of museum operations while enhancing the traditional functions.

In Italy, many initiatives have been promoted by the national Ministry of Culture over the past twenty years. The latest one dates back to July 2019 with the approval of the Three-Year Plan for the Digitalization and Innovation of Museums, aimed at supporting the digitalization process. Although the pandemic crisis has given a boost to digitalization, this topic has not been widely investigated in the literature. The contribution of [11] applied a mixed methodological approach consisting of a descriptive comparison of three museums acting in the Apulia region. Moreover, the study of [8] analyzed the use of information technologies and innovation processes in the Uffizi Gallery in Florence (Tuscany region), based on the Virtual Value Chain Model.

As far as is known, no work in this context has ever dealt with in-depth research at national level on the determinants of digital technologies. In such perspective, this paper aims to estimate the probability for museums to achieve a growing level of technological innovation (no, medium/low, high digitalization), by applying a multinomial multilevel ordered model. The net

effect deriving from the contribution of some key factors which might affect the level of digitalization will be also evaluated.

After a brief theoretical description of the multilevel logit ordered model (Sect. 2), some details will be provided on the dataset and on the methodology applied (Sect. 3) and the corresponding results will be given in Sect. 4.

## 2. Some theoretical hints on the multinomial ordered logit model

The multinomial ordered logistic model is a cumulative regression model which connects an ordinal variable of multiple categories to a set of indipendent variables (or covariates) ([4]). The first papers appeared in the literature concern multilevel regression models for ordinal data ([9; 10]). Additional examples of ordered logistic regression models can be found in [1; 2; 3; 5].

Let $Y_{ijk}$ be a multinomial ordered response variable with values $s = 1, 2, ..., t$ (response categories), where $t$ is selected as the reference category and with the index $i$ ($i = 1, \ldots, n_{jk}$) denoting the level 1 unit, the index $j$ ($j = 1, \ldots, N_k$) corresponding to the level 2 unit and the index $k$ ($k = 1, \ldots, K$) representing the level 3 unit.

Given the probability $\pi_{ijk}^{(s)}$ that the $i$-th first level unit presents a response variable value of $s$, let $\{X_{1\cdot}, X_{2\cdot}, \ldots, X_{H\cdot}\}$ be a set of covariates which influences the dependent response variable.

In order to consider the ordering, the identified model is based on the cumulative response probabilities defined as follows:

$$E(Y_{ijk}^{(s)}) = \gamma_{ijk}^{(s)} = \sum_{h=1}^{s} \pi_{ijk}^{(h)}, \qquad s = 1, 2, \ldots, t-1$$

where $Y_{ijk}^{(s)}$ describes the observed cumulative proportions for the $i$-th unit.

The category probabilities can be expressed in terms of the cumulative probabilities, thus obtaining:

$$\pi_{ijk}^{(h)} = \gamma_{ijk}^{(h)} - \gamma_{ijk}^{(h-1)}, \qquad 1 < h < t$$

$$\pi_{ijk}^{(1)} = \gamma_{ijk}^{(1)}; \qquad \gamma_{ijk}^{(t)} = 1$$

A multinomial ordered model with a logit link can be formalized as:

$$\gamma_{ijk}^{(s)} = \left\{ 1 + \exp\left[ -\left( \beta_{0jk}^{(s)} + \sum_{h=1}^{H} \beta_{hjk}^{(s)} x_{hijk} \right) \right] \right\}^{-1}$$

or analogously:

$$\eta_{ijk} = \log it(\gamma_{ijk}^{(s)}) = \beta_{0jk}^{(s)} + \sum_{h=1}^{H} \beta_{hjk}^{(s)} x_{hijk}$$

with

- $\beta_{hjk}^{(s)} = \beta_{h}^{(s)} + \nu_{hk}^{(s)} + u_{hjk}^{(s)}, \qquad h = 0, 1, \ldots, H$
- $\begin{bmatrix} \nu_{0k}^{(s)} \\ \nu_{1k}^{(s)} \\ \vdots \\ \nu_{hk}^{(s)} \\ \vdots \\ \nu_{Hk}^{(s)} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{\Omega}_{\nu}),$
- $\begin{bmatrix} u_{0jk}^{(s)} \\ u_{1jk}^{(s)} \\ \vdots \\ u_{hjk}^{(s)} \\ \vdots \\ u_{Hjk}^{(s)} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{\Omega}_{u}).$

DRAFT

This implies that increasing values of the linear components are associated with increasing probabilities as $s$ grows ([12]). All the terms $\beta_{hjk}^{(s)}$ vary across categories ($s = 1, 2, \ldots, t-1$); analogously for the random-effect variance term ([5]).

## 3. Data and methods

The microdata used in this paper concern a census survey on the Italian public and private museums carried out by the Italian National Institute of Statistics (ISTAT) in 2018. In this context, a multilevel multinomial ordered model has been implemented in order to measure the probability for the museum of being without significant digital innovations, or adopting medium/low or high level of digitalization. In particular, three hierarchical levels have been analyzed: *the first level*, that is the museums (3,217 museums); *the second level*, corresponding to the Italian provinces where the museums are placed (108 provinces); *the third level*, concerning the Italian regions where the museums are located (20 regions).

The choice of three levels of aggregation is justified by the underlying hierarchical structure of data, where the regions represent the highest level in which the cultural heritage can offer different opportunities; on the other hand, the museums are considered as the lowest level of nesting. A detailed descriptive analysis on the ISTAT microdata has been conducted on the features of museums, distinguished by management, access, visits, staff, financial resources, structures, support of fruition, relationship with the territory, activities and services.

The multinomial scheme of the model allows to catch the effects that regional and provincial characteristics might have on the level of digitalization, which represents the dependent variable sorted into three categories (none, medium/low, high digitalization). More specifically, the level of digitalization has been recoded considering up to 15 digital technologies, including digital inventory, digital catalogue, video and audio guides, applications for smartphones and tablets, among others. From the descriptive analysis, the following covariates have been derived:

- network of museums, access absolutely free or with admission fee, research activities, partnership, guided tours, exhibition space, type of institution, percentage of Italians vs Foreigners (Individual covariates);
- number of tourist accommodation establishments (Provincial-level covariate);
- expenditure for recreation, culture and religion, Gross domestic product per capita, current prices (Regional-level covariates).

The selected covariates have been considered for evaluating their effects on the probability of catching ever-growing levels of digitalization for museums.

DRAFT

## 4. Results and discussion

According to the modeling results, it has to be pointed out that apart from the covariate "Access", which has a negative impact on the probability to have a medium/low or high level of digitalization, all the other covariates listed in Sect. 3 produce a positive effect on the probability to introduce both a medium/low and a high level of digitalization. In addition, from the estimates of the multinomial ordered logit model, it has to be highlighted that it is much more probable that museums in Italy are likely to introduce a medium/low level of digitalization (with estimated values ranging from 0.495 to 0.714) than a high level of digitalization (values from 0.141 to 0.495).

For what concerns the regional-level, Lazio, Tuscany, Emilia-Romagna, Trentino-Alto Adige, Lombardy and the islands of Sardinia and Sicily are the Italian regions with the highest estimated probability of adopting at least a medium/low level of digitalization, with a number of structures ranging from 201 to 553 museums. Besides these regions, also Umbria, Apulia and

Basilicata have to be mentioned, with up to 170 museum structures in their territories. As regards the provincial level, the greatest estimated probability of adopting mainly high digitalization can be found for the provinces of Milan, Mantova, Cremona and Como (Lombardy), as well as for Ferrara (Emilia-Romagna), Venice (Veneto), Naples (Campania), Prato (Tuscany) and Rome (Lazio).

**Acknowledgements** This research has been partially supported by the Consortium CUIS (grant given in 2018).

## References

[1] Agresti, A.: Analysis of Ordinal Categorical Data. 2nd ed. Wiley, New York (2010)
[2] Agresti, A., Natarajan, R.: Modeling Clustered Ordered Categorical Data: A Survey. Int. Stat. Rev. **69**(3), 345–371 (2001)
[3] Fullerton, A.S.: A conceptual framework for ordered logistic regression models. Sociol. Methods Res. **38**(2), 306–347 (2009)
[4] Grilli, L., Rampichini, C.: Multilevel models for ordinal data. In: Kenett, R., Salini, S. (eds.) Modern Analysis of Customer Surveys: with Applications using R, pp. 391–411. Wiley, Chichester (2012)
[5] Hedeker, D.: Multilevel models for ordinal and nominal variables. In: de Leeuw, J., Meijer, E. (eds.) Handbook of Multilevel Analysis, pp. 237–274. Springer, New York (2008)
[6] Hooper-Greenhill, E.: Education, communication and interpretation: towards a critical pedagogy in museums. In: Hooper-Greenhill, E. (ed.) The Educational Role of the Museum, pp. 3–27. London, New York, Routledge (1999)
[7] ISTAT. Survey on museums and other cultural institutions: public use micro.stat files, (2018). https://www.istat.it/en/archive
[8] Lazzeretti, L., Sartori, A.: Digitization of Cultural Heritage and Business Model Innovation: The Case of the Uffizi Gallery in Florence. J. Cult. Herit. **14**, 945–970 (2016)
[9] McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd ed. Chapman and Hall, New York (1989)
[10] McKelvey, R.D., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables. J. Math. Sociol. **4**(1), 103–120 (1975)
[11] Raimo, N., De Turi, I., Ricciardelli, A., Vitolla, F.: Digitalization in the cultural industry: evidence from Italian museums. Int. J. Entrepreneurial Behav. Res. (2021). https://doi.org/10.1108/IJEBR-01-2021-0082
[12] Rasbash, J., Steele, F., Browne, W.J., Goldstein, H.: A User's Guide to MLwiN Version 2.10, Centre for Multilevel Modelling, University of Bristol, United Kingdom, 3rd ed. (2009)
[13] Romanelli, M.: Museums creating value and developing intellectual capital by technology: from virtual environments to Big Data. Meditari Account. Res. **26**(3), 483–498 (2018)

DRAFT

# Functional Partial Least-Squares via Regression Splines. An application on Italian Sustainable Development Goals data

Ida Camminatiello[a], Rosaria Lombardo[a], Jean-François Durand[b], and Leonardo S. Alaimo[c]

[a]University of Campania "L.Vanvitelli"; ida.camminatiello@unicampania.it, rosaria.lombardo@unicampania.it
[b]Montpellier II University; jf.durand001@orange.fr
[c]University of Rome "La Sapienza"; leonardo.alaimo@uniroma1.it

DRAFT

**Abstract**

Functional regression is a statistical method that is used to model the relationship between a response variable and a set of predictor variables that are functions. Functional Partial Least-Squares (PLS) regression is a form of functional regression analysis that is particularly useful when the number of predictors is large compared to the number of observations, or when the predictors are highly correlated. The basic idea of functional PLS via regression splines is to transform both response and predictors by using a set of spline-basis functions, such as $B$-spline basis, and then use the standard PLS technique to estimate the optimal transformed predictors. We show its performance on a real data set concerning the sustainable development goals of Agenda 2030.

***Keywords:*** $B$-splines, Nodal coefficients, Partial Least-Squares, Sustainability

## 1. Introduction

Functional data analysis (9) has gained significant attention in recent years due to technological progress allowing for the collection of high-dimensional functional data in various fields (biology, econometrics, environmetrics, sustainability, etc.). Hence, the result of one observation can be viewed as a discretized version of one curve (Near InfraRed spectrum, radar waveform, physiological signals), or of one surface (3D image). This kind of high-dimensional data needs new functional statistical tools. The basic idea of functional regression is to represent the predictor variables as a linear combination of basis functions, such as Fourier basis or $B$-spline basis, and then use standard regression techniques to estimate the coefficients of the basis functions. Unfortunately, these methods often transform the functional model into a multiple regression model with high multicollinearity. To address this issue, several approaches have been proposed, among which functional principal component regression and functional partial least squares (PLS) regression (2; 6; 10; 18; 1; 11) which estimate a scalar response from predictor curves.
Due to the great interest raised by PLS, we focus on functional PLS regression. To our knowledge, there is no work for dealing with both predictor and response functional variables in the PLS regression. In Section 2, we aim to develop an approach for estimating a functional response from functional predictors. The last section concludes with an application on Italian sustainability development goals data.

## 2. Functional Partial Least Squares Regression

Partial least-squares (PLS) is a statistical technique used for regression analysis, where the goal is to find a linear relationship between the response and predictor variables. PLS is particularly useful when the number of predictor variables is large, and there may be some degree of multicollinearity or correlation between them (14; 15; 16).

Similarly, functional partial-least squares via regression splines allows to model the relationship between a response variable and a set of predictor variables highly correlated where the variables of interest are not simple scalars, but instead are spline functions. Different types of PLS functional regression can be discussed, depending on the type of transformation functions for the predictor variables and the nature of the response variable; see (2; 6; 10; 18; 1; 11). For example, when only the predictor variables are functions, Durand (6) proposed an extension of the linear PLS method towards nonlinearity via regression splines and called this method partial least-squares splines with the acronym PLSS.

Here we consider that the response is a spline function or a multifunction valued in a set of some $B$-spline functions, and that the new predictors are $B$-spline functions associated to each original predictor. The PLSS regression model is used to estimate the coefficients of these basis functions. We call this model Functional Partial Least-Squares via regression Splines (FPLSS). The crucial property of the $B$-spline family associated to $x$, the generic variable among the $p$ predictors, is that it constitutes a basis for the linear space of the spline functions $s(x)$ that are piecewise polynomials of degree $d$ that join end to end with some regularity on $K$ points called the knots. So, $s(x)$ is a linear combination of a set of $r = d + 1 + K$ basis functions called the $B$-splines, see (12). The idea behind PLSS and then behind FPLSS is that the estimated functional response becomes a linear combination of $A$ pseudo-predictors $t_1, ..., t_A$, called the components, the latent variables or the base learners, see (3), where $A$ is generally estimated by cross-validation. Therefore, the generic component can be written as

$$t_i = s_1^i(x_1) + ... + s_p^i(x_p), \quad i = 1, ..., A. \tag{1}$$

Given the nonlinear additive relationship between the components $t_i$ ($i = 1, ..., A$) and the predictors $x_j$ ($j = 1, ..., p$), in order to interprete the influence of the predictors on the component $t_i$ we look at the shape of the coordinate functions $\{s_j^i(.)\}_{j=1,...,p}$. These functions are classified in decreasing order of influence according to the range of their values (recall that the predictors are normalized). FPLSS has some challenges, such as the need to choose the sets of appropriate basis functions. We adopt an heuristic strategy consisting in increasing/decreasing progressively the spline parameters, see (6), default values for the knot locations being the quantiles or equally spaced positions. As a summary, the tuning parameters of FPLSS are the degree, the knots (number and location) for the predictors, and the number $A$ of components.

Inherited from linear PLS, functional PLSS can handle

- a large number of predictors possibbly highly correlated and a small number of observations;
- continuous and categorical response and predictor variables;
- nonlinear relationships and bivariate interactions through tensor products of $B$-splines, see (8).

The FPLSS algorithm using basis-spline functions involves the following steps:

- Create different sets of $B$-spline functions to transform both response and predictors.
- Use these $B$-splines and PLS to create a new set of pseudo predictors in small number (the PLSS components) that nonlinearly capture in (1) the variation in the predictors.
- Use the PLSS regression model (2) to predict the functional response.

The FPLSS model can be written as

$$S(\hat{y}(A)) = \sum_{j=1}^{p} \sum_{l=1}^{r_j} \hat{\beta}_l^j(A) B_l^j(x_j), \tag{2}$$

where $S(y)$ is a functional transformation of the response $y$ by splines that aims at producing a good model (goodness of fit and prediction) as well as easy interpretable results.

DRAFT

For example, it can be $S(y) = s_{id}(y)$, the "spline function identity", where $s_{id}(y) = y$ for any $y$. See Shumaker (12) for its construction when $d > 0$ by using the so called "nodal" $\beta$ coefficients defined by the mean of some knot locations. The spline identity has been also used by Durand (4; 2) in nonlinear multiple regression as the initial step of a sequence of transformations of the predictors through regression splines. In FPLSS the choice $S(y) = s_{id}(y)$ leads to FPLSS = PLSS.

Inspired by the rather formal preceding option for $S(y)$, we propose a sensitive strategy for the response transformation consisting in a local departure from the spline identity by slightly perturbing the $\beta$ nodal coefficients of $s_{id}(y)$. Therefore, decision makers have the ability to controll on-line what is changing in the model (1) and (2) by locally modifying the values of the observed response.

At last, a multi-function $S(y)$ valued in a set of binary disjunctive functions ($B$-splines of degree 0) leads to functional discriminant PLS regression, see (3; 5), where the components play the role of discriminant variables of the constructed response groups. Added to cross-validation, a way of assessing the goodness of the model is to check in the $\{t_i\}_{i=1}^{A}$ space, the observations that are well affected to the "true groups" by using their Mahalanobis distances to the centroids.

## 3. Example: Sustainable Development Goals Data

Equation (1) of FPLSS finds a straightforward application in the context of functional discriminant analysis. The response Migration Policy, *MigraPol*, is split up according to 2 break-points (the knots) to decompose the observations (20 Italian regions) into 3 groups characterized by 3 $B$-splines of degree 0 that constitute the new responses $\{MigraPoli\}_{i=1}^{3}$. Each response indicates (by 0 or 1) respectively low, mean and large values of MigraPol. Then, the multiresponse PLSS algorithm has been run using the free open-source package "Boosted PLS Regression" available at *http://www.jf-durand.pls.com*. Here, the aim is to study $p = 49$ socio-economic variables with regard to the 3 response variables *MigraPol1*, *MigraPol2* and *MigraPol3*.

The 2030 Agenda has been designed to be an universal and global action plan to address the challenges of the 21st century. The Sustainable Development Goals (SDGs) are a set of 17 goals. Here, we consider the Italian Strategy for Sustainable Development (see SDGs 2021 Report: Statistical information for the 2030 Agenda in Italy, published by ISTAT) which concerns the measurement of the sustainable development across the 20 Italian regions. In particular, we focus our attention on Migration Policy which belongs to SDG 10 (reducing inequalities) and is described by a composite indicator constructed by using five simple indicators: 1) Permits issued for non-EU citizens; 2) Share of long-term permits; 3) New permits issued; 4) Percentage of permits issued for political asylum and humanitarian reasons; 5) Citizenship acquisitions. Indeed, Migration Policy and SDGs are interconnected, as Migration Policy can both contribute to and hinder the achievement of the SDGs. The determinants or predictors of Mi-

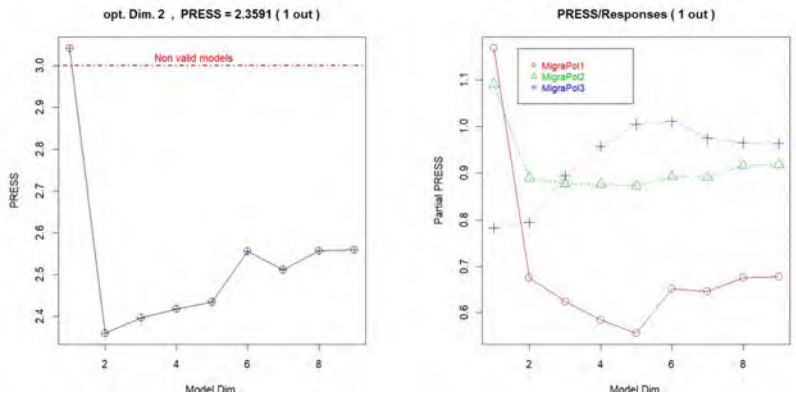



Figure 1: The PRESS criterion for the choice of the number of FPLSS components

gration Policy belonging to SDGs 3, 4 and 8 consist of $p = 49$ indicators. Here for brevity, instead of

DRAFT

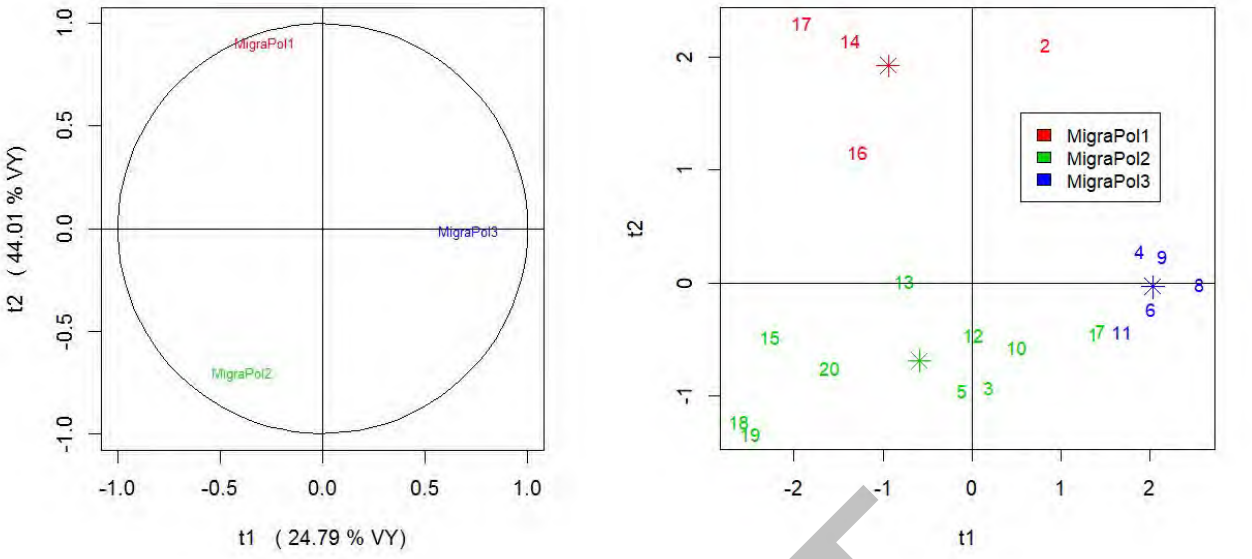



Figure 2: Correlation circle and observation plot. Note that 1=Piemonte, 2=ValleAosta, 3=Liguria, 4=Lombardia, 5=TrentinoAA, 6=Veneto, 7=FriuliVeneziaGiulia, 8=EmiliaRomagna, 9=Toscana, 10=Umbria, 11=Marche, 12=Lazio, 13=Abruzzo, 14=Molise, 15=Campania, 16=Puglia, 17=Basilicata, 18=Calabria, 19=Sicilia, 20=Sardegna.

Table 1: FPLSS model performance in fit

| | $R^2$ of the responses on the t subspaces | | | var Y = 3 | reconstituted var Y=2.5658 | |
|---|---|---|---|---|---|---|
| dimension | *MigraPol1* | *MigraPol2* | *MigraPol3* | Y var | % | cumul. % |
| 1 | 0.081 | 0.154 | 0.509 | 0.7436 | 24.79 | 24.79 |
| 2 | 0.896 | 0.659 | 0.509 | 1.3204 | 44.01 | 68.80 |
| 3 | 0.897 | 0.861 | 0.808 | 0.5018 | 16.73 | 85.53 |

DRAFT

providing the list of these predictors we briefly present the SDGs 3, 4 and 8.
SDG 3: "Ensure healthy lives and promote well-being for all at all ages" (**HealthWell**), described by 12 indicators. SDG 4: "Ensure inclusive and equitable quality Education and promote lifelong learning opportunities for all" (**QualEdu**), explained by 27 indicators. SDG 8: "Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all, and an enhanced productive capacity for least developed regions" (**WorkGrowth**), related to 10 indicators.
After transforming the predictors (by using B-splines of degree 2 with 2 knots at quantiles), it results that the predictive ability of the model is good, and the optimal leave-one-out PRESS suggests to consider 2 components ($PRESS_{total,2} = 2.359$). However, we prefer to take 3 components with similar PRESS ($PRESS_{total,3} = 2.396$) that better discriminate the three groups of observations since three components leads to 0 error of misclassication. So, three components explain 85.53% of the response variability. Table 1 shows the percentage of $S(\hat{y}(A = 3))$ variability with respect to each model dimension. Furthermore, it results that the FPLSS model fits each response very well, i.e. $R^2_{MigraPol1} = 0.897$, $R^2_{MigraPol2} = 0.861$ and $R^2_{MigraPol3} = 0.808$.

When performing FPLSS regression, it is usual to visualize the relationship between the responses and the latent variables by using the correlation circle coupled with the scatterplot of the components (see Figure 2). On the left side of Figure 2, the correlation circle on the plane $t_1, t_2$, shows a high negative correlation between the two responses *MigraPol1* and *MigraPol2* which are strongly correlated with the

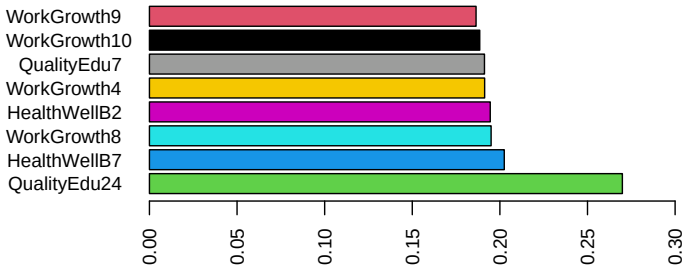



Figure 3: The most important predictors for $t_1$ and *MigraPol3*

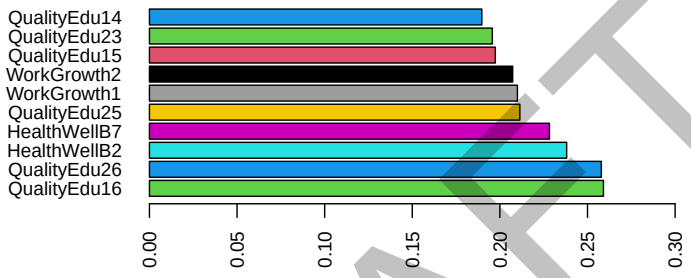



Figure 4: The most important predictors for $t_2$, *MigraPol1* and *MigraPol2*

DRAFT

second component $t_2$. While *MigraPol3* is correlated with the first component $t_1$. Moreover, the right side of Figure 2 shows the FPLSS $(t_1, t_2)$ score plot. We can see three colored groups of regions related to the responses *MigraPol1*, *MigraPol2* and *MigraPol3*, the group centroid is marked with a star. Component $t_1$ separates the blue group (4=Lombardia, 9=Toscana, 8=EmilaRomagna, 6=Veneto, 11=Marche) of high migration policy *MigraPol3* from the two other groups. Also, component $t_2$ separates the red group (4=ValleDAosta, 14=Molise, 16=Puglia and 17=Basilicata) of low migration policy from the large green group (13=Abruzzo, 15=Campania, 20=Sardegna, 18=Calabria, 19=Sicilia, 12=Lazio, 10=Umbria, 5=TrentinoAltoAdige, 3=Liguria, 1=Piemonte, 7=FriuliVeneziaGiulia) of mean migration policy.

In linear PLS, to interpret the role of the predictors on $t_i$ one can look at their correlations displayed on the correlation circles. Here, equation (1) shows a nonlinear additive relationship, so it is informative to look at the coordinate function plots $s_j^i(x_j), j = 1, ..., p$ for $t_i$. However, for the sake of brevity we just display the barplots of the most influent predictors classified in descending order, see Figures 3 and 4. These figures show the most relevant predictors for the components in descending order according to the range of the values of the coordinate spline functions. We can observe a high influence on $t_1$ of:

- the response *MigraPol3* and the predictors **QualityEdu24** (schools with pupils with disabilities), **HealthWellB7** (beds in ordinary hospitalization in public and private healthcare institutions) **WorkGrowth8** (employment rate (20-64 years)), **HealthWellB2** (excess weight), **WorkGrowth4** (employed in fixed-term jobs for at least 5 years), **QualityEdu7** (inadequate listening comprehension of the English language), **WorkGrowth10** (young people who are not working or studying (15-24 years)) and **WorkGrowth9** (young people who are not working or studying).

Furthermore, looking at Figure 4, we can note a high influence on $t_2$ of:

- the responses *MigraPol1* (large positive values) and *MigraPol2* (large negative values), and the predictors **QualityEdu16** (pupils with disabilities), **QualityEdu26** (secondary schools with pupils

with disabilities), **HealthWellB2** (excess weight), **HealthWellB7** (beds in ordinary hospitalization in public and private healthcare institutions), **QualityEdu25** (primary schools with pupils with disabilities), **WorkGrowth1** (annual growth rate of real GDP per employee), **WorkGrowth2** (annual growth rate of value added in volume per employee), **QualityEdu15** (disabled pupils in primary school), **QualityEdu23** (physically accessible schools), and **QualityEdu14** (pupils with disabilities in kindergarten).

In summary, knowing how different (or not) the migration policy is across Italian regions can help policymakers in deciding what reform would be helpful to reduce inequalities.

## References


[1] Boente, G., Vahnovan, A. (2017). Robust estimators in semi-functional partial linear regression models. *Journal of Multivariate Analysis*, **154**. 59–84.

[2] Durand, J.F., Sabatier, R. (1997). Additive splines for partial least squares regression. *Journal of the American Statistical Association*, **92**, 440–467.

[3] Durand, JF, (2008). La régression PLS boostée. *Revue Modulad*, **38**, 63-86.

[4] Durand, JF, (1993). Generalysed principal component analysis with respect to instrumental variables via univariate spline transformations. *Computational Statistics & Dtata Analysis*, **16**, 423-440.

[5] Lombardo, R., Durand, J. F. & Leone, A. P. (2012). Multivariate PLS spline boosting in Agro-Chemistry studies. *Current Analytical Chemistry*, *vol. 8, no 2*, 236-253.

[6] Durand, J.F. (2001). Local polynomial additive regression through PLS and splines: PLSS. *Chemometrics and Intelligent Laboratory Systems*, *58*, 235–246.

[7] Ferraty, F., and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

[8] Lombardo, R., Durand, J. F. & De Veaux, R. (2009). Model building in multivariate additive partial least squares splines via the GCV criterion. *Journal of Chemometrics*, *23*, 605– 617.

[9] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). Springer.

[10] Reiss, P. T., Ogden, R. T. (2007). Functional Principal Component Regression and Functional Partial Least Squares. *Journal of the American Statistical Association*, **102** (479), 984-996.

[11] Saricam, S., Beyaztas, U., Asikgil, B., Shang, H.L. (2022). On partial least-squares estimation in scalar-on-function regression models. *Journal of Chemometrics* doi.org/10.1002/cem.3452.

[12] Shumaker, L. L. (1981). *Spline Functions: Basic Theory.* John Wiley & Sons: New York, Chichester, Brisbane, Toronto.

[13] Tenenhaus, M. (1998). *La Régression PLS, Théorie et Pratique*. Editions Technip: Paris.

[14] Wold, H. (1966). *Estimation of principal components and related models by iterative least squares*. Multivariate Analysis. (Eds.) P.R. Krishnaiah, New York: Academic Press, 391–420.

[15] Wold, H. (1975). *Soft modelling by latent variables: Non linear Iterative Partial Least Squares approach*. Perspectives in Probability and Statistics: Papers in honour of Bartelett. (Eds.) J. Gani, London: Academic Press, 117–142.

[16] Wold, H. (1985). *Partial Least Squares*. Encyclopedia of Statistical Sciences (vol. 6). (Eds.) S. Kotz, N. L. Johnson, New York: Wiley, 581–591.

[17] Wold, S. (1978). Cross-validation estimation of the number of components in factor and principal components analysis. *Technometrics*, *24*, 397–405.

[18] Zhou, Z. (2021)Fast implementation of partial least squares for function-on-function regression. *Journal of Multivariate Analysis*, **185** 104769.


DRAFT

# Assessing multidimensional poverty of the Italian provinces during Covid-19: a small area estimation approach

Mariateresa Ciommi[a], Chiara Gigliarano[b], Francesca Mariani[a], and Gloria Polinesi[a]

[a]Università Politecnica delle Marche; m.ciommi@staff.univpm.it, f.mariani@staff.univpm.it, g.polinesi@staff.univpm.it
[b]LIUC - Università Carlo Cattaneo; cgigliarano@liuc.it

DRAFT

**Abstract**

The aim of this paper is to analyse the effect of Covid-19 on multidimensional poverty in Italy and its provinces by measuring changes in individual poverty before and during the pandemic outbreak. To capture the multidimensional nature of poverty, we consider different dimensions: economic well-being, health condition, education, neighborhood quality, subjective well-being.

The empirical application is based on micro-data from the 'Aspects of daily life' (AVQ) survey by Istat for the years 2019 (pre-Covid period) and 2020 (Covid period).

Since survey direct estimates are reliable only at regional (NUTS 2) level, the introduction of small area estimation (SAE) techniques becomes of crucial importance to monitor and contrast the phenomenon at a finer geographical level.

***Keywords:*** Multidimensional poverty index, composite indicators, small area estimation, EBLUP estimator.

## 1. Introduction

In the recent years there has been a considerable agreement that poverty is a multidimensional phenomenon, which cannot be adequately explained by considering only monetary variables. (1) confirms the existence of significant mismatches between monetary poverty and multiple deprivations and, therefore, the need to include many aspects of poverty, such as health, education, housing, satisfaction with life, security.

Few empirical studies have analyzed multidimensional poverty in Italy at sub-national level; see, among others, (5), (4) and (3). In this paper, we propose a deeper attempt in this direction by looking at the provincial level changes in the multidimensional poverty due to Covid-19 pandemic.

The empirical application is based on micro-data from the 'Aspects of daily life' (AVQ) Italian survey over the years 2019 (pre-Covid period) and 2020 (Covid period).

Since the AVQ survey is planned to obtain precise estimates at regional level, some alternative solutions should be evaluated to estimate poverty at provincial level. More specifically, two main possible strategies can be employed: i) increasing the sample size of AVQ for the specific province of interest (oversampling) so that direct estimates become reliable and ii) apply small area estimation (SAE) techniques; see, among others, (12) and (9).

We apply the Fay-Herriot (FH) small area estimation model proposed by (6) to estimate dashboard indicators at provincial level. Then, we propose an aggregation technique based on the Adjusted Mazziotta-Pareto Index (see (7) and (8)) for the FH estimates in order to obtain reliable estimates of the multidimensional poverty for the local areas of interest as in (11).

Therefore, the aim is to enhance the knowledge of the spatial distribution of the multidimensional poverty index at local level in Italy, focusing on Italian provinces, in order to help the policy maker to address resources towards the areas where the phenomenon is strongly present.

The empirical analysis reveals that the traditional Italian divide North vs. South becomes less clear when analyzing poverty at provincial level and for the different domains of poverty.

The remainder of the paper is as follows. Sections 2. and 3. describe data and methodologies considered in the analysis. Section 4. is devoted to results of the empirical application. Finally, Section 5. draws some conclusions.

## 2. Data

Our analysis is based on data from the 'Aspects of daily life' (AVQ) survey conducted by Istat, focusing on the years 2019 and 2020. The units of analysis are the Italian households and the original samples include 19536 and 18529 households, for the two years considered respectively.

To assess the multidimensional poverty over 105 Italian provinces[1], we consider five domains (health, education, economic well-being, neighborhood quality and subjective well-being), each composed of different dashboard indicators as described in Table 1.

In order to classify a household as deprived in each of the 13 dashboard indicators, we first identify the deprivation cut-offs following the approach proposed in (5), which we have slightly modified due to the data availability at provincial level. The deprivation cut-offs refers to the head of the household.

Table 1: Multidimensional poverty framework: domains, dashboard indicators, deprivation cut-offs.

| **Domains** | **Deprivation Indicators** | **Deprivation cut-offs** |
|---|---|---|
| *Health* | Nutrition | A person is deprived if s/he consumes less than 3 portions of fruits or vegetables a day. |
| *Education* | Educational deprivation | A person is deprived if s/he has not completed higher-secondary school. |
| | Cultural deprivation | A person is deprived if in the 12 months before the interview s/he has joined less than 2 among the following activities: 1) at least once to cinema, theatre, exhibitions and museums, archaeological sites, monuments, concerts of classical music, opera, concerts of other kind of music; 2) read the newspaper at least once a week; 3) read at least a book. |
| *Economic well-being* | Material deprivation | A person is deprived if s/he possesses less than 4 out of 6 following items: washing machine, color tv, scooter/moto or car, phone, personal computer. |
| | Housing deprivation | A person is deprived if s/he experiences 3 or more among the following deprivations related to the house: overcrowding; distance from basic services (pharmacy, shops, school); overall poor condition of the floors and/or walls; expenses too high; house not owned). |
| | Gas | A person is deprived if the house is not served by methane gas. |
| | Water | A person is deprived if s/he declares to have irregularities in their water supply. |
| | Unemployment | A person is deprived if s/he is unemployed. |
| | Financial distress | A person is deprived if her/his economic sources are not sufficient to make ends meet. |
| *Neighbourhood quality* | Noise<br>Crime<br>Pollution | A person is deprived if the area in which s/he lives is declared to be very noisy; at risk of crime; polluted. |
| *Subjective well-being* | Life satisfaction and future expectations | A person is deprived if experiences 3 or more deprivations related to personal satisfaction (life, economic situation, health, familiar and friends relationship, leisure and future expectations). |

[1]The provinces of Vibo-Valentia and Benevento have been excluded from the analysis due the very high percentage of missing data.

DRAFT

## 3. Methodology

Small area estimation (SAE) combines survey data with auxiliary variables of the population of interest to break down regional estimates into sub-regional ones. These variables are commonly obtained from population censuses or from administrative registers. If auxiliary information are available at unit-level (e.g., individual or household-level) one can consider unit-level SAE models[2]. When, on the contrary, the auxiliary data are available only at area-level (e.g., district, municipality or provincial level), one can use area-level SAE models.

We follow the latter approach and apply the Fay-Herriot area-level small area estimation model introduced by (6), which proposes the empirical best linear unbiased predictor (EBLUP) estimator to obtain estimates of the poverty dashboard indicators at provincial level[3].

Consider a finite population $U$, partitioned into $d = 1, \ldots, D$ mutually exclusive and exhaustive areas (in our case, provinces); the Fay-Herriot (FH) model is defined in two stages.

Let $\hat{\delta}_d^{DIR}$ be a direct estimator of $\delta_d$, the parameter of inferential interest for area (province) $d$. In the first stage, we assume that $\hat{\delta}_d^{DIR}$ is an unbiased estimator of $\delta_d$:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \qquad e_d \overset{ind}{\sim} N(0, \psi_d), \tag{1}$$

where $\psi_d$ is the sampling variance of the direct estimator $\hat{\delta}_d^{DIR}$ given $\delta_d$, assumed to be known for all $d = 1, \ldots, D$.

In the second stage, we assume that the area parameters $\delta_d$ are linearly related with a p-vector $x_d$ of area-level auxiliary variables as follows:

$$\delta_d = x_d^T \beta + u_d, \qquad u_d \overset{ind}{\sim} N(0, A). \tag{2}$$

Model (1) is known as *sampling model* because it represents the uncertainty due to the fact that $\delta_d$ is unobservable and the direct estimator is based on the sample data, $\hat{\delta}_d^{DIR}$. While, model (2) is called *linking model* because it relates all areas through the common regression coefficients $\beta$, allowing us to borrow strength from all areas.

Combining the two model components (1) and (2), we obtain the linear mixed model:

$$\hat{\delta}_d^{DIR} = x_d^T \beta + u_d + e_d, \tag{3}$$

where $u_d$ is independent of $e_d$.

The EBLUP estimator is the combination of the direct and the regression-synthetic estimators:

$$\hat{\delta}_d^{EBLUP} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d) x_d^T \hat{\beta} \tag{4}$$

where $\hat{\gamma} = \frac{\hat{A}}{\hat{A} + \hat{\psi}_d}$ represents the shrinkage factor.

For each of the 13 dashboard indicators of deprivation $j$ and for each province $d$ we obtain an EBLUP estimate, denoted with $\hat{\delta}_{dj}^{EBLUP}$. For sake of simplicity we will denote estimator $\hat{\delta}_{dj}^{EBLUP}$ with $\hat{\delta}_{dj}$.

Then, in order to obtain a composite indicator of multidimensional poverty, we aggregate the EBLUP provincial estimates $\hat{\delta}_{dj}$, where $d$ indicates the province and $j$ the dashboard indicator, using the Adjusted Mazziotta-Pareto Index (AMPI); see (8).

The AMPI approach considers the normalized provincial estimates $r_{dj}$ defined as:

$$r_{dj} = \frac{\hat{\delta}_{dj} - Min(\hat{\delta}_{dj})}{Max(\hat{\delta}_{dj}) - Min(\hat{\delta}_{dj})} 60 + 70. \tag{5}$$

Denoting with $Ref_{\delta_j}$ the reference value for the indicator $j$, the 'goalposts' $Max(\hat{\delta}_{dj})$ and $Min(\hat{\delta}_{dj})$ are defined as:

$$\begin{cases} Min(\hat{\delta}_{dj}) = Ref_{\delta_j} - \Delta \\ Max(\hat{\delta}_{dj}) = Ref_{\delta_j} + \Delta \end{cases} \tag{6}$$

[2]The Battese-Harter-Fuller (BHF) model is one of the most used unit-level small area estimation model (2).

[3]See (10) for notation details.

DRAFT

where $\Delta = \frac{Sup_{\delta_j} - Inf_{\delta_j}}{2}$. $Inf_{\delta_j}$ and $Sup_{\delta_j}$ are the overall minimum and maximum of the indicator $\hat{\delta}_{dj}$ across all provinces and the two years considered. To facilitate the interpretation of results, we choose as $Ref_{\delta_j}$ the mean of the year 2019. The normalized values $r_{dj}$ will fall approximately in the range $(70; 130)$, where 100 represents the reference value.

Denoting with $M_d$ and $S_d$ , respectively, the mean and standard deviation of the normalized indicators $r_{dj}$ of province $d$, the multidimensional poverty index of province $d$ is computed as:

$$AMPI_d = M_d + S_d \times cv_d \tag{7}$$

where $cv_d = \frac{S_d}{M_d}$ is the coefficient of variation for the local unit $d$; see (8).

## 4. Empirical findings

Figures 1 and 2 illustrate a geographical representations of how the multidimensional poverty indicator as well as the domain-specific deprivation indices are distributed across the Italian provinces in the years 2019 and 2020, respectively. In each panel, value 100 corresponds to the Italian average in 2019, and darker colors indicate higher levels of deprivation.

Comparing Figures 1(d) and 2(d) it emerges how health deprivation strongly worsened during the Covid-19 pandemic especially in Sicily, Sardinia, Lombardy and Piedmont and for the provinces located along the Tyrrhenian coast. On the contrary, Figures 1(e) and 2(e) reveal that life satisfaction and future expectations have not been affected by Covid-19.

Overall, the multidimensional poverty index (Figures 1(a) vs. 2(a)) does not seem to have deteriorated after the first year of the pandemic, although its levels remain higher in the South of Italy. We note that some domains of poverty (economic, education, health) have worsened in many provinces, but their effects have been compensated by the opposite trend registered for subjective well-being and quality of neighborhood. Anyway, the result refers only to the first year of the pandemic. In order to have a complete overview of the effect of Covid-19, one should analyze more recent data.

DRAFT

## 5. Conclusion

In this paper we have estimated the multidimensional poverty of the Italian households during the period of Covid-19 at provincial level. Since the AVQ survey is planned to obtain precise estimates at regional level, we have applied the FH small area estimation model in order to obtain reliable estimates of the dashboard indicators of deprivation at this finer geographical level. The overall composite indicator of multidimensional poverty has been then obtained using the AMPI approach.

Empirical findings suggest that, among the different domains of poverty considered, health deprivation has strongly worsened during the first year of the pandemic especially for provinces of the Islands (Sicily and Sardinia), Lombardy, Piedmont and the whole region located along the Tyrrhenian coast. Subjective deprivation, on the contrary, has reduced in the period of analysis. Overall, data reveals that the multidimensional poverty has not worsened relevantly, although the provinces of southern Italy are multidimensionally poorer than those of the north.

Further research could be devoted to the use of different techniques of aggregation of the dashboard indicators.

## Acknowledgments

This research has been founded by Fondazione Cariplo, under the project *POST-COVID: POverty and vulnerability Scenarios in The era of Covid-19: how the pandemic is affecting the well-being of the Italians* - rif. 2020-4216.

The data used in this paper are from Istat's Aspects of Daily Life ('Aspetti di vita quotidiana') survey. The elaborations were carried out at Istat's ADELE ('Analisi dei Dati ELEmentari') laboratory

and in compliance with the legislation on protection of statistical confidentiality and personal data. The responsibility for results and opinions expressed in the paper rests solely with the authors. The analyses have been carried out without sampling weights.

## References


[1] Atkinson, A. B.: Multidimensional deprivation: contrasting social welfare and counting approaches. The Journal of Economic Inequality **1 (1)**, 5–65 (2003)

[2] Battese, G. E., Harter, R. M., Fuller, W. A.: An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association **83 (401)**, 28–36 (1988)

[3] Betti, G., Cheli, B., Lemmi, A., Verma, V.: The fuzzy set approach to multidimensional poverty: the case of Italy in the 1990s. In: Kakwani, N., Silber, J. (eds). Quantitative Approaches to Multidimensional Poverty Measurement. Palgrave Macmillan, London (2008).

[4] Coromaldi, M., Zoli, M.: Deriving multidimensional poverty indicators: Methodological issues and an empirical analysis for Italy. Social Indicators Research **107**, 37–54 (2012)

[5] De Rosa, D.: Are Italians getting multidimensionally poorer? Evidence on the lack of equitable and sustainable well-being. Italian Economic Journal **8 (1)**, 145–174 (2022)

[6] Fay, R. E., Herriot, R. A.: Estimation of income from small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association **74 (366)**, 269–277 (1979)

[7] Mazziotta, M., Pareto, A.: On a generalized non-compensatory composite index for measuring socio-economic phenomena. Social Indicators Research **127 (3)**, 983–1003 (2016)

[8] Mazziotta, M., Pareto, A.: Measuring well-being over time: the adjusted Mazziotta-Pareto index versus other non-compensatory indices. Social Indicators Research **136 (3)**, 967–976 (2018)

[9] Molina, I., Rao, J. N. K.: Small area estimation of poverty indicators. The Canadian Journal of Statistics **38 (3)**, 369–385 (2010)

[10] Molina, I., Marhuenda, Y.: Sae: an R package for small area estimation. The R Journal **7 (1)**, 81 (2015)

[11] Pratesi, M., Quattrociocchi, L., Bertarelli, G., Gemignani, A., Giusti, C.: Spatial distribution of multidimensional educational poverty in Italy using small area estimation. Social Indicators Research **156**, 563–586 (2021)

[12] Rao, J. N. K.: Some new developments in small area estimation. JIRSS (2003)


DRAFT

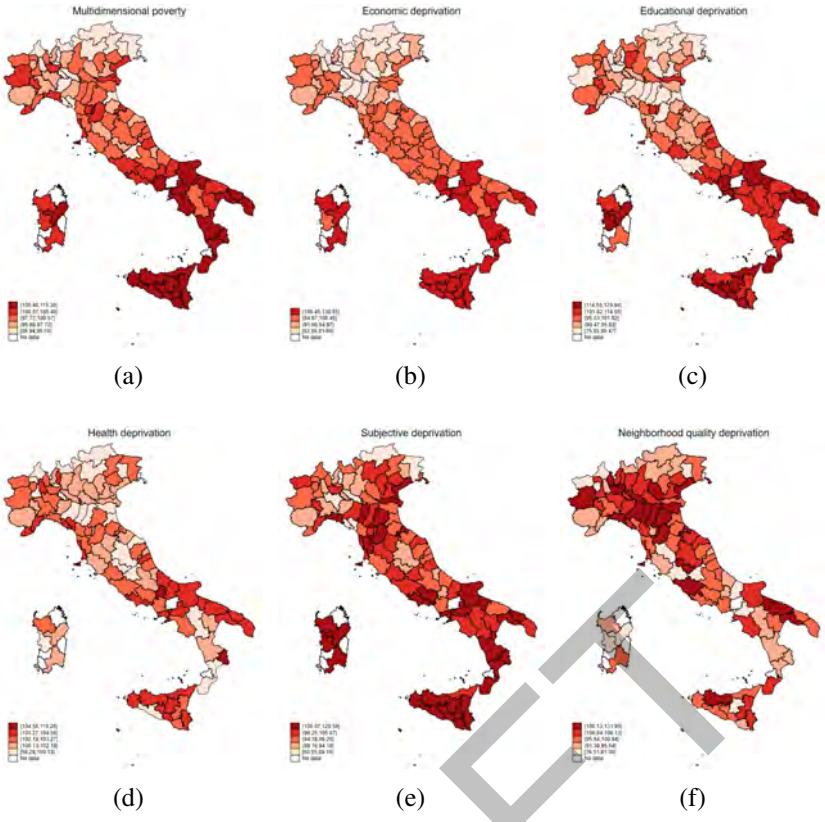



Figure 1: Overall deprivation and by specific dimension (year 2019).

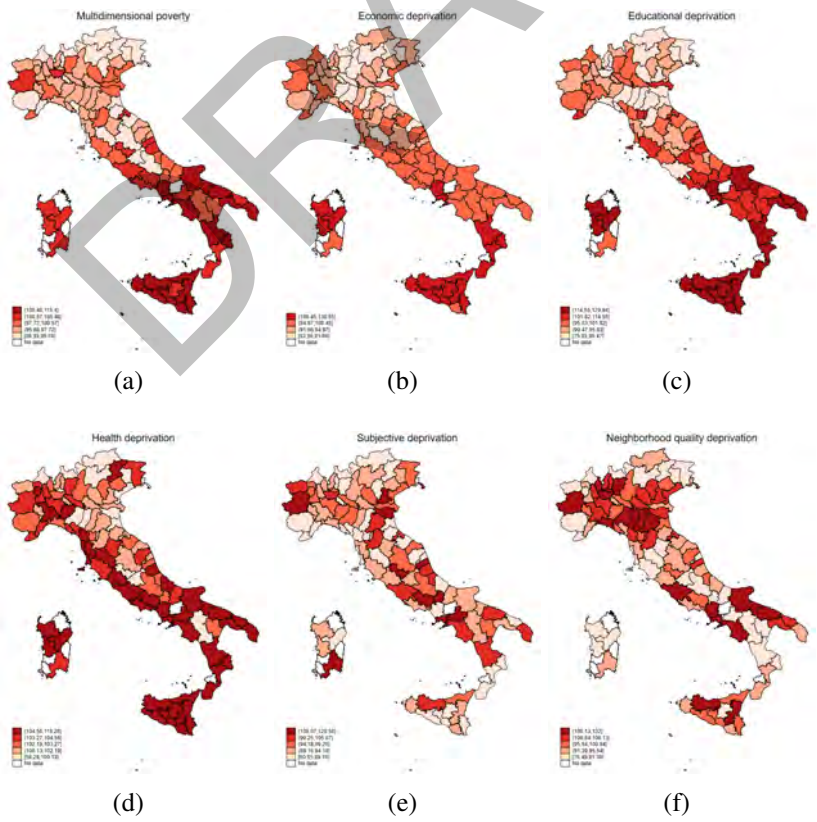



Figure 2: Overall deprivation and by specific dimension (year 2020).

DRAFT

# The fuzzy set approach as statistical learning for the analysis of multidimensional well-being

Gianni Betti[a], Federico Crescenzi[b], Antonella D'Agostino[c], and Laura Neri[a]

[a] University of Siena; Gianni.betti@unisi.it,laura.neri@unisi.it
[b] University of Tuscia: Federico.crescenzi@unitus.it
[c] Parthenope University of Naples; antonella.dagostino@uniparthenope.it

**Abstract**

Monitoring living conditions at local level is becoming an important goal for policy maker - especially after the pandemic – in order to identify and measure the incidence of new vulnerable people and worsening of existing vulnerabilities. However, its measurement is complex and multi-faceted, and, as current and future living conditions depend upon a great number of variables. Moreover, there is no doubt that pandemic impacted differentially in different areas and sub-groups of population. Against this background, the aim of this article is to provide a better understanding of the changes in living conditions in Tuscany at the local level. In order to achieve it, the paper used a multidimensional and fuzzy approach to measure living conditions in Tuscany in a dynamic perspective. The study is based on two *ad hoc* surveys conducted in 2021 and 2022.

**Keywords:** living conditions, fuzzy approach, pandemic, local level, Tuscany

DRAFT

## 1. Introduction

The intent of this study is to measure how and how much the consequences of the COVID-19 pandemic affected socio-economic living conditions of the Tuscan household population going below the regional level and using a multidimensional perspective. This exercise will inform us whether households, at the end of 2022, experiment a phase of improvement of their living conditions with respect their situation during the pandemic in 2021, or the new economic shocks due to the war, the rising of energy prices and inflation worsening existing socio-economic vulnerabilities. The final aim is to obtain new information for designing preventive and protective measures at local level. From a methodological point of view, we start from the consideration that the traditional poverty measures lack to properly estimate the vulnerability of households towards poverty. The reason behind this inability is mainly the failure of the existing measures to recognize the graduality inside the concept of poverty (Chiappero-Martinetti,1994; Cheli and Lemmi,1995). The Integrated Fuzzy and Relative (IFR) approach was developed in this perspective because the graduality within a vague concept can well be represented by the idea of fuzzy logic (Zadeh, 1965). In other words, IFR approach is based on the assumption that poverty is a multidimensional phenomenon and a vague predicate that manifests itself in different shades and degrees (fuzzy concept) rather than an attribute that is simply present or absent for individuals in the population (Betti et al., 2006). The IFR approach in a "modern statistical language" can be also defined an unsupervised learning method since it addresses a data reduction issue. IFR method does not presume any structure of the data. In fact, several elementary indicators represent the potential set of indicators that are single observable manifestations of the multidimensional concept of poverty and explorative factor analysis is performed for discovering if the multidimensional concept should be broken down into more than one

dimension. The empirical analysis is based on two *ad hoc* and cross-sectional sample surveys conducted in 2021 and 2022 by the Regional Institute for Economic Planning of Tuscany in collaboration with the University of Siena.

## 2. Data

Data on living conditions of resident households in Tuscany between pandemic and post-pandemic period have been drawn from two cross-sectional surveys. The first survey was conducted in September 2021 approximatively after eighteen months from the pandemic beginning, while the second in October 2022, when the pandemic issues was reduced, but households were in difficulty with the rising energy prices and inflation. Both surveys focus on the economic and social features of households, with particular attention to the current economic situation and prospects. The survey design of the two surveys is very similar: households have been drawn to achieve representativity at NUTS-3 level (the sample size is reported in Table 1); most of the interviews was conducted by C.A.T.I. and C.A.M.I. methods, interviewing one adult household member; the two questionnaires share most of the questions. As regards to some specific features of the 2022 survey it is worth to highlight that some interviews (29%) were conducted by C.A.W.I., and that the survey 2022 includes a set of questions as regards to the family's strategies to face the high energy price. A weighting adjustment procedure has been performed so that the sample totals conform to the population totals of specific domains obtained as an aggregation of functional geographies, clustering according to economic characterization of the so-called Local Labour Market Areas. Such grouping, in six different areas, refers also to the levels of employment and of the remuneration of productive factors (labour and capital) and consequently to different level of wellbeing. As regards item nonresponses, missing data have been imputed by deductive imputations based on logical or mathematical relationships between the variables, where it was possible. Item nonresponses relative to some quantitative and qualitative variables were imputed with stochastic imputation methods, assuming fully conditional specification[1]. The largest number of missing values (respectively 14.5% for 2021 and 10.7% for 2022) were registered for the only question adopted to collect the approximative monthly total net household income. The approximative values collected lead to a bracket distribution, and continuous values within each bracket have been imputed considering the kernel density estimate of the empirical distribution. Based on the total household disposable income, we retrieved the equivalized income using the OECD-modified equivalence scale. The poverty line was taken as the 60% of the median of the equivalised income distribution respectively for 2021 and 2022.

DRAFT

Table I: Sample sizes by Province (surveys 2021 and 2022).

| | Prato | Massa | Livorno | Grosseto | Pistoia | Arezzo | Lucca | Siena | Pisa | Firenze | Tuscany |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2021 | 83 | 94 | 164 | 166 | 175 | 207 | 263 | 320 | 336 | 691 | 2499 |
| 2022 | 142 | 248 | 265 | 306 | 200 | 182 | 437 | 268 | 328 | 641 | 3017 |

Several non-monetary indicators (binary variables) were collected in both surveys. These indicators are based on the standard questions as regard to affordability, such as affordability to eat nutritional meals, to keep household adequately warm and the capacity to cover costs for health, for one week holiday, for cinema or theatre, for eating out once a month, to cover costs for transport, for children clothes, toys, or specific children food; to cover costs for education such as taxes, books and materials and finally and then ability to cope with unexpected expenses of different amount. When necessary, these indicators have been transformed according to a positive polarity: the indicator is one in a situation of vulnerability to poverty or zero otherwise.

[1] FCS method of the MI procedure of the SAS software.

## 3. Methodology

The Integrated Fuzzy and Relative (IFR) assumes that poverty is a multidimensional phenomenon and a fuzzy concept rather than an attribute that is simply present or absent for individuals in the population, as the traditional poverty approach assume. In particular, it belongs to the fuzzy methods whose membership function is based on a distribution function (Betti et al., 2023). For a detailed discussion of IFR methodology reader is referred to Betti et al. (2006). Very briefly, IFR approach includes a non-fixed value of poverty risk and deprivation, through the introduction of a membership function, i.e. a quantitative specification of individual/household degrees of poverty and deprivation depending on the other individuals or households included in the analysis. A membership function's value of 0 is always associated with the lowest risk of poverty and deprivation, whereas a value of 1 is associated with the highest risk. Furthermore, the multidimensional framework of the IFR approach works up on several non-monetary indicators, assumed to be the manifest representation of a restricted number of underlying dimensions of deprivation, besides a monetary indicator based on the equivalent disposable income. In this paper, let FM be the membership function defined for the monetary variable (i.e. household income) and $FS_j$ be the membership function referring to dimension j ($j = 1..D$).

## 4. Results

### 4.1 Traditional approach

The so-called traditional approach has been performed by calculating the Eurostat "At-risk-of-poverty" rate (ARPR), based on the 60% of the median household equivalised income. Although the both median and mean income have been substantially stable from 2021 and 2022 (see Table II), the ARPR increased from 11.58% in 2021 to 15.79% in 2022. The deeper analysis conducted taking into account other EU Laeken indicators, such as 40%, 50% and 70% of the equivalised income, have confirmed this change in the reported income values (only) in the left-side if the income distribution itself: in fact, all such additional traditional indicators have increased from 2021 to 2022.

Table II: Traditional approach: the Laeken indicators (2021 and 2022)

| | Mean | Median | 40% median | 50% median | **60% median** | 70% median |
|---|---|---|---|---|---|---|
| 2021 | 1295 | 1196 | 2.41 % | 5.72 % | **11.58 %** | 20.20 % |
| 2022 | 1377 | 1207 | 4.93 % | 9.29 % | **15.79 %** | 24.13 % |

Such traditional approaches to the measurement of poverty have received some critics in the literature of the last three decades; i) first of all, they depend on a poverty line, and thus dichotomise the population into the poor and the non-poor (Cheli and Lemmi, 1995); ii) they are mainly unidimensional – i.e. they refer to only one proxy of poverty, namely, low income or consumption expenditure (Bourguignon and Chakravarty,1999; Anand and Sen, 1997); iii) in the research on poverty dynamics, spells usually replaced households or individuals as units of analysis, which led to a concentration on the duration of poverty and loss of sight of its severity (Ashworth et al., 1993); iv) moreover, small changes in the monetary variable may let individuals cross the poverty line, thus overestimating transitory poverty (Cheli and Betti, 1999), which may have been the case of the situation in Tuscany from 2021 and 2022. As introduced in the previous sections, the fuzzy and multidimensional approach to poverty and well-being measurement could be a statistical learning method able to overcome some of such critics.

### 4.2 Multidimensional and fuzzy approach

We used data collected in 2021 for identifying significant dimensions of our data. Accordingly, we used Exploratory Factor Analysis (EFA) for determining whether the 12 items listed in Table III can be arranged in a restricted number of dimensions. Table III reports the final structure provided by data. We identified three dimensions reflecting i) Basic needs & social life, ii) More specific Utilities and iii) Financial issues whose membership functions are FS1, FS2 and FS3, respectively. The last dimension represents the membership function based on the monetary variable (FM). These dimensions have been assumed also for the 2022 survey, in order to perform a longitudinal analysis.

DRAFT

At regional level we observed an unnoticeable difference between 2021 and 2022 in the FM indicator. FM indicator increases indeed from 0.116 to 0.117, although the so-called traditional approach has shown a worsen situation based on the poverty lines and the solely left-side part of the equivalised income distribution. Whereas results suggest an improvement of living conditions in two dimensions (Basic needs & Social life and Specific Utilities) because FS1 and FS2 indicators decrease from 0. 241 to 0.176 and from 0.1394 to 0.0650, respectively. Financial issue (FS3) dimension deserves particular attention because its value is the only one increased from 0.115 to 0.142 in the two years period. These general findings stress that the new current socio-economic shocks seem to contribute to increase final insecurity more than other aspects, or at least its perception. The perception of financial insecurity could also have caused a sort of underreporting in the left-side of the income distribution (without affecting its median or mean); this perception is clearly an important aspect that can contribute to other detrimental outcomes such us undermined basic psychological needs and lowered well-being (measured in terms of self-esteem, depression, and anxiety).

Table III: Dimension and items

| Membership function label | Dimension descripti | Items |
|---|---|---|
| FS1 | Basic needs & Social life | Meals with meat or fish // Household adequately warm // cover costs for health// cover costs for 1 week holiday// cover costs for cinema, theatre, eating out once a month; |
| FS2 | Specific utilities | Costs for: children (clothes, toys, child's food)// education (taxes, books and materials) // transports |
| FS3 | Financial issues | Inability to cope with unexpected expenses: 5,000, 2,000, 800 Euros; |
| FS4 | Monetary dimension | Equivalised monthly income |

DRAFT

It was also fairly clear that not all Provinces showed the same intensity of the multi-dimensional indicators and the same variation from 2021 to 2022 (see Figure 1)[2]. For instance, in 2021, the FM indicator present values larger than the regional level (0.116) in Lucca and Grosseto while Siena, Firenze and Arezzo present figures below the regional estimate. The same pattern is confirmed in 2022. Looking at difference over the period, results show light improvements for Arezzo, Pisa and Siena, while the monetary situation seems to get worse for Firenze, Livorno, Pistoia and Prato. Other interesting findings can be stressed by observing the other dimensions and comparing different provinces characterized by different level of urbanisation. For instance, Florence and Grosseto share a similar value of FS3 in 2021, whereas in 2022 household living in Florence, province characterized by high urban functions and widespread tertiarization, suffer less of financial insecurity that people living in Grosseto that is instead a rural district.

## 5. Conclusion

The consequent picture therefore is that even if the families have below of the poverty threshold, their economic behaviour could be curbed, by the financial fragility, and also by fear for the near future.
Finally, although the presented findings highlight some interesting information for policy maker by comparing living conditions and subnational level in Tuscany between 2021 and 2022, further analyses at finer territorial levels or peculiar groups of the population are necessary for designing preventive and protective measures at local level. These analyses will be the natural development of the preliminary results discussed in this paper.

[2] The estimates for Massa Carrara and Prato should be used with caution, according to Statistics Canada recommendation (Neri & Crescenzi, 2022).

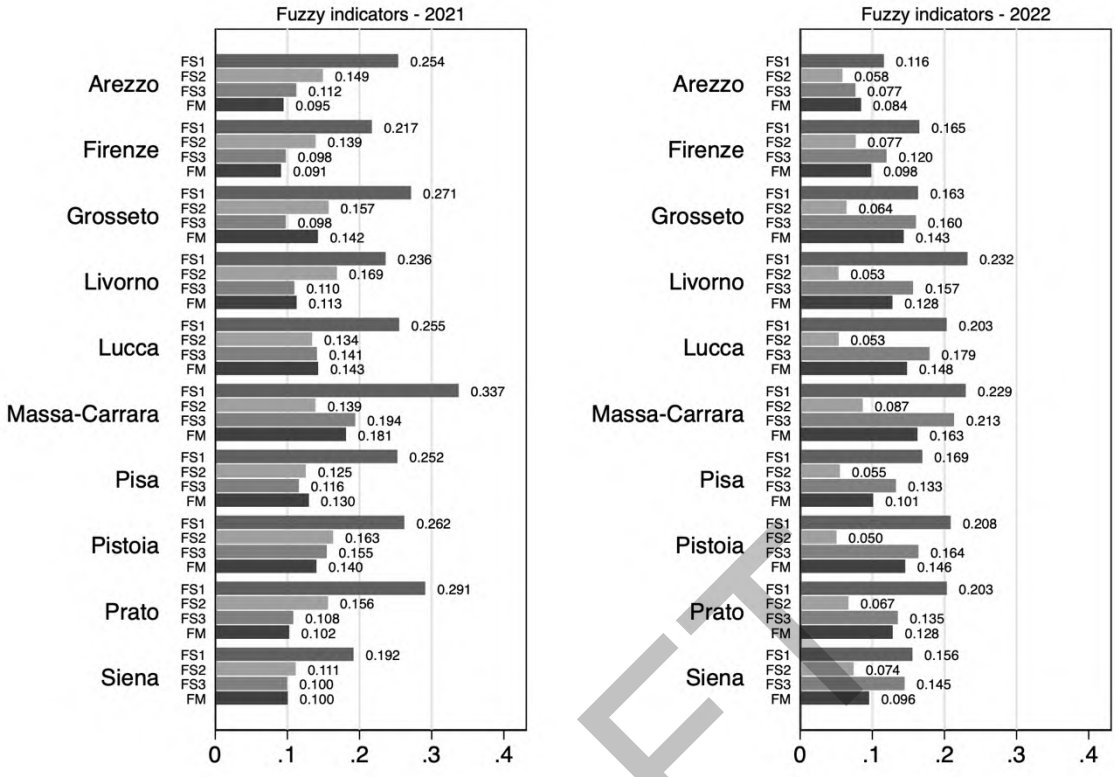



Figure 1: Fuzzy supplementary and Fuzzy monetary indicators at NUTS-3 level

DRAFT

## References


[1] Anand, S., Sen, A.K. (1997), *Concepts of human development and poverty: a multidimensional perspective*, in UNDP, Human Development Papers, 1997: Poverty and Human Development, United Nations Development Programme, New York.

[2] Ashworth, K., Walker R., Hill, M. (1993), A new approach to poverty dynamics, *Bulletin of Sociological Methodology*, 38(1), 14–37.

[3] Betti, G., Cheli B., Lemmi, A., Verma, V. (2006), Multidimensional and longitudinal poverty: an integrated fuzzy approach, in A. Lemmi and G. Betti (eds.), Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 111–137, Springer, New York.

[4] Betti, G., D'Agostino, A., Lemmi, A., Neri, L. (2023), The Fuzzy Approach to Poverty Measurement" in J. Silber (ed.) *Research handbook on measuring poverty and deprivation*, Edward Elgar Publishing Limited.

[5] Bourguignon, F., Chakravarty, S.R. (1999), A family of multidimensional poverty measures, in D.J. Slottje (ed.), *Advances in Econometrics, Income Distribution and Scientific Methodology*, Physica-Verlag, Heidelberg, Germany.

[6] Cheli, B., Betti, G. (1999), Fuzzy analysis of poverty dynamics on an Italian pseudo panel,1985–1994, *Metron*, 57(1–2), 85–105.

[7] Cheli, B., Lemmi, A. (1995), A totally fuzzy and relative approach to the multidimensional analysis of poverty, *Economic Notes*, 24, 115–134.

[8] Chiappero-Martinetti, E. (1994), A new approach to evaluation of well-being and poverty by fuzzy set theory, Giornale Degli Economisti e Annali di Economia, 53, 367–388.

[9] Neri, L., Crescenzi, F. (2022), Conference paper, presented at the International Conference on Regional Science, Granada, 2022.

[10] Zadeh, L.A. (1965), Fuzzy sets, *Information and Control*, 8, 338–353.

# A Bayesian framework for early cancer screening

Sally Paganin[a] and Jeff Miller[a]

[a]Harvard T.H. Chan School of Public Health, 655 Huntington Ave, Boston MA 02115;
spaganin@hsph.harvard.edu

**Abstract**

There is growing interest in developing tools for cancer screening and monitoring based on the analysis of DNA sequencing data derived from non-invasive procedures such as blood samples. At early cancer stages, such samples contain DNA from a majority of normal cells and a low fraction of tumor cells. Cancer presence can be assessed by measuring allelic imbalance: since a person inherits one allele from each parent, the allele proportion at heterozygous loci is close to 0.5 in normal cells, whereas significant deviations from 0.5 are indicative of the presence of cancer. To efficiently and sensitively detect such deviations, we model the allele proportions over the genome via a Bayesian hierarchical Hidden Markov Model.

***Keywords:*** Allelic imbalance, Bayesian modeling, biological priors, NIMBLE.

DRAFT

## 1. Introduction

Cancer screening and monitoring are crucial for early detection and better treatment outcomes. With the advancement of next-generation sequencing (NGS) technologies, there has been a growing interest in developing minimally invasive procedures. One of the most promising technologies is the use of *cell-free DNA* (cfDNA), which refers to DNA fragments that can be found in the bloodstream. These fragments are released by cells during their normal life cycle. The cfDNA can be extracted from a bodily fluid sample, such as blood, making it a minimally invasive and accessible method for cancer screening.

Cancer is a genetic disease defined by alterations of the DNA that occurs in some cells. It has been shown that fragments of this altered DNA can also be found in the bloodstream and thus potentially used to detect cancer. Such tumor-derived DNA is commonly referred to as *circulating tumor DNA* (ctDNA). The main challenge in using ctDNA as a cancer biomarker is whether ctDNA can be distinguished from cfDNA. Moreover, in most cancer patients, ctDNA constitutes a very small proportion of the cfDNA, especially when cancer is at its early stages. Hence, advanced statistical methods are needed to extract the signal from the noise at such a small resolution.

We propose a hierarchical Bayesian framework for genomic signals of cancer to efficiently enable cancer detection at early stages. We focus on measuring allelic imbalance: since a person inherits one allele from each parent, the allele proportion at heterozygous loci is typically distributed around 0.5 when DNA fragments come from healthy cells. Small significant deviations from 0.5 can be indicative of DNA alterations, and hence the presence of DNA from cancerous cells.

## 2. Allelic imbalance in genomic data

Evidence for cancer presence in genomic data is typically inferred by modeling different data summaries derived from sequencing data. This kind of data is derived from a multi-stage process where

DNA strands are unraveled, cut into pieces and amplified, so that, at the end of the process, we have multiple observations at each genomic position (locus). These pieces of DNA are referred to as *reads*, and to be analyzed, researchers need to know where they are located on the genome. Reads are typically aligned using a reference genome, which provides a standardized map of the entire genetic makeup of an organism. At the end of the process, genomic data consists of multiple observations at each locus of the nucleotide bases (A, T, C, G) that makes up the DNA code.

The term *allele* indicates one of two or more versions of a DNA base. Since a person inherits half of their genome from each parent, at each locus we can either observe the same allele (homozygous) or two different ones (heterozygous). For healthy human cells, allele frequencies at heterozygous loci are expected to be distributed around $0.5$. If one allele has a higher frequency than the other—i.e., the two alleles are expressed differently—one can individuate a major and a minor allele. *Allelic imbalance* describes the situation where the allele frequency deviates from the expected $0.5$ value. This can result from deletion or duplication of part of a chromosome and is often related to cancer development.

Information about which is the minor or major allele is not directly available in genomic data, as alleles are distinguished using the reference genome used in the alignment phase. The term *reference allele* indicates the base that is found in the reference genome at a certain position, while the *alternative allele* refers to any base, other than the reference observed at the locus. This is also known as *B-allele* and we will use this term in the manuscript. Since B-alleles are defined with respect to the reference genome, they can either represent the major or minor allele.

Information related to parental origin can be used to aid the correct identification of minor alleles. It is known from biology, that stretches of the genome containing variants tend to be inherited together from one of the parents. These stretches are called *haplotypes*, and we can obtain an estimate of the haplotype for each locus by using a *phaser algorithm* (e.g., SHAPEIT2 (4)). This is based on a Hidden Markov model that uses use information about genotype and genomic locations, which can be considered as "external data" for the allele counts.

DRAFT

## 3. A Hidden Markov Model for allele proportions

We model observations within each chromosome independently and illustrate the model for one generic chromosome. Let $y_i$ and $n_i$ denote the number of B-alleles and the total number of reads at each locus $i$, for $i = 1, \dots, m$. We want to estimate the minor allele proportion, which can be described as a piecewise function over the chromosome. Sections of the chromosome where it is locally constant are referred to as *segments*. Segments are typically unknown and estimated via a statistical model; here we assume that they are given and ignore the associated uncertainty. We introduce a variable $s_i \in \{1, \dots, J\}$ indicating which segment locus $i$ belongs to. Interest is in estimating $\theta_j \in (0, 0.5], j = 1, \dots, J$, which is the average minor allele proportion for each segment.

To do so, we model the number of B-alleles assuming a Binomial distribution at each genomic position

$$y_i \sim \text{Bin}(n_i, \theta_{s_i}), i = 1, \dots, m, s_i \in \{1, \dots, J\}. \tag{1}$$

As explained in Section 2, the observed number of B-alleles does not necessarily reflect the minor allele proportions. This means that the observed number of B-alleles $y_i$ at some loci corresponds the number of major alleles, and should be considered to estimate $1 - \theta_{s_i}$, rather than $\theta_{s_i}$. To reflect this, we introduce an indicator variable $h_i \in \{0, 1\}$ encoding whether or not the B-allele is the minor allele, so that

$$(\theta_{s_i} \mid h_i) = \theta_{s_i}^{h_i}(1 - \theta_{s_i})^{(1-h_i)} \quad i = 1, \dots, m. \tag{2}$$

Finally, we include in the model the haplotype information obtained from the phasing algorithm via a Hidden Markov Model. We do this by interpreting the haplotype as a prior estimate $\hat{h}_i$ for the latent variable $h_i$, for $i = 1, \dots, m$. Although $\hat{\boldsymbol{h}}$ cannot be expected to perfectly recover $\boldsymbol{h}$, it can be expected to either match or not match $\boldsymbol{h}$ for long stretches of the genome. To account for occasions of mismatch, we introduce a variable $w_i = I(h_i \neq \hat{h}_i)$ indicating whether or not the haplotype of the minor allele

coincides with the estimate, with $w_i \in \{0, 1\}$. In other words, $w_i$ tells us when we should switch the label given by $\hat{h}_i$. We model dependence between loci within the chromosome assuming a Markovian model

$$
\begin{aligned}
(\theta_{s_i} \mid h_i) &= \theta_{s_i}^{h_i}(1-\theta_{s_i})^{(1-h_i)} \\
(h_i \mid \hat{h}_i, w_i) &= \begin{cases} \hat{h}_i & \text{if } w_i = 0, \\ 1-\hat{h}_i & \text{if } w_i = 1, \end{cases} \\
\Pr(w_i = t \mid w_{i-1} = v) &= \pi_{tv} \quad t, v \in \{0, 1\}, \quad i = 1, \ldots, m.
\end{aligned}
$$

The model above can be interpreted as a Hidden Markov Model, where $\boldsymbol{w}$ is the vector of latent states with 2 possible states, $\mathbf{T}_w = \{\pi_{tv}\}_{t,v\in\{0,1\}}$ the transition matrix, and $\boldsymbol{y}$ the vector of emissions.

**Prior choice.** We use a spike-and-slab prior to distingush between regions of standard (healthy) behavior and regions of allelic imbalance. Conceptually, as spike distribution $\mathrm{p}_{0.5}(\cdot)$ we look for a distribution concentrated around $0.5$, representing the behavior of allele proportions as in a healthy sample, while we denote as $\mathrm{p}_{AI}(\cdot)$ the slab distribution over $(0, 0.5]$ for situations of allelic imbalance.

$$
\begin{aligned}
(\theta_{s_i} | s_i = j) = \theta_j, \quad & i = 1, \ldots, m, \\
\theta_j \sim \nu_j \mathrm{p}_{0.5}(\theta_j) + (1-\nu_j)&\mathrm{p}_{AI}(\theta_j) \\
\nu_j \sim \text{Bern}(0.5), \quad & j = 1, \ldots, J.
\end{aligned} \tag{3}
$$

We use $\mathrm{p}_{0.5}(\cdot) = Unif(0.47, 0.5)$ as spike distribution, while for the slab distribution, we use $\mathrm{p}_{AI}(\cdot) = \text{Beta}_{(0,0.5)}(2, 2)$, where $\text{Beta}_{(a,b)}$ indicate a truncated Beta distribution in the $(a, b)$ interval. To elicit the spike distribution, one can also calibrate the model on control samples containing DNA from healthy cells.

We assume a symmetric transition matrix $\mathbf{T}_w$, with $\pi_{00} = \pi_{11} = 1 - p$ and $\pi_{01} = \pi_{10} = p$. Since we expect $\hat{\boldsymbol{h}}$ and $\boldsymbol{h}$ to match for long stretches of the genome, probabilities $\pi_{01} = \pi_{10}$ can be interpreted as the probability of an error in the haplotype estimate. Hence we set $p = 1/1000$ which corresponds to an estimated error rate for the haplotype phaser (1).

**Posterior computation.** We estimate the model via MCMC, making use of the nimble software (2; 3), which provides a flexible `R` framework for hierarchical models. Nimble comes with a suite of built-in customizable samplers (e.g., conjugate, Metropolis-Hastings, automated slice samplers, etc.) requiring only the coding of new sampling algorithms. To estimate the model in (1)-(3) we use a combination of conjugate Metropolis-Hastings, altogether with a custom algorithm to sample the latent states for the Hidden Markov Model component. A naive implementation that samples each latent state via Gibbs sampling tends to be inefficient and has been often replaced by alternative sampling algorithms in the literature. We consider here the approach in (5) which uses the forward-backward (FB) algorithm to sample the latent states sequentially from their joint posterior distribution at each MCMC iteration.

## 4. Application to laboratory experiment data

In this application, we use data derived from samples generated in a laboratory experiment using cell lines, which are cultures of cells that can be propagated multiple times while maintaining the same characteristics. The DNA extracted from two cancer cell lines (HCC38, HCC1143) was mixed with DNA from matched B-lymphocyte cell lines (HCC38BL, HCC1143BL) in various proportions measured in volume, to mimic mixed blood samples from subjects developing cancer. Here we show some preliminary results from our model for chromosome 1 across 3 different samples, containing respectively about $3\%, 10\%, 20\%$ DNA from cancerous cells. In this type of application, a sample containing $20\%$ tumor DNA tends to provide strong evidence for allelic imbalance for the majority of chromosomes.

Chromosome 1 comprises 3827 heterozygous loci, divided into 48 segments of different sizes. Observed raw B-allele frequencies (BAF) are reported in the top row of Figure 1, while in the bottom row,

DRAFT

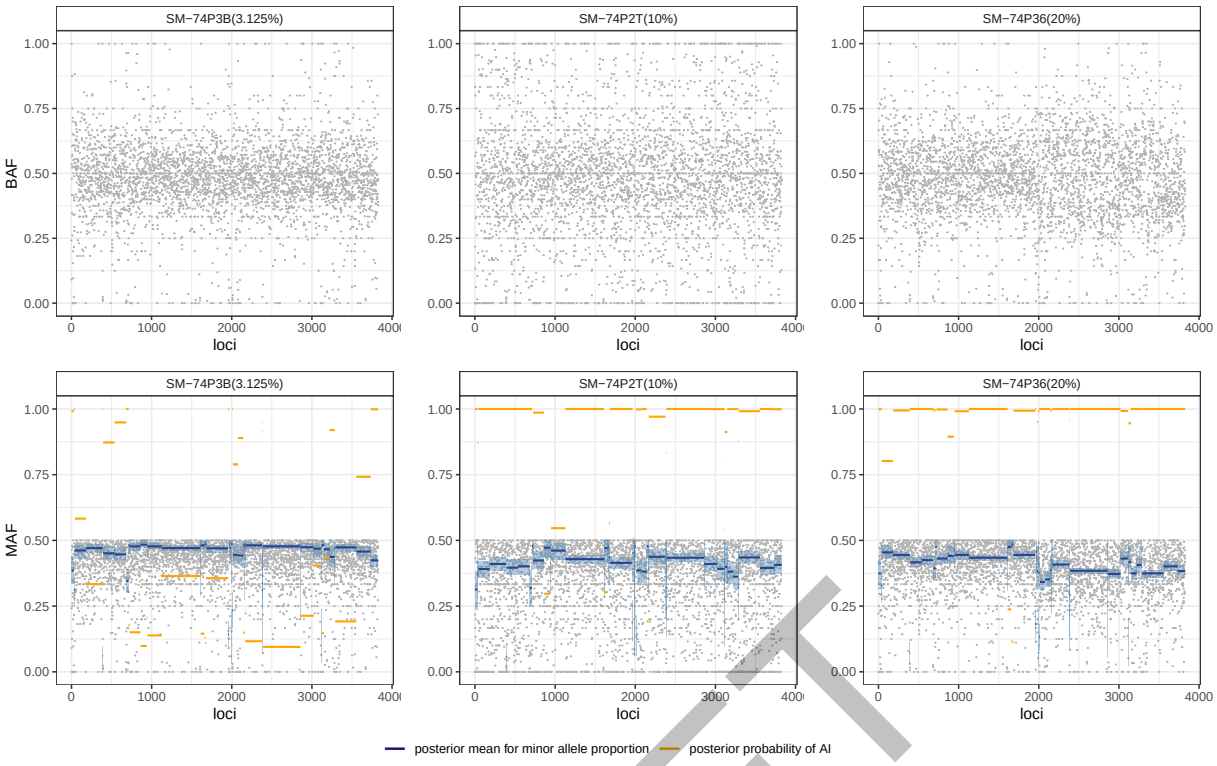



Figure 1: Top row. Observed B-allele frequency for 3 laboratory samples containing respectively $3\%, 10\%, 20\%$ DNA from cancerous cells. Bottom row. Results from our model. Gray dots are the MAF based on $\hat{\boldsymbol{w}}$. Blue lines represent posterior means for the minor allele proportions for each segment, while shaded blue areas represent high-density credible intervals at $95\%$. Yellow lines quantify the posterior probability of allelic imbalance for each segment.

DRAFT

we show the results from our model. For each sample, adjust the raw BAF to show the minor allele frequencies (MAF) based one the posterior estimate for the sequence of switches $\boldsymbol{w} = \{w_i\}$. We use the maximum a posterior, in this case corresponding to the sequence of $w_i$'s that occurs in the posterior samples for the higher number of the iterations. We also plot the posterior means for the minor allele proportions for each segment, along with the highest posterior density intervals at $95\%$. Finally, we show the posterior probability of allelic imbalance for each segment, i.e. the posterior probability that the segment mean is a draw from the slab distribution $p_{AI}(\cdot)$.

It can be seen that there is an increase in the number of segments with a high probability of allelic imbalance probability as the percentage of tumor DNA increases. For the sample containing $20\%$ tumor DNA, almost all segments are in allelic imbalance.

## 5. Discussion

In this work, we present a Bayesian model for allele proportions to detect allelic imbalance in samples containing signals from both healthy and cancerous cells, presenting preliminary results on laboratory experiments. Immediate extensions of the model will focus on including multiple samples from the same subject; although the segment's allele proportions can vary across samples, the haplotype information is shared and multiple samples can aid a better identification of the minor allele and, consequently, higher power in detecting allelic imbalance.

Finally, although the model is carefully constructed accounting for biological information related to the data, there are still factors that have not been considered in this application. For example, in genomic data there is a tendency to systematically observe more reads containing a reference allele

than an alternative one, due to different factors intrinsic to the sequencing process. The introduction of genomic covariates when modeling allele proportion can aid mitigation of such bias.

## References


[1] Choi, Y., Chan, A., Kirkness, E., Telenti, A. & Schork, N. Comparison of phasing strategies for whole human genomes. *PLOS Genet*. **14**(4), 1–26, (2018)

[2] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: Writing statistical algorithms for general model structures with nimble. *J. Comput. Graph. Stat.*, 26(2):403–413, (2017).

[3] P. de Valpine, C. Paciorek, D. Turek, N. Michaud, C. Anderson-Bergman, F. Obermeyer, C. Wehrhahn Cortes, A. Rodríguez, D. Temple Lang, S. Paganin, and J. Hug. Nimble: MCMC, particle filtering, and programmable hierarchical modeling, (2022).

[4] O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J., Rudan, I., McQuillan, R., Fraser, R., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A., Vitart, V., Navarro, P., Zagury, J., Wilson, J., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. & Marchini, J. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genet.* **10**, 1-21 (2014).

[5] Steven L Scott. Bayesian methods for Hidden Markov Models: recursive computing in the 21st century.J AM STAT ASSOC *J. Am. Stat. Assoc.*, 97(457):337–351, (2002).


DRAFT

# Linear models with assumptions-free residuals: a Bayesian Nonparametric approach.

Filippo Ascolani[a] and Valentina Ghidini[a]

[a]Bocconi University, Milan; filippo.ascolani@phd.unibocconi.it, valentina.ghidini@phd.unibocconi.it

**Abstract**

Assumptions on residuals limit the application of simple-to-interpret models to real world phenomena, especially when they are not mere measures of error but rather encode assumptions about the framework. For this reason, this work proposes a Bayesian nonparametric approach to relax some common assumptions on errors (such as homoskedasticity and symmetry), exploiting a nonparametric prior on the space of distributions of residuals. In particular, we design a methodology aimed at creating a more flexible model, yet retaining desirable parametric properties (such as interpretability). The final application involves a linear model, where the residual distribution is heavy tailed.

***Keywords:*** Bayesian nonparametrics, linear model, flexible residuals

DRAFT

## 1. Introduction and Motivation

Linear models are ubiquitous in statistics and machine learning. They start from the collection of $n$ observations $(y_i, X_i')$, where $y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^{p\times 1}$ and the assumption

$$y_i = X_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim f, \quad i = 1,\dots,n \tag{1}$$

with $\beta = (\beta_1,\dots,\beta_p)$ unknown parameters, with $p < n$, and $f \in \mathscr{F}$, where

$$\mathscr{F} = \left\{ g \mid \int x g(x)\,\mathrm{d}x = 0, \int x^2 g(x)\,\mathrm{d}x < \infty \right\}.$$

From a classical perspective, the advantage of the formulation in (1), with the general assumption of zero mean and finite variance of the residuals (called second order conditions) is that the estimate of $\beta$ that minimizes the square distance can be computed with the well known formula

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y \tag{2}$$

with $X = [X_1',\dots,X_n'] \in \mathbb{R}^{n\times p}$ and $Y = (y_1,\dots,y_n)'$. In other words, under model (1) and quadratic loss, (2) is the optimal estimate of $\beta$. However, in order to perform statistical inference (e.g. computing confidence intervals), a distributional assumption on $\varepsilon_i$ is required and the most common choice is given by $\varepsilon_i \sim N(0,\sigma^2)$. This setting is extensively studied and is endowed with many nice statistical and computational properties. Nevertheless, placing such a restriction poses many issues: not only the coverage of the confidence intervals and the power of the statistical tests will depend crucially on the distance between the data generating distribution and the Gaussian case, but also the symmetry of the residuals

is required, together with the assumption of light tails. This can be particularly relevant in the economic and econometric literature, where the $\varepsilon_i$ are not just mere errors, but are often a crucial component of the model. For instance, in the well known Capital Asset Pricing Model (3), assuming symmetric residuals implies symmetric returns, that can be an unrealistic assumption (10). Therefore, it makes sense to look for methodologies more capable of exploring the space of the residual distributions.

The goal of this work is to propose a novel methodology, based on a suitable countable mixture of normal distributions, that combines the theoretical generality of the Bayesian nonparametric (BNP) approach (see (5) for a review) with the appealing simplicity of linear models. On the one hand it will have better performances than the OLS estimate, both in terms of point estimate of $\beta$ and coverage, even with extreme residual distributions, on the other it will provide the researcher with simple ways of performing inference, similarly to the classical case. Finally, all the parameters involved will have an intuitive role, making prior elicitation less mysterious; in particular, it will allow us to easily study the possible asymmetry of the residuals.

The problem of modelling the residual through a BNP approach has been already discussed in the literature (e.g. (6; 9)): however the proposed methodologies either assume symmetry of the residuals, or can be difficult to interpret for a practitioner.

The rest of the paper is organised as follows. In Section 2 the methodology is introduced and thoroughly discussed. In Section 3 an algorithm to estimate the proposed model is presented and a simulation study is conducted. The paper ends with a discussion.

DRAFT

# 2. Methodology: theory

## 2.1 Formulation

In the Bayesian context, after setting the model (1), the statistician places a probability distribution, called *prior*, on the parameters (the coefficients $\beta$'s, for instance); then, instead of considering the Maximum Likelihood Estimates, inference is made through the *posterior distribution*, that is the distribution of the parameters given the observed data. In other words, we update the prior beliefs through the collected observations. The main advantage of this framework is that the researcher is endowed with an entire distribution for the parameter instead of a single numerical estimate: in this way, for instance, evaluation of the uncertainty of the analysis becomes immediate. See (5) for more details on Bayesian inference.

Considering more formally our problem, we start by the standard setting of Bayesian linear models:

$$\begin{aligned} y_i \,|\beta &= X_i'\beta + \varepsilon_i, \quad i = 1,\dots,n \\ \beta &\sim N_p(\cdot \mid b_0, B_0) \end{aligned} \tag{3}$$

where $N_p(\cdot \mid b_0, B_0)$ is the $p$-variate normal distribution, with mean vector $b_0 \in \mathbb{R}^p$ and covariance matrix $B_0$. The characteristic feature is given by the assumption on the residuals. Indeed, denoting $\mathrm{TN}(\cdot \mid \mu, \sigma_0^2)$ the normal distribution truncated on the interval $[0,\infty)$ and $\mathrm{IG}(\cdot \mid s, S)$ the inverse gamma distribution, we have

$$\begin{aligned} &\varepsilon_i \,|\mu_i, \tau_{1i}^2, \tau_{2i}^2 \overset{\text{ind}}{\sim} \frac{1}{2}N(\cdot \mid \mu_i, \tau_{1i}^2) + \frac{1}{2}N(\cdot \mid -\mu_i, \tau_{2i}^2) \\ &(\mu_i, \tau_{1i}^2, \tau_{2i}^2) \mid P \overset{\text{iid}}{\sim} P \\ &P \sim \mathrm{DP}(P_0, \theta), \quad P_0(\cdot) = \mathrm{TN}(\cdot \mid \mu_0, \sigma_0^2) \times \mathrm{IG}(\cdot \mid s_1, S_1) \times \mathrm{IG}(\cdot \mid s_2, S_2), \end{aligned} \tag{4}$$

where $\mathrm{DP}(P_0, \theta)$ is a *Dirichlet process* (DP) (introduced by (4)), with probability distribution $P_0$, called *baseline distribution*, and $\theta > 0$.

In the setting of (4), conditional to $P$, for the $i$-th residual a triplet $(\mu_i, \tau_{1i}^2, \tau_{2i}^2)$ is sampled from $P$ and the distribution of $\varepsilon_i$ is given by a mixture of two Gaussian distributions with opposite means and (possibly) different variances; thus the mean of $\varepsilon_i$ is zero, but the symmetry around the origin is not required. Since the Dirichlet process is almost surely discrete (see (4)), considering (4) and integrating

out $P$, the unconditional density of $\varepsilon_i$ is given by

$$\varepsilon_i \sim \sum_{j\geq 1} W_j \left[\frac{1}{2}N(\cdot \mid \mu_i, \tau_{1i}^2) + \frac{1}{2}N(\cdot \mid -\mu_i, \tau_{2i}^2)\right], \quad (\mu_i, \tau_{1i}^2, \tau_{2i}^2) \overset{\text{iid}}{\sim} P_0, \tag{5}$$

where $\{W_j\}_{j\geq 1}$ are random probability weights. Therefore, the distribution of the residual is assumed to be a countable mixture of a mixture of two Gaussians.

# 3. Methodology: computations

## 3.1 Algorithm

Considering model specification (3) with residuals (4), it is in general not possible to compute analytically the density $p(\beta \mid Y, X)$ with an explicit formula.

This issue holds for the vast majority of Bayesian models, especially in nonparametric contexts. However, many algorithms have been developed and studied to collect a sample $\{\beta_t\}_{t\geq 0}$ such that $\beta_t \sim p(\beta \mid Y, X)$, at least approximately. The idea is to create a Markov chain (hence the name Markov Chain Monte Carlo), whose invariant distribution is given by the posterior distribution of interest. In this paper we construct a chain $\{\beta_t, P_t\}_{t\geq 0}$ through the Gibbs sampler methodology (see (2) and (8) for more details). The procedure works as follows

1. Initialize $\beta_0$ with random starting values.
2. For $t = 1, \ldots, B$ do:
   (a) Sample $P = P_t$ from its conditional distribution given $(\beta_{t-1}, Y, X)$.
   (b) Sample $\beta_t$ from the conditional distribution given $(P_t, Y, X)$.

In order to sample $P_t$, that is a infinite dimensional object, we use a suitable truncation and employ the blocked Gibbs sampler proposed in (7). Thus, all the quantities of interest can be approximated through a simple Monte Carlo approach, i.e.

$$\mathbb{E}[h(\beta) \mid Y, X] \approx \frac{1}{B}\sum_{t=1}^{B} h(\beta_t).$$

In this way, for instance, the posterior mean and higher moments can be computed with $h(x) = x^k$. Similarly, credible intervals for the quantities of interest can be estimated by suitable empirical quantiles of $\{\beta_t\}_{t\geq 0}$.

## 3.2 Application - Heavy tailed residuals

To assess the performance of the proposed method, we consider a number of simulation examples varying the distribution of the residuals $\varepsilon$, and here we report the results regarding heavy tailed residuals $\varepsilon \sim T_q$, where $T_q$ denotes a Student's t distribution with $q$ degrees of freedom. The true coefficients are

$$\beta = (-1, -1, -1, 0, 1, 1, 1, 1, 1, 1)$$

The goodness of the model is measured with the mean width of empirical credible intervals, and the mean euclidean distance between the true coefficients and the estimated ones; then, we compute the median of those quantities over 100 randomly generated samples and compared with the ones obtained using the standard linear OLS regression.

In Figure 1 and 2, we can see the results for simulated residuals $\varepsilon \sim T_{1.1}$. Figure 1 displays the comparison between the point and interval estimates of the coefficients provided by the proposed method and by OLS: in particular, we can see that our method returns estimates (blue line) much closer to the true $\beta$ (black line), while OLS (red line) can not retrieve most of the estimates in a precise way.

DRAFT

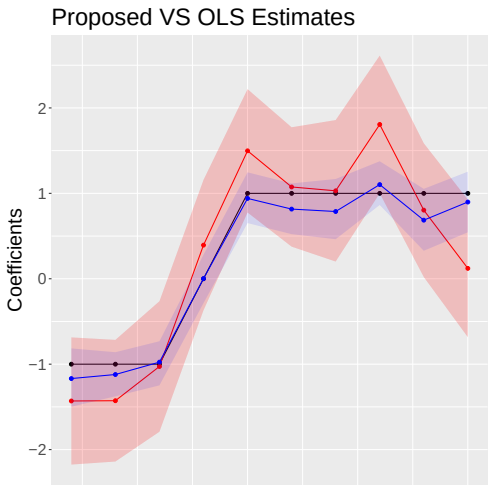



Figure 1: True coefficients (black), estimated with the proposed algorithm (blue), OLS Regression (red). True residual Density: $\varepsilon \sim T_{1.1}$.

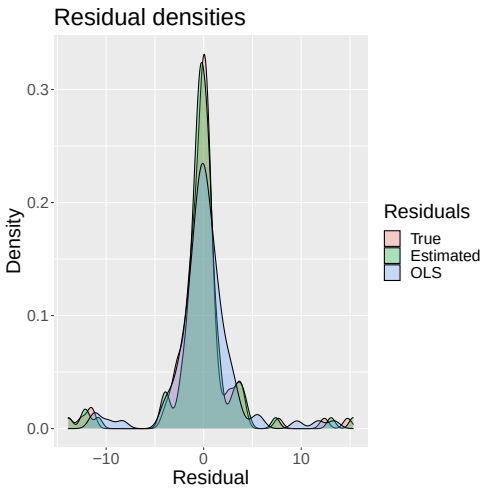



Figure 2: Residual Density: true (orange), posterior estimation with the proposed algorithm (green), estimation with OLS (blue). True residual Density: $\varepsilon \sim T_{1.1}$.

Moreover, the OLS confidence intervals are much wider than the posterior credible intervals computed with the proposed algorithm (confidence and credible level: 0.95). In Figure 2 we can see the difference in the estimated law of the residuals: the posterior density provided by the Gibbs sampler in Section 3.1 (green curve) is much closer to the true law from which the residuals are simulated from (orange curve). On the contrary, residuals from OLS can not retrieve the heavy tailed behaviour of the distribution in a proper manner.

Finally, in Table 1 we can see the mean interval width and the mean distance from the true coefficients obtained by 100 generated samples according to the same model. In particular, we can see that for residuals distributed according to a Student $T$ with 1.1 and 2.1 degrees of freedom, the proposed method returns more precise estimates, closer to the true coefficients and with narrower intervals. Observe that these are the most problematic cases, with the heaviest tails.

However, for 50 degrees of freedom, the performance of the Bayesian method and of the OLS is perfectly comparable: this is because a Student's t distribution with 50 degrees of freedom is close to a Gaussian, which is the case where the OLS model is perfectly specified. Thus, we can conclude that our method performs better for heavy-tailed residuals, and it is also a good alternative to OLS even in the case of Gaussian residuals.

Table 1: Comparison of our method with classical OLS

| Df | Mean interval width | Mean interval width (OLS) | Mean distance from true coefficients | Mean OLS distance from true coefficients |
|---|---|---|---|---|
| 1.10 | 0.69 | 8.86 | 0.25 | 2.05 |
| 2.10 | 0.55 | 1.13 | 0.19 | 0.34 |
| 3 | 0.51 | 0.69 | 0.16 | 0.20 |
| 50 | 0.39 | 0.42 | 0.14 | 0.13 |

DRAFT

## 4. Discusssion

We have proposed a new methodology to avoid restrictive assumptions on the residuals of a linear model: it works by combining the interpretability of parametric models with the great flexibility guaranteed by Bayesian nonparametric methods. The methodology is shown to empirically outperform the MLE estimate, when the residuals are heavy tailed.

Future work will focus on how to exploit the discreteness of the realisation from a Dirichlet process; indeed, it can be used to identify outliers or to identify relevant subgroups of the observations. The methodology could be also applied to hierarchical models, to combine information from different populations. The results will be detailed and developed in (1).

## References


[1] ASCOLANI, F. and GHIDINI, V. (2023+). Generalized Linear Models with assumptions-free residual. *Work in progress.*

[2] CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *Am. Stat.* **46**, 167.

[3] FAMA, E. F. and FRENCH, K. R. (2004). The Capital Asset Pricing Model: Theory and Evidence. *J. Econ. Perspect.* **18**, 25–46.

[4] FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

[5] GHOSAL, S. and VAN DER VAART, A. (2017). Fundamentals of Nonparametric Bayesian Inference. *Cambridge Series in Statistical and Probabilistic Mathematics.*

[6] HANSON, T. and JOHNSON, W. (2002). Modelling Regression Error with a Mixture of Polya Trees. *J. Am. Stat. Assoc.* **97**, 1020–1033.

[7] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *J. Am. Stat. Assoc.* **96**, 161–173.

[8] LENIN, D. A., PERES, Y., WILMER, E. L. (2008). Markov Chains and Mixing Times. *American Mathematical Society.*

[9] PATI, D. and DUNSON, D. B. (2014). Bayesian nonparametric regression with varying residual density. *Ann. Inst. Stat. Math.* **66**, 1–31.

[10] PREMARATNE, G. and BERA, A. K. (2000). Modeling Asymmetry and Excess Kurtosis in Stock Return Data. *Office of Research Working Paper Number* 00 − 0123*, University of Illinois.*


DRAFT

# Applications of data visualization for industry

Martina Dossi[a], Stefano Sangaletti[a], Marilena Di Bari[a], and Federica Bruschini[a]

[a]Prometeia s.p.a.; martina.dossi@prometeia.com, stefano.sangaletti@prometeia.com, marilena.dibari@prometeia.com, federica.bruschini@prometeia.com

**Abstract**

Data visualization is a powerful tool for conveying data and information in a way that is both accessible and engaging for any kind of organization or business. In a world where data is being generated and accumulated at an unprecedented rate, data visualization has become increasingly essential to make sense of it and support decision-making processes at all levels. Through a number of case studies, this paper presents the perspective of a Data Science team that works in the financial services industry and has acquired a solid experience in data visualization applications at all stages of analytical projects.

***Keywords:*** data visualization, data science, crisp-dm, visual analytics

DRAFT

## 1. Introduction

Data visualization is the rigorous and scientific representation of quantitative information in the form of visual elements. As argued by Tufte in his principles of graphical excellence, "it is a matter of *substance*, *statistics*, and *design*" that allows communicating complex ideas with clarity, consistency and efficiency [8]. In the process of making data easily accessible to a broader audience and for a variety of purposes, data visualization plays a key role in the democratisation of information. It does not merely translate into an attempt at simplifying reality, but rather at capturing and explaining phenomena in all their nuances and implications, bridging people of different cultures and knowledge backgrounds.

The act of visualizing data entails the constant search for new structures to produce valuable revelations of the data in a visual manner. They might be functional to showcase hidden connections, discover trends and patterns, or convey stories that spark discussions and even determine courses of action. Widely accepted definitions of data visualization delineate two distinct approaches: explorative [3] and explanatory [9].

- Explorative data visualizations support all stages of a data analysis that focus on deepening the understanding of the data, in terms of their underlying structure, quality and multidimensional relationships.
- Explanatory data visualizations are used to provide guidance, reveal insights and communicate results on a certain topic when the research about it is more mature, stable and ready to be shared with some generic audience.

Besides this broad distinction, the two approaches are often combined to serve a spectrum of different needs and uses that are intertwined with its inherent multidisciplinarity, as many disciplines actually contribute to defining data visualization (e.g. statistics, graphical design, computer science, communication, psychology, semiotics). In the private sector, data visualization has become a common language among executives, business leads and data analysts to cooperate, interpret data and disseminate information that might have far-reaching business impact. This paper illustrates applications of data visualization from the point of view of an analytical team that leverages data science techniques to support clients in the financial services industry.

## 2. Data visualization in data science projects

Data science projects are aimed at solving typically complex and real-world problems by means of data-driven techniques and domain knowledge considerations. Data visualization is tailored to underpin a deeper interpretation of data and models as the project evolves during its life cycle and has to adapt to the variability of the context to which it is applied. Over the years, several methodologies have been proposed to standardize data science workflows through comprehensive, application- and technology-agnostic frameworks [2]. Among these, the CRISP-DM (CRoss-Industry Standard Process for Data Mining) paradigm [6] is a popular standard. It prescribes the steps to be taken, along with practical advice, in a structured yet flexible manner that is well suited to the needs of different use cases. Despite more recent advancements, we still refer to the original methodology, as newer approaches are mostly aimed at integrating team-based project management practices [5], which are outside the scope of this paper. The CRISP-DM is a circular and iterative process model consisting of six interconnected phases, as described in Fig. 1.

DRAFT

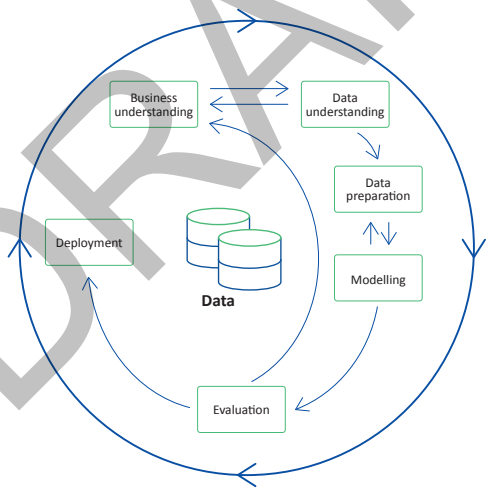



Figure 1: Phases of the CRISP-DM process model.

- *Business understanding*: comprehension of the context and customer's needs, definition of business requirements and conception of a preliminary project plan.
- *Data understanding*: initial data exploration towards a general understanding of the available data and its sources.
- *Data preparation*: data analysis and processing to produce a clean version of the original raw data, to be used for the subsequent phases.
- *Modelling*: training and evaluation of multiple models with chosen performance metrics.
- *Evaluation*: interpretation of the results and selection of the model that best suits the initial business requirements.
- *Deployment*: deployment of the completed pipeline and sharing of the results with external people.

## 3. Case studies description

This section draws on the Prometeia s.p.a. Data Science Team (in the following DST) experience to illustrate applications of data visualization that have proven particularly beneficial and effective in bringing innovation to the analytical projects undertaken. The first two examples (Fig. 2, Fig. 3) are exploratory charts that provide non-standard perspectives on model evaluation by enlightening on complementary aspects: model sensitivity and interpretability. The last three (Fig. 4, Fig. 5, Fig. 6) are well-rounded explanatory data visualizations, developed to be precious allies for decision-making processes either for internal or external stakeholders. Most of the presented charts have been developed with Python and its graphical libraries (e.g. *Seaborn*[1], *Matplotlib*[2]) – apart from a few exceptions that will be specified accordingly.

**Model sensitivity analysis** The case study presented in Fig. 2 refers to the outcome of a binary classification model for predicting the probability of default (probability of not repairing a loan) for some customers within the year following the period of model estimation. The idea is to compare the past trend for the predicted probabilities of the two classes to understand the sensitivity of the model in detecting long-term signals, i.e. how far in advance the model is able to detect significant deviations in the behaviour of the two populations, which might indeed turn useful in other applications of risk estimation. Hereafter, we refer to the defaulting class as the positive class and to the other one as the negative class – respectively in red and green in the chart. The two bands describe the model scores distribution over time for two samples of clients drawn in December 2018 from the test set. They are both delimited by the 25th and 75th percentiles – lower and upper bounds – with a straight line in the middle representing the median. In this case, the trend analysis outlines that the model is able to discriminate between the two classes even years before the default date. In fact, the scores for defaulting clients are systematically higher than those of the other class and the gap between them not only increases with the approaching of the default event but also appears at an unexpectedly early stage, leading to conclude that the model is able to detect never-seen-before signals.

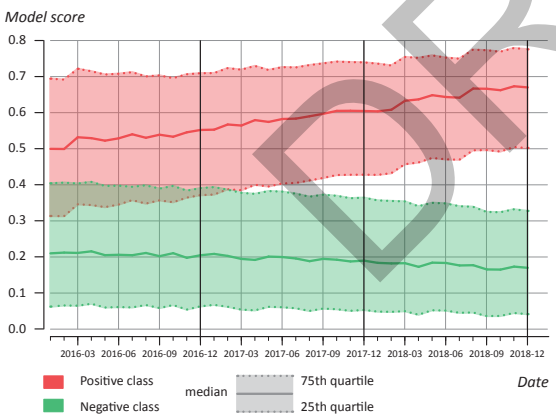



Figure 2: Model scores distribution for out-of-sample data over time, for both the positive and negative class.

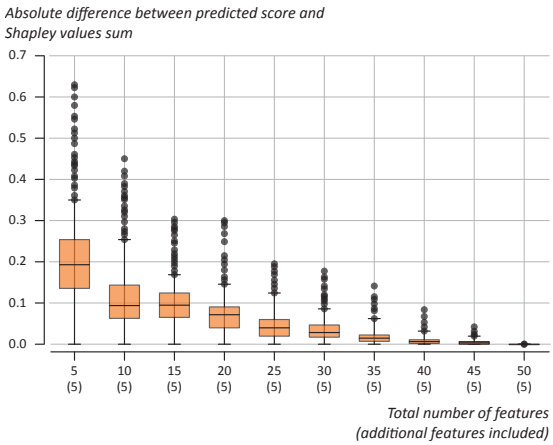



Figure 3: Fraction of the model output explained by the Shapley values sum as the number of included features increases.

**Feature selection for model explainability** The chart reported in Fig. 3 leverages SHAP (SHapley Additive exPlanations) [4] to provide guidance on the number of features actually needed to explain an adequate fraction of the model output. SHAP is a method of model interpretability widely used to

[1] https://seaborn.pydata.org/
[2] https://matplotlib.org/

DRAFT

facilitate the usage of *black-box* models in all those domains where explaining the model outcome is a strong requirement. For each feature, every single data instance receives a Shapley value that expresses the weight of that feature on its *local* model prediction, with the constraint that all the Shapley values must sum to the score. In other words, every observation has its own personal feature importance ranking. The idea is thus to support the human comprehension of the model considering – ideally – only a few *ad-hoc* features for interpreting any single prediction while retaining most of its accuracy. The box plots summarise this concept by showing the distribution of the absolute deviation of the Shapley values sum from the predicted scores, as the number of included features increases. The first box plot, for instance, considers only the five most relevant features in terms of Shapley values for every unit. With five features rather than fifty (the complete feature set), it is possible to explain 75% of the predictions with an approximation error of about 25%, which is the absolute deviation on the y-axis. As features are added to the model, the Shapley values' sum comes closer to the model score, until they are all included.

**Visualizing process mining results for credit requests** Process mining is the integration of data science and process management techniques to support the analysis of operational processes, with the final goal of characterising the processes and identifying eventual bottlenecks or opportunities for increasing efficiency. In the scenario presented, we are considering an application to the credit risk management area, where credit requests are taken over by specialised officers from the collection of the initial documentation to the final outcome of approval or rejection. In this case, data visualization exploits the results of a process mining pipeline carried out through the Python library *pm4py*[3] by providing a graphical representation of the process analysed, which is the whole life cycle of credit requests (Fig. 4). While these requests are expected to follow straightforward paths – with just a few variations – the map clearly outlines the presence of three main workflows: the most standard and linear one lies in the centre, while the other two, less common and therefore hard-to-detect, lie on the edges. This example emphasises how the visual display of tabular data brings nontrivial information to light, enabling the prompt detection of anomalies that can trigger further investigations. For instance, additional charts and metrics may reveal interesting insights about the average execution time of each phase and the presence of loops, thus explaining unusual behaviours and providing data-driven best practices to handle them. Data visualization fills a strategical gap in the bank's risk infrastructure by empowering credit officers with insights that actively support their daily work and speed up business decisions, such as actions to reduce delays when processing a practice, leading to increased efficiency with a timely and standardised organization of the credit approval process.

DRAFT

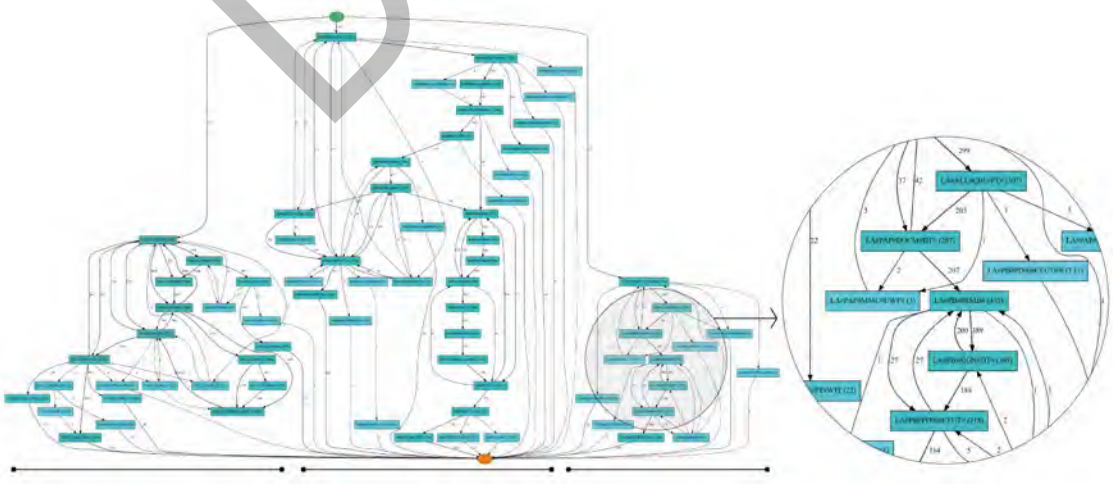


Figure 4: The life cycle of credit requests, grouped in three main buckets.

[3] https://pm4py.fit.fraunhofer.de/

**A dashboard to centralise scattered data and highlight event connections** Behind data visualization artifacts, there is usually a tangled journey of data gathering, cleaning, aggregation, and analysis to move from non-structured data sources to a structured collection of data, ready to be translated into some visual representation. However, all this underlying work is rarely evident in the data visualizations that come at the end of a project's life cycle and tend to be more result-oriented. In contrast, the complex process of data aggregation is at the heart of the presented case study, where we developed a *Power BI*[4] dashboard (Fig. 5) for the risk and compliance function of a bank. The objective was to connect several business processes and explain them through a number of summary statistics and key indicators. The project started with an extensive work of data understanding in order to link all the data sources used by analysts in their daily activities, followed by a stage of data cleaning and homogenisation. The purpose was to provide them with a centralised tool in which all their single reports would converge, with the immediate effect of abating their workload and preventing undesired errors either due to manual operations of data linkage or inconsistency between data sources. Besides connecting all the data, the dashboard presents several indicators to interpret phenomena more efficiently, identify emerging trends and retrace the underlying chain of events, supporting the comparison of interconnected statistics that may guide more reliable analyses about potential cause-effect relationships rather than isolated metrics.

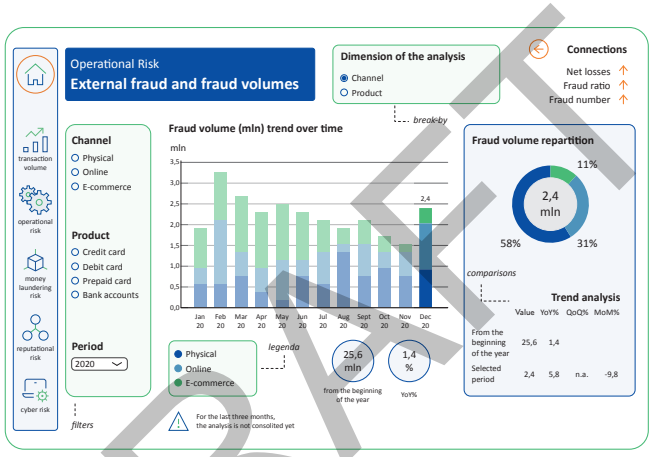



Figure 5: An example of a dashboard page.

**Skills mapping to foster cohesion within a team** The skill map is a powerful tool the DST has developed to depict the landscape of people skills over time and support several other internal processes, ranging from upskilling to resource allocation. It has been conceived as a guided self-assessment exercise with about 80 technical skills arranged in 10 macro groups (e.g. *programming* is a *macro-skill* that includes skills such as *Python*, *R*, *Bash* and others), to be updated on a 6-month basis. For each skill, people are asked to assess their proficiency level on a scale from 1 to 5 (from *no experience* to *expert*). Data visualisation allows to effectively summarise the results of every assessment both individually and in aggregated form. While the former enables personalised and professional growth initiatives, retention strategies and more optimised resource allocation for project teams, the latter reveals any potential skill gap to be handled in the hiring or upskilling process. The representation in Fig. 6 is a comprehensive view of the team composition in terms of proficiency level in all the assessed skills. The graph, produced in *Gephi* [1], depicts data-driven skills clusters from the questionnaire results at a given snapshot date. It originated as a tool to investigate the team composition and then turned into an experiment to explore whether the team intrinsically embeds current professional roles as per the market. Each graph node is a skill connected by edges that are weighted by the cosine similarity of the vectors of proficiency levels. Overlapping circles represent our attempt to find professional roles by skills, which we do not expect to reflect widespread established roles but ones tailored to our reality. The cluster at the centre, associated with the data scientist label, consists of skills that characterise this professional role and are shared by

[4]https://powerbi.microsoft.com/en-au/

DRAFT

most analysts in the team, thus typical among more functions. The other clusters of professional figures encompass more field-oriented competencies, e.g. text analysts relate to a pertinent cluster. In general, the more peripheral the skill clusters are, the more peculiar and less widespread these skills also are.

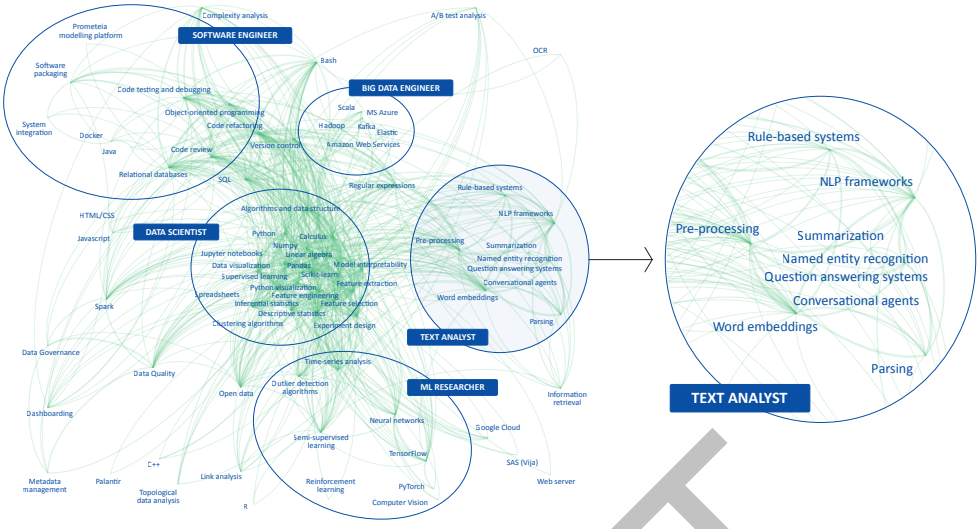



Figure 6: Skill map results: data-driven skill clusters and professional roles overlapped.

## 4. Conclusion

This paper advocates for a conscious use of data visualization as a communication vehicle to fill the gap between different professional roles. The DST constantly relies on visual insights to support any phase of project development and guide business decisions, striving to structure visualization tools that accomplish a variety of goals in the most reproducible and domain-agnostic manner. Further progress is expected toward more interactive visualizations capable of increasing engagement in the end users.

**Acknowledgments** Thanks to Prometeia s.p.a. and the Data Science Team for supporting this work, especially to Federico Crecchi and Maurizio Monaco.

DRAFT

## References

[1] Bastian M., Heymann S., Jacomy M.: Gephi: an open source software for exploring and manipulating networks. Int. AAAI Conference on Weblogs and Social Media. (2009)
[2] Cavaller, V.: Dimensional Taxonomy of Data Visualization: A Proposal From Communication Sciences Tackling Complexity. Frontiers in research metrics and analytics, 6, p.643533 (2021)
[3] Knaflic, C.N.: Storytelling with data: A data visualization guide for business professionals. John Wiley and Sons (2015)
[4] Lundberg, S.M., Lee S.I.: A unified approach to interpreting model predictions. Advances in neural inf. processing systems, 30 (2017)
[5] Martinez, I., Viles, E. and Olaizola, I.G.: Data science methodologies: current challenges and future approaches. Big Data Research, 24, p. 100183 (2021)
[6] Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S. and Wirth, R.: Crisp-Dm 1.0—Step-by-step data mining guide. Cris. Consort, p.76 (2000)
[7] Schmidt, J: Usage of Visualization Techniques in Data Science Workflows. (2020)
[8] Tufte, E.: The Visual Display of Quantitative Inf. Graphic Press, Cheshire (2001)
[9] Yau, N.: Data points: visualization that means something. John Wiley and Sons (2013)

# Some Notes on the Use of the Circular Boxplot

Giovanni C. Porzio[a] and Davide Buttarazzi[a]

[a]Sample Survey and Data Analysis Lab, University of Cassino and Southern Lazio;
porzio@unicas.it, davidebuttarazzi@outlook.com

**Abstract**

Data lying on the boundary of the unit circle are generally referred to as circular data. Recorded as angles or unit norm vectors on a plane, they are mainly used to describe directions and time cycles in a variety of fields.

Because of their bounded and nonlinear nature, circular data require specific tools to be analyzed. To exemplify, if the standard sample mean is applied to a set of angles, the final result could be very disappointing. Specific visualization tools are also required. If, for instance, the aim is to visualize a set of angles, their histogram can be misleading again.

Standard visualization tools for directional data include the circular histogram and the rose diagram. Alternative tools are the circular nonparametric density estimator and the CircSiZer map. One more tool, introduced five years ago, is the circular boxplot.

Since its introduction, this latter has been exploited to visualize circular distributions in a series of different application settings and with different goals. Five years later, we aim to provide a critical review of the ways this boxplot has been employed, and to highlight the pros and cons of its use in practice. A modified version particularly conceived to deal with anomalous data will also be discussed.

***Keywords:*** Directional statistics, directional data depth, visualization.

DRAFT

## References

[1] Ameijeiras-Alonso, J., Crujeiras, R.M.: Flexible Circular Modeling: A Case Study of Car Accidents. In: SenGupta, A., Arnold, B.C. (eds.) Directional Statistics for Innovative Applications, pp. 93-116. Springer, Singapore (2022)

[2] Buttarazzi, D., Pandolfo, G., Porzio, G. C.: A Boxplot for Circular Data. Biometrics **74**, 1492-1501 (2018)

[3] Li, C., Chiang, K.P., Laws, E.A., Liu, X., Chen, J., Huang, Y., Chen, B., Tsai, A.Y., Huang, B.: Quasi-Antiphase Diel Patterns of Abundance and Cell Size/Biomass of Picophytoplankton in the Oligotrophic Ocean. Geophys. Res. Lett. **49**, e2022GL097753 (2022) doi: 10.1029/2022GL097753

[4] Rodriguez-Quintero, W.D., Moreno-Chacón, M., Carrasco-Urra, F., Saldaña, A.: From dark to darkness, negative phototropism influences the support-tree location of the massive woody climber Hydrangea serratifolia (Hydrangeaceae) in a Chilean temperate rainforest, Plant Signal. Behav. **17** (2022) doi: 10.1080/15592324.2022.2122244.

[5] Salimpour, Y., Mills, K.A., Hwang, B.Y., Anderson, W.S.: Phase-targeted stimulation modulates phase-amplitude coupling in the motor cortex of the human brain. Brain Stimul. **15**, 152-163 (2022) doi: 10.1016/j.brs.2021.11.019.

# TERRA: a smart visualization tool for international trade in goods statistics

Francesco Amato[a], Mauro Bruno[a] and Maria Serena Causo[a]

[a] Istat; framato@istat.it, mbruno@istat.it, causo@istat.it

**Abstract**

International trade in goods statistics is a rich data source, freely available on Eurostat open-data portal and updated at high timeliness. Data users can explore several aspects of the statistical domain, but, as it happens when many paths depart in front of us without no clear road signs, some users could find themselves "lost in data". In such a complex context, a smart data visualization tool, i.e., TERRA, can bring a relevant added value to data dissemination providing relevant timely indicators and a first approach to exploratory trade data analysis. In this work we describe the core functionalities of TERRA and provide a brief overview of the architectural components of the dashboard. The latest version of TERRA will be soon published as an experimental statistic by Istat.

**Keywords:** Smart data visualization, External Trade, COMEXT, Graph analysis, Time series

DRAFT

## 1. Introduction

International trade in goods statistics published by Eurostat measure the value and quantity of goods traded between the EU Member States (intra-EU trade) and goods traded by the EU Member States with non-EU countries (extra-EU trade). Data are compiled by Member States in harmonised way based on concepts and definitions set out in EU legislation and are characterised by many dimensions of analysis and high granularity level not only in terms of geographical destination of trade flows, but also in terms of traded products. Indeed, trade data are disseminated at monthly time frequency at the most detailed level of the following product nomenclatures: the Combined Nomenclature (CN), the Standard International Trade Classification (SITC), the Broad Economic Categories classification (BEC), the Classification of Products by Activity (CPA) and the Standard Goods Classification for Transport Statistics/Revised (NST/R). Moreover, trade flows are classified by mode of transport, providing relevant information for defining transportation policy, monitoring international transport routes, and assessing the impact of trade on the environment.

As international trade forms a major part of the world economy, statistics on trade in goods are an instrument of primary importance for numerous users [1, 2], including public and private sector decision makers. For example, international trade in goods statistics are valuable in order to:

- inform on recent and long-term developments in trade and economy;
- help EU businesses conduct market research and define their commercial strategy;
- enable EU authorities to prepare multilateral and bilateral negotiations under the common commercial policy;
- enable EU authorities to evaluate the progress of the Single Market and the integration of EU economies;
- enable EU authorities to define and implement anti-dumping policies;

- provide an essential source of information for other statistical domains, as Balance of Payment (BoP) statistics or national accounts.

With such a rich data source, freely available and updated at high timeliness, data users can explore several aspects of international trade, but, as it can happen when many paths depart in front of us without no clear road signs, some users could find themselves "lost in data". A smart visualisation tool, presenting trade data in suitable graphical form that illustrates the evolution of trade flows not only in terms of trading volumes, but also in terms of composition of the basket of traded goods, and that provide, moreover, tools for analysis embedded in the system, can bring relevant added value to data dissemination, and facilitate a first approach to exploratory trade data analysis. TERRA (imporT ExpoRt netwoRk Analysis) is a data visualization tool specifically designed at Istat for this purpose.

## 2. The dashboard TERRA

TERRA monthly processes about one billion records relating to the commercial exchange of goods with foreign countries, produced by the 27 Member States according to harmonized methodologies and publicly available on Eurostat's COMEXT[1] database. Therefore, official estimates on trade flows in monetary value and in physical quantities at the maximum granularity in temporal resolution (monthly frequency), characteristics of the traded product, trading partner countries, mode of transport, provide TERRA with the source of information for further elaboration.

The main functions implemented in TERRA[2] allow to analyse the impact of shocks in means of transport and the effects of interruptions in trade relations between countries for specific products with social network analysis techniques, offering a set of global indicators, typical of graph analysis. The dashboard is publicly available at the following link: https://terra.statlab.it/.

### 2.1 Economic and international trade indicators

This section provides a map showing, for each Member State, macroeconomic indicators, main imported and exported products, main trading partner countries (as shown in Figure 1). A time-lapse functionality displays the time evolution of year-on-year changes in traded values for the last three years, giving a simple and clear picture of country trade effectiveness.

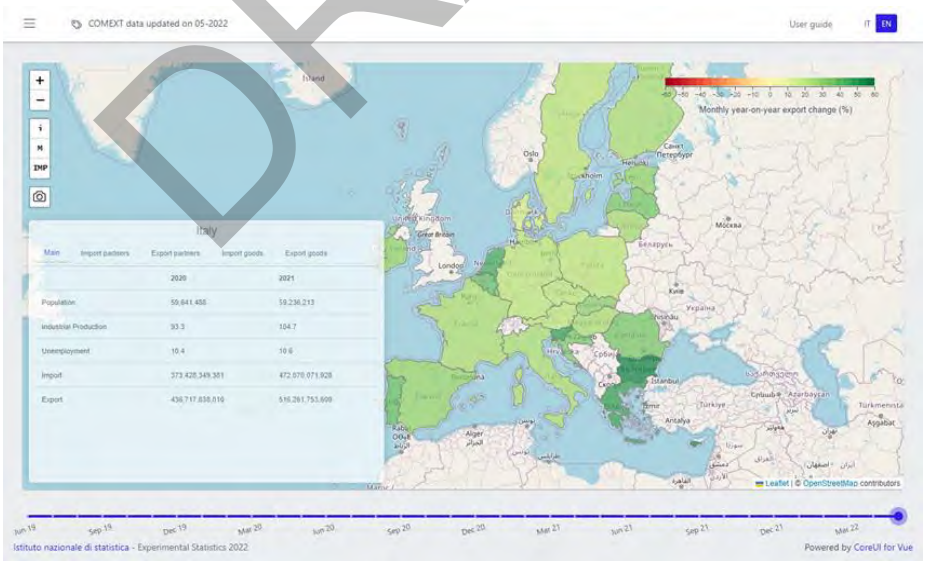



Figure 1: TERRA interactive map

[1] COMEXT database can be accessed using the Eurostat bulk download tool, available from: https://ec.europa.eu/eurostat/data/bulkdownload

[2] TERRA is an open-source project, the source code is publicly available from: https://github.com/istat-methodology/terra

DRAFT

## 2.2 International trade EU - Extra EU relations

This section displays graphs representing the network of international trade between EU and extra-EU countries by product and mode of transport, together with relevant global and local graph indicators [3,4,5]. In the right panel (as shown in Figure 2), through the drop-down menus, the user can select the parameters of the analysis. Based on the values chosen, the graph is created, and the respective centrality measures are calculated. The main indicators displayed by TERRA are the following:

- *Product spread*: global measure corresponding to the graph density and representing how much the product is spread through the graph.
- *Vulnerability*: local measure corresponding to (1 - the indegree centrality[3])
- *Export strength*: local measure corresponding to the outdegree centrality[4].
- *Hubness*: local measure corresponding to the closeness centrality for each country.

The tool contains a panel showing, in tabular form, the metrics of the graph for each country: *Export strength*, *Hubness*, *Vulnerability*. The user can sort the results by clicking on the arrows in the numerical columns. It is also possible to filter the results by code and country name.

The tool allows to explore possible new scenarios by deleting edges, i.e., interrupting trade paths between specific countries and for specific modes of transport. An animation tool is provided to follow evolution of the graph structure over time. An example of the analysis that can be performed using TERRA is displayed in Figure 2.

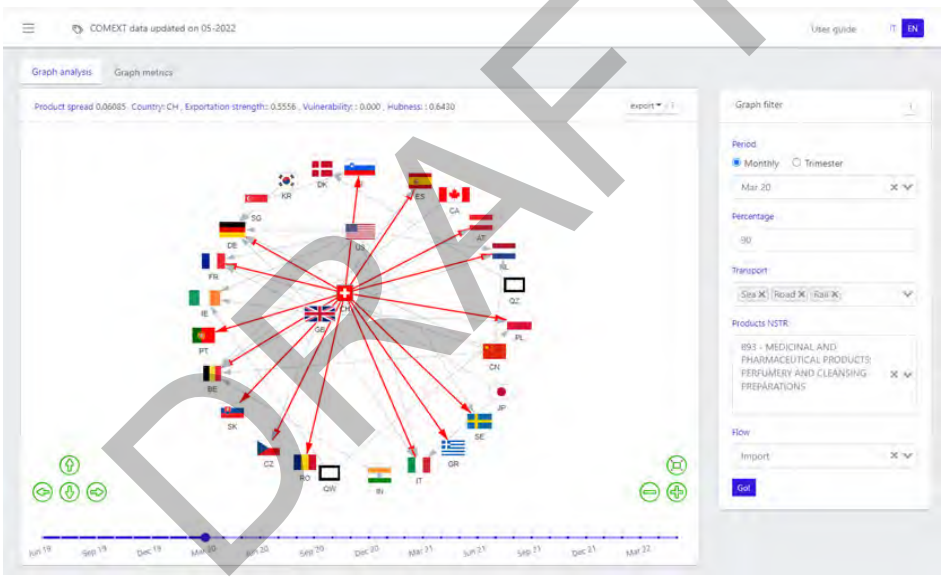



Figure 2: TERRA graph analysis

## 2.3 Time series visualization

Tool for visualizing times series monthly traded values and quantities between partner countries and for products (see Figure 3). Series autocorrelation functions and Q-Q plots are provided, allowing a first approach to exploratory time series analysis. The section provides users with a large number series representing bilateral trade for specific products. Series and indicators can be downloaded in several formats suitable for further analysis.

[3] The indegree centrality is the number of connections that point inward at a vertex.
[4] The outdegree centrality is the number of connections that originate at a vertex and point outward to other vertices.

DRAFT

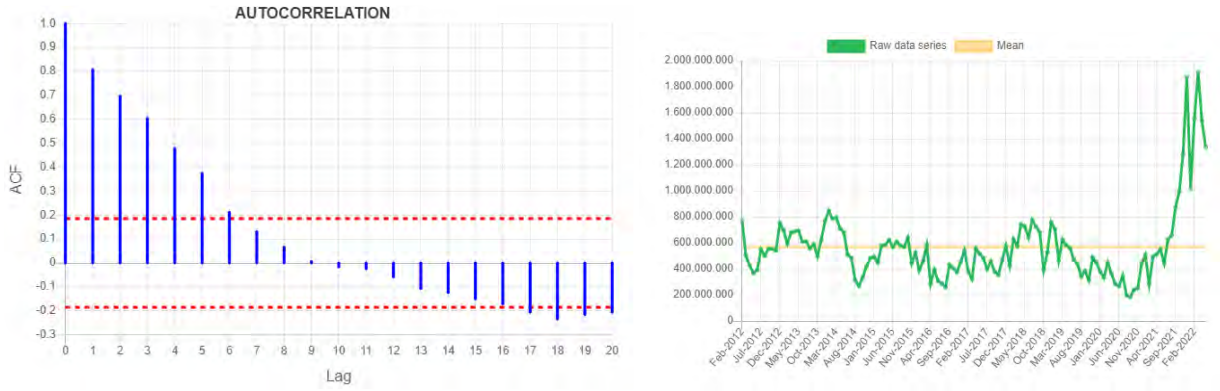



Figure 3: TERRA time series visualization

## 2.4 Basket of products

In the section relating to the basket of traded products, the trends of the monthly shares, in value or quantity, of imported and exported products as defined according to the two-digit breakdown of the CPA 2.1 classification are represented for each Member State, together with the related trend variations. The interest of the section lies in the possibility of monitoring which types of products are more traded in periods of markets non-equilibrium, associated with crisis and subsequent economic recovery phases. An example, concerning the increase in the trade of essential products during the first phase of the COVID pandemic, is displayed in Figure 4. The section allows an easy country level comparison of different trade strategies put in place as a reaction to external shocks, or as a long-term trend. As an example, in Figure 4 year-on-year percentage changes in the share of Italian exported product are displayed. One can notice that in spring 2020, first phase of Covid-19 pandemic, the share of exported basic necessities (basic pharmaceutical products, food products, products of agriculture) increased with respect to the previous year, and, at the same time, export of less necessary products (motor vehicles, leather, wearing apparel, textiles, furniture) faced a contraction. The opposite phenomenon is evident in the 2021 economic recovery phase, where the gain in share for motor vehicles, furniture, leather and wearing apparel, is very evident.

DRAFT

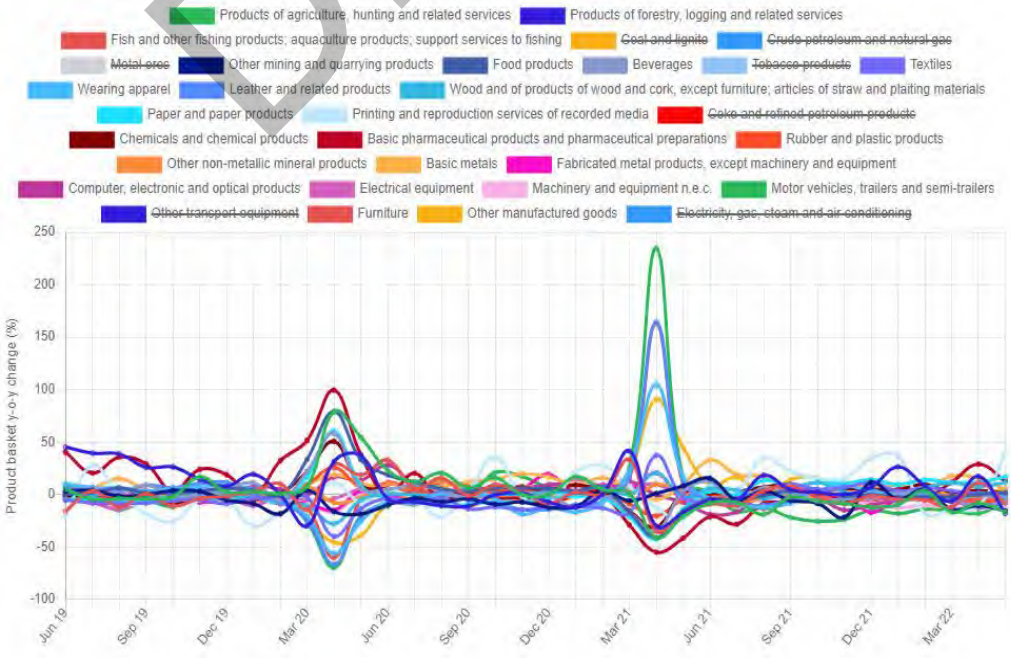



Figure 4: TERRA, share variation in the basket of Italian exported products

## 3. Architecture

The architectural solution adopted for the implementation of TERRA arise from the need to make the data analysis timely and available to the stakeholders a few hours after Eurostat COMEXT monthly data is published. The application retrieves and processes each month data covering a time series of 10 years. Further, TERRA allows to select two different type of monthly time series - raw data and yearly variation, until the latest released update. This corresponds to processing about 1 billion records each month.

To get and process such a large amount of data, we realized a specific batch program (Batch processing in the left panel of Figure 5), implemented in Python language. The batch script is scheduled to start every 24th day of the month, and, automatically, downloads file, performs processing, produces outputs and, finally, updates the data stored on the server.

To speed up the elaboration time, the script runs in cloud platform, and data are processed using high-performance algorithms, using the library PySpark on Apache Spark Framework. The script is configured to access to COMEXT bulk download section of EUROSTAT portal and starts a parallel process of downloading 136 files of monthly products data, 2 files for annual products data, 36 files for monthly transport data and the files containing their classifications. Because of the heterogeneity and complexity of the data analysis algorithms, we chose a microservices architectural paradigm [6]. This approach of dividing the functions performed by the application into n-services allows the principle of individual responsibility: each service is implemented and distributed separately from the others.

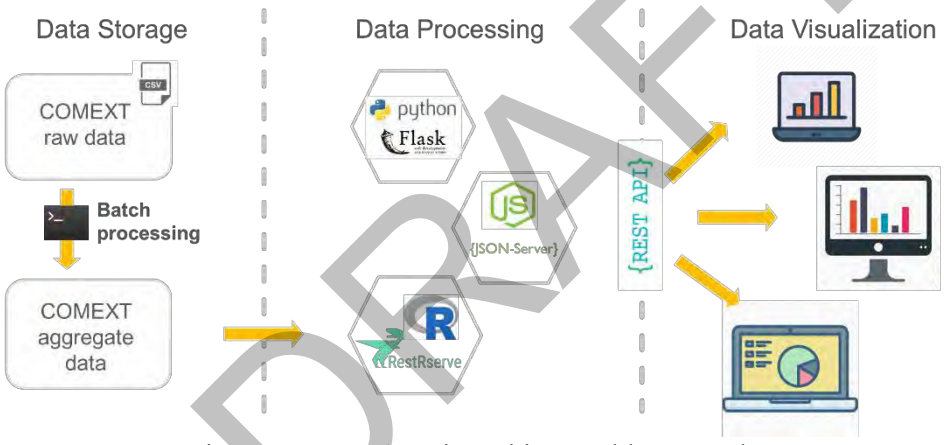



Figure 5: TERRA main architectural layers and components

Figure 5 sketches the main architectural components of TERRA. Three different layers can be identified in the architecture: data storage, data processing, and data visualization.

- **Data storage layer**, contains the raw data downloaded from COMEXT portal, as described above. Further the Batch processing stores in this layer the aggregated data that are further processed by the microservices in the data processing layer.
- **Data processing layer**, or backend, includes the three microservices: the first is devoted to the processing of Python scripts; in the same way, the second is dedicated to the processing of scripts in R. Finally, the third microservice exposes static data, such as classifications and metadata. The frontend communicates with the microservices through requests according to the HTTP protocol, exchanging messages in JSON format.
- **Data visualization layer**, or frontend, includes the web component as the user interface. This application was implemented as a Single Page Application (SPA) using open-source web frameworks. The use of these modern technologies makes the application responsive, i.e., the layout adapts to the size of the display, therefore it is possible to access the dashboard both from PCs and mobile devices.

DRAFT

## 4. Conclusions

TERRA is the first example of smart visualization tool for international trade statistics specifically developed with the aim of bridging the gap between simple graphical web-based tools for data visualization and more complex data processing systems. The result is a system, which is user friendly as a simple graphical tool, but in the other hand has a hidden complex core for massive data processing and statistical indicators calculation. Although further improvements can be introduced by enlarging the set of synthetic indicators provided, the product has been so far welcomed by the statistical community and can be used as a prototype for applications in different statistical domains.

## References


[1] Schmitz H.: Value Chain Analysis for Policy-makers and Practitioners. International Labour Organization (2005)

[2] Jara A., Escaith H.:Global Value Chains, International Trade Statistics and Policymaking in a Flattening World. In: World Economics, 13 (4) (2012)

[3] Opsahl, T., Agneessens, F., and Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32 (3), 245-251 (2010)

[4] Fruchterman, T.M., Reingold E.M.: Graph drawing by force-directed placement. Software: Practice and Experience, 21 (11), 1129 -1164 (1991)

[5] De Benedictis L., Nenci S., Santoni G., Tajoli L., Vicarelli C.: Network Analysis of World Trade using the BACI-CEPII dataset, CEPII Document de travail 24 August (2013)

[6] De Lauretis L.: From Monolithic Architecture to Microservices Architecture. IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), DOI: 10.1109/ISSREW.2019.00050 (2019)


DRAFT

# Clustering of Distributional data based on LDQ transformation

Gianmarco Borrata [b] and Rosanna Verde[a]

[a]Dept. of Mathematics and Physics, University of Campania Luigi Vanvitelli, 81100 Caserta, Italy;
rosanna.verde@unicampania.it
[b]Dep. Social Science, University of Naples Federico II, Vico Monte della Pietà, 1, 80138 Napoli;
gianmarco.borrata@unina.it

**Abstract**

In this paper, we present a new clustering algorithm for data represented in the form of distributions. The contribution of this work, compared to clustering methods for distributional data ((7)), is to consider a suitable data transformation. We referred to a logarithmic transformation of the derived quantile functions (LDQ) associated with the distributional data. This transformation makes it possible to provide a mapping of the density functions in a Hilbert space. The proposed method is based on the classical dynamic clustering algorithm (4) on the LDQ data transformation. The results are then compared with those of an extension of the k-means clustering algorithm on distributional data. Finally, given the different partitions provided by the data distribution, such as quantile and LDQ functions, we propose to combine the partitions obtained on the different transformed data to obtain a single partition. Preliminary results on real environmental data have confirmed the effectiveness of the proposed method.

***Keywords:*** Distributional data, Dynamic clustering, LDQ transformation

DRAFT

## 1. Introduction

In this work, we present a new clustering algorithm for data represented in the form of distributions. The input consists of a set of $N$ observed with respect to $X_1, \ldots, X_p$ distributional variables. Each individual is represented by $p$ $x_{i1}, \ldots, x_{ip}$, probability functions or estimated ones, like histograms. Distributional Data are defined according to (2) and in the recent years many Data analysis techniques have been developed on such kind of data, some references are available in (3) and Several clustering algorithms were proposed for a partitioning a set of distributional data into subgroups using a K-means like algorithm. The centroids of the clusters are represented by barycentric functions according to the appropriate dissimilarity measure. Elements are assigned to clusters with respect to the minimum distance from the centroid consistent with the algorithm's optimised criterion of minimum internal variability of clusters.

Previous works, such as (1) and (8), proposed clustering models for distributional data based on the Euclidean and the Wasserstein metric in a linear space. Recent developments in distributional data analysis (DDA), have introduced a transformation of the quantile functions into the Logarithm of the derivative of the quantile functions (9). This transformation allows for mapping density functions into a Hilbert space addressing some issues in DDA. As is well known, a vector representation of complex data (such as distributions) in a Hilbert space allows the definition of numerical operators (e.g. the inner

product) and distance measures to compare distributions. The LDQ transformation can return the results of the analysis in the same form as the original data through an inverse transformation.

Based on this new transformation of the distributional data, we propose a clustering method by performing a partitioning of the set of objects in $k$ clusters according to the minimum internal variability of the clusters formed by LDQ functions. The centroids of the clusters are defined as the average of the LDQ functions of the elements of each cluster and the dissimilarity measure for assigning elements to clusters is taken as the Euclidean distance between the LDQ functions.

The clustering process is performed in two alternating phases. Once a predefined number of $K$ clusters has been set, the first stage involves the representation of the clusters with the averages of the LDQ functions of the cluster elements; in the second stage, the elements are assigned to the clusters according to the minimum of the sum of the squared distances between the LDQ functions of the cluster elements and the respective centroid. This process is repeated until convergence to stable clusters.

The outcome of the partition obtained by this new clustering method has been compared with the results of the clustering method on DD based on the Wasserstein squared distance. It was worth noting that the results of the achieved partitions are different, due to the kind of LDQ transformation which enounces the characteristics of the distributions for effect of the derivative of the quintile functions and of the logarithmical transformation of the derivatives of the quantiles. Some interesting considerations on the different similarity of the distributional data captured by this LDQ transformation made one consider the idea of combining the two distributional data descriptions through a partitioning strategy. Therefore, the proposal has considered a combination of the two partitions by associating the descriptors with a weight related to the importance given to the original distributions by the quantiles and the LQD transformed functions, in the partitioning process.

Preliminary results on the environmental data observed over time, aggregated in the form of distributions, made it possible to assess the effectiveness of the proposed strategy and to indicate different weight settings in the combined partition strategy.

DRAFT

## 2. Distributional-valued data

Let $\Omega$ be a set of objects. A modal (2) variable $X$ with domain $\mathscr{D}$ on the set $\Omega$ is a mapping $\Omega \to \mathscr{M}$ of all possible measures $\pi$ on $\mathscr{D}$ : $\omega \to X(\omega) \in \mathscr{M}$, for $\omega \in \Omega$.

Histograms are a suitable way of representing aggregated data, like distributions. In the context of SDA, a variable $X$ is considered a histogram-valued variable if each object $\omega$ is represented by an estimate distribution in the form of a histogram (2; 3).

Formally, let's $x$ be a realization of $X$ with support $D = [a_1, b_H]$, that is partitioned into a set of contiguous no-overlapped intervals (or bins) $I_h = [a_h, b_h]$, $\forall h : I_h \subseteq D$.

A non negative weight $\pi_h$ (a probability or a relative frequency) is associated to each $I_h$. A histogram data is defined by a sequence of intervals (bins) with associated respective weights:

$$x_i = X(\omega_i) = [(I_1, \pi_1), \ldots, (I_h, \pi_h), \ldots, (I_H, \pi_H)] \tag{1}$$

A cumulative distribution function $F(x)$ is associated to each $x$. $F(x)$ is a continuous function and, for cumulative distribution function is a piece-wise function, strictly increasing in the interval $[a_1, b_H]$; $0 \leq F(x) \leq 1$; $F(x)$ is differentiable on $[a_1, b_H]$.
The inverse of a distribution function $F^{-1}(x) = Q(t)$ is denoted as a quantile function. It is a monotone increasing function with support in $[0,1]$ (for a histogram, it is a piece-wise function). $Q(t)$ is also differentiable on $[0,1]$.

More recently a suitable transformation has been introduced for mapping probability densities to a Hilbert space of functions through a continuous and invertible map (9). A probability density function is firstly transformed by the derivative of its quantile function:

$$q(t) = \frac{dQ(t)}{dt} = \frac{dF^{-1}(t)}{dt} = \frac{1}{f(Q(t))}$$

which is strictly positive and continuous in its domain $[0,1]$, and then, by the logarithmic transformation of $q(t)$, as follows:

$$l(t) = \ln q(t) = \ln \frac{1}{f(Q(t))} = -\ln f(Q(t))$$

Let $l(t)$ denote the Logarithm Derivative of the Quantile function (LDQ). On these functions it is possible to define the addition and scalar multiplication operations, the inner product and the Euclidean norm. The squared Euclidean distance between two LDQ functions $l_1(t)$ and $l_2(t)$ is given by:

$$d^2_{LDQ}(l_1(t), l_2(t)) = \|l_1(t) - l_2(t)\|_2 = \int_0^1 (l_1(t) - l_2(t))^2 \, dt$$

The main problem with LDQ transformation is that it loses information on the location parameters of the density distribution. In fact, the derivatives of two quantile functions that differ by a constant term are equal.

## 2.1 Clustering algorithm of distributional data

In the framework of distributional-valued data analysis, many statistical analysis methods have been proposed, as clustering methods, regression models, factorial approaches, and many others. A reference book on these themes is (3).

Clustering methods for DD have been proposed mainly based on the classical Dynamic Clustering Algorithm (DCA).

DCA looks for the partition $P \in P_k$ of $\Omega$ in $k$ classes, among all the possible partitions $P_k$, and the vector $L \in L_k$ of $k$ prototypes representing the classes in $P$, such that, the following $\Delta$ fitting criterion between $L$ and $P$ is minimized:

$$\Delta(P^*, L^*) = Min\{\Delta(P,L) \mid P \in P_k, L \in L_k\}. \tag{2}$$

Such a criterion is defined as the sum of dissimilarity or distance measures $d(x_i, G_h)$ of fitting between the elements $x_i$ belonging to a class $C_h \in P$ and the representative of the class $G_h \in L$:

$$\Delta(P,L) = \sum_{h=1}^{k} \sum_{x_i \in C_h} d(x_i, G_h). \tag{3}$$

Generally the criterion $\Delta(P,L)$ is based on an additive distance on the $p$ descriptors. As the criterion is additive, the optimization problem can be solved for each variable $X_j$ (for $j = 1, \ldots, p$).

The representative or prototype $G_h$ associated to a class $C_h$ is an element or a function of the elements $\omega_i \in \Omega$ belonging to $C_h$. The prototype of a cluster $C_h$ of the partition $P$ is defined as the element $G_h$ which minimize the following function:

$$f(G_h) = min_{j=1,\ldots,k} \sum_{x_i \in C_h} d^2(x_i, G_j) \tag{4}$$

- The algorithm is initialized by generating $k$ random clusters or, alternatively, $k$ random prototypes.

The criterion function $\Delta(P,L)$ is optimised in two alternating steps:

1 - a representation step $(P, L^*)$: the protopypes of the clusters are computed as the elements which minimizing the $f(G_h)$;
2 - an assignment step $(P^*, L)$: the elements are assigned to the clusters according to the minimum distance $d(x_i, G_h)$ to the prototype $G_h$:

DRAFT

$$x_i \to C_h \quad \text{if} \quad d(x_i, G_h) < d(x_i, G_{h'}) \quad \text{with} \quad h \neq h'$$

According to the above Clustering algorithm, we assume to partition the set of elements of $\Omega$, on the distributional data transformed in LDQ functions $l_i(t) \in \Lambda$

The DCA performs the best partition of the set $\Lambda$ alternating the 1 - representation step and the 2 - assignment step. The Euclidean distance $d_{LDQ}$ between LDQ functions is assumed as the dissimilarity measure in the algorithm.

1 - The prototype $L_h$ of the cluster $C_h$ is defined as the average $L_h(t)$ of the $l_i(t)$ for $i \in C_h$ as the minimizing elements of the function **??**.

2 - the $l_i(t)$ are assigned to the class $C_h$ if $d_{LDQ}(l_i, L_h) < d_{LDQ}(l_i, L'_h)$ with $h \neq h'$

The main clustering approach on DD (7) is based on a suitable metric for comparing distributions, the quadratic $2-norm$ Wasserstein distance (known as the Mallow distance). In one-dimensional space, this metric corresponds to the Euclidean distance between the two quantile functions, $Q_i(t)$ and $Q_{i'}(t)$, associated with the representation of the two objects $\omega_i, \omega_{i'}$ :

$$d_W(x_i, x_{i'}) = \int_0^1 (Q_i(t) - Q_{i'}(t))^2 dt. \tag{5}$$

We have observed on synthetic and real data that the partitions $P_k$'s obtained with the two approaches are different. That is due to the different information about the characteristics of the distributional data expressed by the quantile and the LDQ transformation functions. In order to preserve both types of information, we introduce a combination of clustering approaches by collaborating the two types of functions. The criterion function is thus expressed as follows:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{x_i \in C_h} [\alpha_{1h} \cdot d_{LDQ}(l_i, L_h) + \alpha_{2h} \cdot d_W(Q_i, G_h)] \tag{6}$$

where: $\alpha_{1h} \in A_1$ and $\alpha_{2h} \in A_2$ are two sets of $k$ weights related to the different importance assumed by the two representation functions of the distributional data, that is the LDQ and quantile functions.

A double set of prototypes $G_1, \ldots, G_k$ and $L_1, \ldots, L_k$, respectively associated to the quantile function $Q_i \in \Omega$ and to the LDQ $l_i(t) \in \Lambda$, are obtained as solutions of the optimizing process at the step 1.

The assignment of the elements $\omega_i$ to a cluster $C_h$ is performed according to the minimum weighted distances of the associated functions from the respective prototypes:

$$(\alpha_{1h} \cdot d_{LDQ}(l_i, L_h) + \alpha_{2h} \cdot d_W(Q_i, G_h)) < (\alpha_{1h'} \cdot d_{LDQ}(l_i, L_{h'}) + \alpha_{2h'} \cdot d_W(Q_i, G'_h))$$

at the step 2. The algorithm is iterated until a stable partitions in $k$ clusters of the DD in $\Omega$ is reached. The convergence of the criterion $\Delta(P, L)$ is guaranteed by the decreasing of the criterion in the alternating steps (the demonstration is referred to the classical DCA one (5)).

## 3. Preliminary results on real data

The performance of the proposed method was tested on real data. The dataset consists of 59 observations from various monitoring stations collecting environmental data on an hourly basis, located in different regions of Italy. Four variables were considered: X1 (Benzene), X2 (NO2), X3 (PM2.5) and X4 (PM10), each represented by a density function. In order to compare the clustering results calculated with respect to the two sets of functions, a concordance measure is required.

The Rand index is a measure of similarity between two clusters of data, ranging from 0 to 1. When the two partitions are in perfect agreement, the Rand index is equal to 1.

A problem with the Rand index is that the expected value of the Rand index of two random partitions does not assume a constant value (e.g. zero). The corrected Rand index proposed by (6) assumes the generalised hypergeometric distribution as a model of randomness, i.e. the two partitions are chosen at random and the number of objects in the classes and clusters is fixed ((10)).

From the comparison of the quantile and LDQ partitions, the Rand index value is 0.52, with an adjusted

DRAFT

Rand index of 0.04 for k=2. When k=3, the Rand index value increases to 0.64, with an adjusted Rand index of 0.20. Similarly, for k=4, the Rand index value is 0.66, with an adjusted Rand index of 0.08. The partitions are significantly different from each other and capture distinct information. Therefore, we propose combining the partitions obtained from the different data transformations to obtain a single partition.
The better results are reached for $k = 3$ clusters according to the Elbow method. The quality of the partition computed as ratio between the internal and total variability is 0.12 confirming the performance of the method in partitioning the distributional dataset.

## References


[1] Billard L., Diday E. Clustering Methodology for Symbolic Data. John Wiley & Sons Ltd. (2020)
[2] Bock H., Diday E.: Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media, (1999).
[3] Brito P., Dias S.: Analysis of Distributional Data. Chapman Hall (2022)
[4] Diday E.: Introduction à l'analyse factorielle typologique, Revue de Statistique Appliquée, XXII(4), 29-38, ( 1974)
[5] Diday E.,Simon J.C.: Clustering analysis. Digital Pattern Recognition, 47-94 Springer (1980)
[6] Hubert L., Arabie P.: Comparing partitions. Journal of Classification, 2, 193-218 (1985)
[7] Irpino A., Verde R. Dynamic clustering of interval data using a Wasserstein based distance. Pattern Recognition Letters, 29, 11, 1648-1658 (2008)
[8] Dynamic Clustering of Histogram Data Based on Adaptive Squared Wasserstein Distances. Expert Systems with Applications, 7, 41 (2011)
[9] Petersen A., Müller H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics, Ann. Statist. 44(1), 183-218, (2016)
[10] Yeung Ka Yee, Ruzzo W. L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17 (9), 763-774 (2001)


DRAFT

# Dynamic learning from data streams through the combined use of probability density functions and simplicial functional principal component analysis

Francesca Fortuna[a], Fabrizio Maturo[b], and Tonio Di Battista[d]

[a]Roma Tre University, Rome; francesca.fortuna@uniroma3.it
[b]Universitas Mercatorum, Rome; fabrizio.maturo@unimercatorum.it
[d]G. d'Annunzio University, Pescara; tonio.dibattista@unich.it

DRAFT

**Abstract**

This article deals with the challenges of processing and analyzing data streams due to their high speed and large volume, which makes traditional data mining and statistical techniques impractical. The study proposes a multi-step data stream dimensionality reduction procedure using functional data analysis and probability density functions' simplicial functional principal components decomposition in separate time windows to build the so-called simplicial functional classification trees. The simplicial functional classification trees are trained using the scores of the decomposition, and the weighted ensemble's majority vote gives the class labels' final prediction over the whole time domain. The goal is to create a classifier capable of extracting information from data streams and systematically updating the classification rule as new data becomes available without the need to store past data.

***Keywords:*** probability density functions; simplicial functional classification trees; functional principal components; functional data analysis.

## 1. Introduction

Recent advances in computing technology have enabled the collection of a vast amount of data, which arrives continuously as streams. Data streams are realisations of huge amounts of information that are fast, continuous, mutable, ordered, and unbounded. For this reason, data streams need to be processed and analysed once they arrive. Examples of data streams can be found in several fields, such as sensor networks, mobile data collection platforms, traffic management, and bank transactions.
The main problems in mining data streams are the high speed and large volume of the arriving information, and thus it is either unnecessary or impractical to store them in some forms. As a consequence, data streams lead to the curse of dimensionality issues, which makes traditional data mining and statistical multivariate techniques inappropriate. Furthermore, every learned model should be updated continuously and rely more on the most recent data in the stream (17). Thus, research on this topic is lively in the statistical literature (19; 14).
Usually, data streams are processed by computing suitable summaries of the data and splitting the time domain into intervals. Recent studies have proposed using distributional data analysis techniques to deal with clustering data streams (19). Density estimation has also been considered for analysing this type of data (10) as probability density functions (pdfs) make it possible to reduce the dimensionality in the time domain by discovering additional information in specific time intervals. Many authors also have

proposed the use of the functional data analysis approach (FDA) (18) as a dimensional reduction technique able to provide additional insights on phenomena (6; 2; 5). The basic idea of the FDA approach is to treat the entire streams as single functional objects. In this context, FDA has been applied to deal with different practical problems, e.g. unsupervised and supervised classification, outlier detection and forecasting.

Several functional classifiers have been proposed concerning supervised classification, which is the area of interest of this article. The latter offers promising tools to reduce the dimensionality of phenomena and improve accuracy by combining FDA and tree-based methods (9; 13; 15). The limit of these classifiers is that they immediately become obsolete because data is constantly evolving. This research proposes a methodological tool to solve this problem by the joint use of functional classifiers and pdfs. Specifically, a multi-step data stream dimensionality reduction procedure is suggested using the functional data representation. In the first step, data streams are divided into non-overlapping time-based windows and for each a pdf is estimated to capture the main streams' characteristics. Since pdfs are treated as functional data, the Simplicial Functional Principal Component (SFPC) decomposition is used in the second step. In the third step, simplicial functional classification trees (SFCTs) are built for each time window. Functional classification trees (13) are trained using the scores of the SFPC decomposition leading to an ensemble of classifiers. The final prediction of the class label for each stream is given by a weighted majority vote of the ensemble.

The ultimate goal is to create a classifier capable of extracting information from data streams and systematically updating itself as new data becomes available. In particular, a novel algorithm is proposed to update the classification rule by training new SFCTs as new information are available.

The remainder of the paper is the following: the proposed algorithm is presented in Sect. 2. A simulation study is shown in Sect. 3. Conclusions are provided in Sect. 4.

DRAFT

## 2. Materials and Methods

Let $S_i = \{x_{i1}, x_{i2}, ..., x_{it}, ..., x_{iT}\}$ be the $i$-th observed data stream, $i = 1, ..., n$, whose generic element $x_{it}$, $t = 1, ..., T$, is a real-valued scalar observed at time $t$. $S_i$ can be partitioned into a sequence of $J$ equally spaced time-based windows of length $q$ as follows:

$$W_j = \{x_{q\times(j-1)+1}, .., x_t, ....., x_{q\times j}\}, \quad j = 1, ..., J. \tag{1}$$

For each $j$-th window, a probability density function, $\widehat{f}_{ij}(x)$, can be estimated to summarize the behaviour of the $i$-th stream in the time interval $j$. Most existing approaches for estimating the density of data streams are based on the Kernel Density Estimation (KDE) method (17).

A natural way to address density functions is FDA. Most methods in FDA implicitly assume that the data objects can be embedded in the $L_2$ space, which can be used as embedding for unconstrained data. However, the $L_2$ space becomes meaningless in the presence of density data (3). Indeed, pdfs represent a special case of functional data because they are constrained functions (18):

$$f(x) \geq 0, \quad \int_{\mathcal{I}} f(x)\, dx = 1 \Rightarrow f(x) \notin L^2. \tag{2}$$

where $\mathcal{I} \subset R$ is a closed interval with log in $L_2$.

Thus, features of pdfs are accounted for in the Bayes spaces, $B^2$, which generalizes the Aitchison geometry for compositional data (1) to the functional setting. Hence, pdfs can be interpreted as functional compositional data, i.e. parts of some whole, which carry only relative information (3; 4; 11; 16). $\mathcal{B}^2$ is the space of (equivalent classes) of positive functions integrating to a constant, with square-integrable logarithm:

$$B^2 = \left\{ f : f > 0, \int_{\mathcal{I}} f(x)\, dx = c, \, log(f) \in L^2 \right\}. \tag{3}$$

An isometric isomorphism between $B^2$ and $L^2$ is defined by the Centered log-ratio (clr) transformation:

$$clr(f)(x) = f_c(x) = log\, f(x) - \frac{1}{\eta}\int_I log\, f(s)\, ds; \tag{4}$$

where $\eta$ stands for the length of the interval pdf support $I$; and, by construction, clr-transformed data have zero integral.

To reduce the dimensionality of a set of pdfs, we consider the Simplicial Functional Principal Component (SFPC) decomposition (11), which reformulates FPC decomposition in terms of Bayes spaces. Starting from a sample of $n$ centred pdfs:

$$\tilde{f}_{ij}(x) = \hat{f}_{ij}(x) \ominus \bar{f}_j(x); \quad \bar{f}_j(x) = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \hat{f}_{ij}(x),$$

For a specific $j$-window, SFPCs capture the main modes of variability of the densities, finding the SFPCs, $\zeta_k \in B^2(I)$, maximizing the following objective function over $\zeta \in B^2(I)$:

$$\frac{1}{n}\sum_{i=1}^{n}\langle \tilde{f}_i; \zeta\rangle^2_{B^2}, \quad subject\, to \quad \|\zeta\| = 1;\ \langle \zeta_k, \zeta_{k'}\rangle_{B^2} = 0,\ k < k'. \tag{5}$$

Hron et al. (2016) proved that the maximization problem in (5) can be performed by solving an equivalent FPC decomposition in $L^2$, by considering the clr-transformed pdfs.

$$\frac{1}{n}\sum_{i=1}^{n}\langle clr(\tilde{f}_i); \xi\rangle^2_{L^2},\ subject\, to\, \|\xi\|_{L^2} = 1;\ \langle \xi_k, \xi_{k'}\rangle_{L^2} = 0,\ k < k'; \quad \int_I \xi_k = 0. \tag{6}$$

Eq. (6) is solved by the eigenfunctions, $\xi_k$ of the sample covariance operator of the clr-transformed data, $V_{clr}$, obtaining the FPCs $\xi_k$, that are linked to the SFPCs by the relation:

$$\zeta_k = clr^{-1}(\xi_k); \tag{7}$$

and the scores, $v_{ik}$, that are equivalent to the scores obtained in the SFPC decomposition:

$$v_{ik} = \int_{\mathcal{I}} f_c(x)\xi_k(x)\, dx.$$

In the functional supervised classification context, the starting point is a training set of functions whose labels are known a-priori. Hence, we consider a functional data-set of the following form: $\{y_i, f_{ij}(x)\}$, where $f_{ij}(x)$ is a predictor pdf, and $y_i$ is the scalar response value observed at sample $i = 1, ..., n$, which could be either numeric or categorical, and that, in this context, is assumed to be constant over time (so, it does not depend on $j$). For simplicity, we will consider the case where $y$ is a binary variable. The known labels of the functions are used to train a functional classifier, which is a classification rule that can be exploited to predict the class labels of new functional observations In the literature, several classifiers have been proposed for functional supervised classification, e.g. Logistic Classifier, k-Nearest Neighbor Classifier, Maximum Depth Classifier, and Kernel Classifier (8; 12; 14; 15).

The basic idea of this paper is to create a functional classification rule that can be updated when new data are available. In particular, our proposal is to introduce the so-called Simplicial Functional Classification Trees (SFCTs), which extends the Functional Classification Trees (FCTs) proposed by (13; 14) to the context of pdfs by using the scores of SFPCs computed on the clr-transformed pdfs. Therefore, the coefficients of SFPC decomposition are used as new features to predict the response.

The proposed algorithm is composed by the following steps:

1. Given $S_i = \{x_{i1}, x_{i2}, ..., x_{it}, ...., x_{iT}\}$, create $J$ equally spaced time-based windows of length $q$;
2. For each $W_j$ window ($j = 1, ..., J$), convert raw data into pdfs;
3. Compute the clr transformation and implement the SFPCs using the clr transformation for the $j$-th time window;

DRAFT

4. Use the scores of the $j$-th SFPCs to build the $j$-th SFCT for the $j$-th time window;
5. Delete the original data and pdf and store only the $j$-th SFCT;
6. Predict the curves' labels using a weighted majority vote from the ensemble, where the weights give more importance to the most recent SFCTs. Thus, the posterior probability for each class is computed as follows:

$$P(\hat{y}_i = G_c \mid W_1, W_2, ..., W_J) = \frac{\sum_{j=1}^{J} I_j(y_i = G_c) w_j}{J},$$

where $G_c$ represents the possible labels of $Y$, i.e. the groups to predict, and the weigths are computed as follows:

$$w_j = (1/J)^{(1/j)}$$

7. Evaluate, the performance of the functional classifiers.
8. After observing $q$ new time observations, repeat steps 2-7 for each new time window $W_{j'}$, with $j'$ $\in$[J+1,+$\infty$[, and update the functional classification rule updating the ensembles.

## 3. Simulation study

Data streams are simulated by considering $n = 200$ observations, for $T = 600$ instants of time (days), with $n_1 = 100$ observations with label 1 (red) and $n_2 = 100$ observations with label 2 (black). We assume that the class membership does not change over time. As a consequence, future observations that are added are used to refine the predictive power of the classifier based on the new temporal information. The simulation scheme for functional data belonging to two classes is the same as used in (14), following scenario 3. The charts of the first row of Fig. 1 show how data flowing over time are split over time into windows of the same widths. Once the width of the window has been fixed, each new set of data is aggregated with those of the previous intervals. The second row of Fig. 1 shows the second step of the algorithm, where the time data of each series are converted into pdfs, individually. The third row of Fig. 1 highlights the correspondence between the original data in each window and the clr transformations of each interval. The charts in the last row of Fig. 1 illustrate the forth step of the algorithm, i.e. SFCTs are computed separately for each window and pruned using the cost-complexity pruning.

Table 1 illustrates the results of the SFCTs of each single window and the proposed algorithm. As we can see, SFCTs performance improves as new data is added, and the final classification rule $W_{all}$, which is based on the time-weighted majority vote in the ensemble of SFTCs, proves to significantly refine the results. Indeed, the proposed algorithm has the power of the ensemble methods in terms of variability reduction and considers the peculiarities of the functions in single windows by capturing the distinctive features of the original classes in detail and specific time intervals.

| $W_j$ | Accuracy |
|---|---|
| $W_1$ | 0.81 |
| $W_2$ | 0.83 |
| $W_3$ | 0.86 |
| $W_4$ | 0.86 |
| $W_5$ | 0.87 |
| $W_{all}$ | 0.94 |

Table 1: Results of the SFCTs for each window and using the proposed weighted ensemble $W_{all}$.

DRAFT

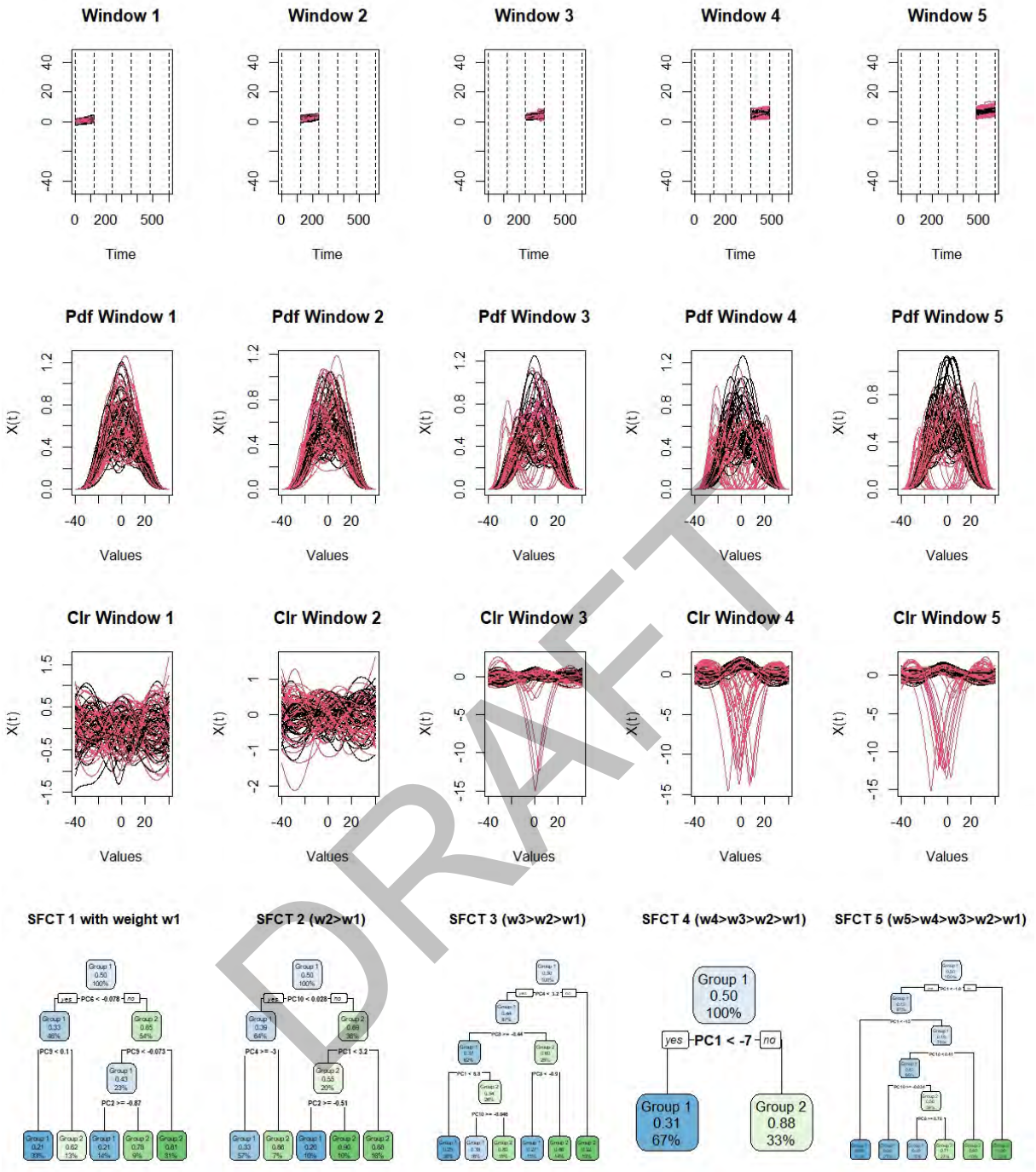



Figure 1: The visual explanation of the proposed algorithm.

## 4. Conclusions

In this paper, we have proposed a new algorithm to classify data streams via a dynamic classifier based on the joint use of FDA and density functions. Our approach is based on the so-called Simplicial Functional Classification Trees (SFCTs) which extends Classification Trees (CTs) to the context of pdfs viewed as functional data. The accuracy of the functional classifier is evaluated via bootstrap but different methods can be used, such as via cross-validation or a functional test set as in (14). The main objective of this study is to propose a classification method for high-dimensional data that achieves excellent levels of precision and is able to eliminate past datastreams by keeping only the necessary information of the

DRAFT

classification rule. In this way, the functional classification rule can be updated as new data arrives without the need to keep the huge amounts of data that are difficult to store due to their size.

## References


[1] Aitchison, J.: The Statistical Analysis of Compositional Data. In: Monographs on Statistics and Applied Probability, Chapman and Hall Ltd., London, UK (1986)

[2] Aguilera-Morillo, M., Aguilera, A., Escabias, M., Valderrama, MJ.: Penalized spline approaches for functional logit regression. Test. **22**(2), 251- 277 (2012)

[3] Delicado, P.: Dimensionality reduction when data are density functions. Computational Statistics & Data Analysis, **55**, 401–420 (2011)

[4] Egozcue, J., Diaz-Barrero, J., Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, English Series, **22**, 1175–1182 (2006)

[5] Febrero-Bande, M, de la Fuente, MO.: Statistical computing in functional data analysis: the R package fda.usc. J Stat Softw **5**(4): 1–28 (2012)

[6] Ferraty, F, Vieu, P.: Nonparametric Functional Data Analysis. New York, NY: Springer (2006)

[7] Fortuna, F, Maturo, F, Di Battista, T.: Clustering functional data streams: Unsupervised classification of soccer top players based on Google trends. Quality and Reliability Engineering International. **34**(7), 1448–60 (2018)

[8] Garcia, M.L.L., Garcia-Rodenas, R., Gomez, A.G.: K-means algorithms for functional data. Neurocomputing, **151**, 231-245 (2015)

[9] Gregorutti, B., Michel, B., Saint-Pierre, P.: Grouped variable importance with random forests and application to multiple functional data analysis. Comput Stat Data Anal. **90**, 15–35 (2015)

[10] Heinz, C., Seeger, B.: Cluster kernels: Resource-aware kernel density estimators over streaming data. TKDE **20**, 880–893 (2008)

[11] Hron, K., Menafoglio, A., Templ, M., Hruzova, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis, **94**, 330–350 (2016)

[12] Jacques, J., Preda, C.: Functional data clustering: a survey. Advances in Data Analysis and Classification **8**(3), 231-255 (2013)

[13] Maturo, F., Verde, R.: Pooling random forest and functional data analysis for biomedical signals supervised classification: Theory and application to electrocardiogram data. Statistics in Medicine. **41**(12), 2247–2275 (2022). doi: 10.1002/sim.9353

[14] Maturo, F., Verde, R.: Supervised classification of curves via a combined use of functional data analysis and tree-based methods. Computational Statistics (2023). **38**, 419–459. doi: 10.1007/s00180-022-01236-1

[15] Maturo, F., Verde, R.: Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. Computational Statistics (2022). doi: 10.1007/s00180-022-01259-8

[16] Menafoglio, A., Secchi, P., Guadagnini, A.: A class-kriging predictor for functional compositions with application to particle-size curves in heterogeneous aquifers. Mathematical Geosciences **48**, 463-485 (2014)

[17] Qahtan, A., Wang, S., Zhang, X.: Efficient Estimation of Dynamic Density Functions with Applications in Data Streams (2018) 10.1007/978-3-319-89803-11

[18] Ramsay, J, Silverman, B.: Functional Data Analysis. 2nd ed. New York, NY: Springer (2005)

[19] Verde, R., Irpino, A., Balzanella, A.: Dimension Reduction Techniques for Distributional Symbolic Data. IEEE Trans Cybern. **46 (2)**, 344–355 (2016)

[20] Zhou, A., Cai, Z., Wei, L., Qian, W.: M-kernel merging: towards density estimation over data streams. DASFAA 2003. 285-292 (2003)


DRAFT

# Social networks and loneliness among older migrants in Italy

Viviana Amati[a], Eralba Cela[b], and Elisa Barbiano di Belgiojoso[a]

[a] University of Milano-Bicocca; viviana.amati@unimib.it,elisa.barbiano@unimib.it
[b] University of Milan; eralba.cela@unimi.it

**Abstract**

Ageing and migration are two of the major demographic trends in Western societies. Their intersection generates a considerable diversity of older migrants' groups that range from the more privileged international retirement migrants to the more disadvantaged refugees and 'zero generation' migrants. Different groups present wide variations in their needs and vulnerabilities. Nonetheless, older migrants are still an under researched group mainly because scholars' and policy-makers' attention has focused primarily on issues of border control, integration and the second generation, and the myth of return that projects migrants re-settled in their home countries. Empirical evidence shows however that retirement may trigger return migration. Still, most ageing labour migrants in Western countries stay put or adopt pendular strategies, travelling back and forth. In this paper we focus on the overlooked topic of loneliness among older migrants residing in Italy. We rely on a unique survey carried out by ISTAT during 2011–2012, specifically addressing migrant population.

**Keywords:** Loneliness, older migrants, social networks, Italy

DRAFT

## 1. Introduction

Loneliness has been defined as an unpleasant experience of a perceived mismatch between the available and desired number and quality of social relations. Unlike social isolation, which refers to the objective lack or the limited number of social relationships, loneliness represents a subjective feeling thatmight be present even when an individual is surrounded by an extensive network of family and friends or might be absent even if one is socially isolated.

Loneliness is not an individual problem only. It has multiple societal implications because of substantial negative consequences on physical and mental health and is associated with a lower quality of life.

Thus far, most of the research on loneliness among older adults is concentrated on the native population. In contrast, individuals with a migratory background have not been looked at extensively, although there has been a consistent increase in recent years (Fokkema & Ciobanu 2021). In the Italian context, the issue of loneliness among migrants (and even older ones) is rarely analysed. With the few exceptions of the studies of Cela and Fokkema (2017) and Cela and Barbiano di Belgiojoso (2021), no other studies have been carried out so far on this topic.

In the present study, we aim to fill this critical gap by focusing on loneliness among older migrants in Italy and analysing the impact of different types of support networks, namely the instrumental/support and emotional networks, while controlling for standard socio-economic variables.

## 2. Determinants of loneliness

The few existing quantitative studies investigating loneliness among older migrants have shown that, on average, older migrants are more likely to feel lonely than native peers (e.g., de Jong Gierveld et al., 2015). The main factors that explain the higher loneliness prevalence are related to general and specific risk factors associated with the status of being a migrant.

Among the general risk factors, empirical studies have shown that several demographic characteristics, such as gender, ethnicity, and marital status, are important predictors of loneliness. Health is also a significant determinant as it enables or hampers participation in different social activities; being healthy enables individuals to engage in differentiated social activities with friends and in community life, associations, and religious activities, whereas poor health is associated with a high risk of loneliness. Related to this, empirical evidence shows that older migrants are more likely to experience (earlier) poor objective and subjective health than their native peers (Cela & Barbiano di Belgiojoso, 2021).

Examples of migrant-specific risk factors are family disruption, language and cultural barriers, length of permanence in the host country, lack of social support, discrimination, and hostility in the host country, to name a few.

Among these migrant-specific factors, language proficiency and migration duration are critical determinants of the heterogeneity and size of migrants' support networks (Djundeva & Ellwardt, 2020). Specifically, those migrants who have spent more time in the host country might be less prone to feel lonely. They are more proficient in the local language, have had more opportunities to establish their local networks and frequently have other family members living nearby (Cela & Fokkema, 2017).

Recent studies (de Jong Gierveld et al., 2018; Djundeva & Ellwardt, 2020) have investigated the role of social networks in protecting (or not) against loneliness. They found out that being engaged in narrowed and poor-quality social contacts regarding composition and functioning are essential predictors of loneliness. Although there is no ideal size for social networks, having heterogeneous networks and frequent contacts with others is usually considered a safety net, as different types of relations play different roles and may result in various meeting opportunities and support exchanges (Litwin & Stoeckel, 2013). Among the web of social networks, family ties have a key role; the presence of a partner is the most crucial protective factor, as the partner represents the primary source of emotional and practical support. In contrast, the condition of being a widow or divorced or living alone often lead to loneliness. Parents, siblings and kin are also sources of emotional and instrumental support, especially in later life and are, therefore, a protective factor against loneliness. Other sources of support and protection against loneliness are friends and co-workers. Empirical evidence has also shown that participating in volunteerism, community activities, and religious services protects from loneliness as these activities are usually associated with a higher level of (given and received) social support and integration (Luhmann & Hawkley, 2016).

DRAFT

## 3. Data and methods

We rely on unique data from the "Social Condition and Integration among Foreign Citizens" survey administered by the Italian National Institute of Statistics during 2011 - 2012 on a sample of 25,000 individuals living in a household with at least one foreign member. The selected sample addressed specifically migrants, thereby capturing the multifaceted nature of the migration process. We analysed the subsample of 3,104 (12.26%) migrants older than 50.

The survey collected information on traditional socio-economic variables, such as gender, age, migration duration, area of origin, education, and employment. Additional information concerns discrimination, perceived wealth and health, health condition, and difficulties with the Italian language (reading, writing, speaking, and listening). Information on loneliness was collected through the question: "Do you feel lonely?". In the sample, 5.2% (162) do not feel lonely, 14.1% (438) feel lonely a little, 27.4% (852) somewhat lonely, and 53.2% (1652) very much lonely.

Quite a few questions collected information on the social relationships of the migrants. We used this data to construct ego network typologies following the approach presented in Amati et al. (2015) and Pelle & Pappadà (2021). Ego networks are the set of relationships in which an individual (hereafter referred to as

Ego) is embedded. It comprises Ego and its relations with alters along with Ego and alters characteristics. The survey provides information on different sets of alters that include the partner/spouse, parents, children, relatives, friends, co-workers, people who helped Ego with childcare, housekeeping, or important matters and associations with Ego as a member. The granularity of the information varies according to the alter's role and cohabitation with Ego. Thus, the alters in the ego network are represented by alter's categories rather than the individuals belonging to those categories. Based on the data, we defined seven alter categories: partner/spouse, parents, children, relatives, friends, members of the place of worship and others. We established the presence of alters based on cohabitation, and the frequency of contact between the migrant and the alters to whom it is tied. Figure 1 represents the ego network as a graph where the nodes depict Ego and alters while a tie identifies the presence of a support relationship.

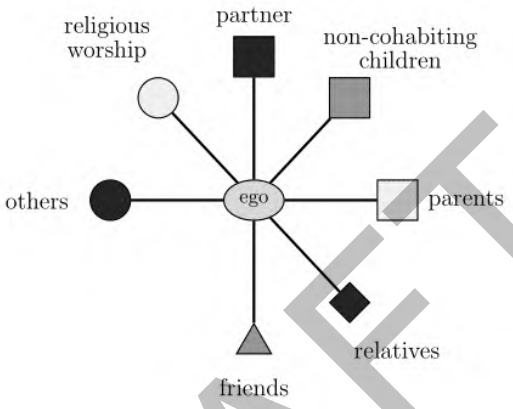



*Figure 1 – Ego network of a migrant. A squared node indicates that the alters belong to the immediate family, a diamond that they are part of the extend family, a triangle that are friends, and a circle that are members of the religious worship or others. The colour of the nodes is used to distinguish the role of the alter.*

DRAFT

We clustered the ego networks into groups to identify network typologies, ranging from the complete network where all the alters are present to sparse networks where only a few are present.

The network typologies were included as explanatory variables in a logistic regression model to make inferences on the effect of social networks on loneliness while controlling for the standard socio-economic variables described above. The dependent variable takes value 0 if migrants do not feel alone and 1 if they feel a little, somewhat, or very much lonely. While, in traditional analyses, the relationships in which Ego is embedded are treated as separate variables, the network typologies have the advantage of summarising those relationships to obtain a more parsimonious model and jointly analyse different combinations of alters.

We repeated the analysis using an adjacent category model to distinguish among the different categories of loneliness. The two models led to the same results.

## 4. Results

Seven different network typologies have been identified, ranging from the complete ego network, where all the alters are present, to the almost empty network comprising only children and friends. Table 1 reports the absolute and relative frequency distributions of the network typologies.

The network typologies are widespread among different groups of migrants and therefore are not specific to, e.g., females or the age of the migrants.

Table 2 reports the estimates of the model coefficients describing the association between feeling lonely and the network typologies. The coefficients are all statistically significant, but the one related to the ego network immediate family (no parents) + friend typology. Compared to the complete ego, we observe that

all the other network typologies have higher odds of feeling alone, as suggested by the positive estimates of the coefficients and the corresponding odds ratios greater than 1. Specifically, the odds of feeling alone are: 1.37 times (37%) larger for those with a Complete network (no parents), 1.52 times (52%) larger for those with an Immediate family + non-kin ego-network, 2.08 times (108%) larger for those with an Immediate family + friends; 2.01 times (101%) larger for those with an Extended family + non-kin groups; and 1.93 times (93%) larger for those with a Children + relatives ego-network.

Table 1 – Absolute and relative frequency distribution of the network typology.

| Network typology | n | % |
|---|---|---|
| Complete network (no others) | 399 | 12.9% |
| Complete network (no parents) | 487 | 15.7% |
| Immediate family + non-kin groups | 240 | 7.7% |
| Immediate family + friends | 351 | 11.3% |
| Immediate family (no parents) + friends | 503 | 16.2% |
| Extended family + non-kin groups | 527 | 17.0% |
| Children + friends | 597 | 19.2% |

Table 2 – Model results for the network typologies: estimates (Est.), odds ratio (OR), p-value (p-value) and percentage change (%change)

| | Est. | OR | p-value | %change |
|---|---|---|---|---|
| Network typologies (ref. Complete network (no other)) | | | | |
| Complete network (no parents) | 0.31 | 1.37 | ** | 36.60 |
| Immediate family + non-kin groups | 0.42 | 1.52 | ** | 51.55 |
| Immediate family + friends | 0.73 | 2.08 | *** | 107.91 |
| Immediate family (no parents) + friends | 0.21 | 1.23 | | 22.99 |
| Extended family + non-kin groups | 0.70 | 2.01 | *** | 101.19 |
| Children + friends | 0.66 | 1.93 | *** | 92.70 |

We controlled the results for common factors such as gender, area of origin, education, economic and health conditions, knowledge of the Italian language and discrimination. Results are in line with previous research. We also considered interactions between the network structures, and gender and migration duration, respectively, but none of the interactions was significant.

The results of our study indicate that there is an association between loneliness and network typologies. However, loneliness might also affect the embeddedness in a network since it might lead to isolation from others. Applying models accounting for endogeneity is the next step of this research.

DRAFT

## References

[1] Amati, V., Rivellini, G., & Zaccarin, S. Potential and effective support networks of young Italian adults. Social Indicators Research, *122*(3), 807-831 (2015).

[2] Cela, E. & Barbiano di Belgiojoso, E. Ageing in a foreign country: determinants of self-rated health among older migrants in Italy, J. Ethnic Mig. Stud., 47:15, 3677-3699, (2021). doi: 10.1080/1369183X.2019.1627863.

[3] Cela, E., & Fokkema, T. Being lonely later in life: A qualitative study among Albanians and Moroccans in Italy. Ageing & Society, *37*(6), 1197–1226. (2017). Doi: 10.1017/S0144686X16000209.

[4] de Jong, G.J., Van der Pas, S., Keating, N. Loneliness of older immigrant groups in Canada: effects of ethnic-cultural background. J. Cross-Cult. Gerontol. 30:251–268 (2015) doi:10.1007/ s10823-015- 9265-x
[5] de Jong Gierveld, J., van Tilburg, T.G., Dykstra, P.A. New ways of theorizing and conducting research in the field of loneliness and social isolation. In: Vangelisti AL, Perlman D (eds) The Cambridge handbook of personal relationships. Cambridge University Press, Cambridge, pp 391–404 (2018)
[6] Djundeva, M., Dykstra, P. A., Fokkema, T. Is Living Alone "Aging Alone"? Solitary Living, Network Types, and Well-Being. J. Gerontol.: Soc. Sciences, gby119 (2018). doi:10.1093/geronb/gby119.
[7] Fokkema, T., Ciobanu, R.O. Older migrants and loneliness: scanning the field and looking forward. Eur. J. Ageing 18, 291–297. (2021). Doi: 10.1007/s10433-021-00646-2
[8] Litwin, H., Stoeckel, J. Confidant Network Types and Well-Being among Older Europeans. The Gerontologist, 54 (5): 762–772 (2013)
[9] Luchetti, M., Lee, J. H., Aschwanden, D., Sesker, A., Strickhouser, J. E., Terracciano, A., Sutin, A. R. (2020). The trajectory of loneliness in response to COVID-19. American Psychologist. doi: 10.1037/amp0000690
[10] Luhmannm, M., Hawkley, L.C. Age differnces in loneliness from late adolescence to oldest old age. Developmental Psychology, 52: 943-959 (2016)
[11] Pelle, E., Pappadà, R. A clustering procedure for mixed-type data to explore ego network typologies: an application to elderly people living alone in Italy. Statistical Methods & Applications, 30(5), 1507-1533 (2021)

DRAFT

# The Italian Decree on Security: An Analysis of the Impact on Asylum Applications

Giorgio Piccitto
Department of Political and Social Sciences - University of Bologna - - Strada Maggiore 45, Bologna - giorgio.piccitto3@unibo.it

**Keywords:** AsylumAapplications; Decree on Security; Asylum Recognitions;

DRAFT

## 1. Introduction

Despite the existence of a normative space as the European Union (EU), the rates of recognition for asylum seekers' protection statuses remarkably vary across different countries (Van Wolleghem, Sicakkan, 2022). Several studies have attempted to examine the potential explanations of these different between-countries rates of recognition, considering both economics and political reasons, without finding any clear driver (Toshkov, 2014), up to the point that it has been commented that «asylum decisions in Western Europe are highly arbitrary» (Bronkhorst, 1991, p.151). At the same time, it has been argued that changes in asylum policies within a given country may be associated with changes in recognition rates on asylum application flows (Thielemann, 2003), modifying the level of deterrence characterizing a determined socio-economic context (Holzer et al., 2000).
On December 3rd 2018, the Italian government approved the so-called "Immigration and security" decree. Among its intentions, regarding a wide range of matters (contrast to mafia and terrorism, urban security), the decree remarkably changed the regulation of asylum, immigration and citizenship (De Petris, 2019).
This contribution aims to evaluate the association between the "Immigration and security" decree and recognition rates of asylum applicants in Italy and if and how this association has been conditional on the applicant's socio-demographic characteristics. This goal will be accomplished by performing some statistical analyses on a dataset created from the annual Eurostat database on the first instances decisions on applications.

## 2. Data and method

The data used for this work are recorded by Eurostat at an aggregated level, considering the outcomes of first instances decisions on asylum applications in Italy by applicant's citizenship, age and sex. Age is recorded in the following classes: 1 = 0-13 years old (19.3% of the sample); 2 = 14-17 years old (11.2% of the sample); 3 = 18-34 years old

(39.0% of the sample); 4 = 35 years old or more (30.5% of the sample). Sex is recorded as follows: 1 = female (40.8% of the sample); 1 = male (59.2% of the sample). The dependent variable is a dummy, indicating a positive decision (1) or a rejection(0). We consider all the applications in the period from 2008 to 2021 (N= 667,065). For each combination of the above-mentioned categories, the aggregated database of Eurostat provides the number of applicants rounded up to the nearest 5. As an example, female citizens of Afghanistan aged 0-13 who received a positive response to their asylum application in 2009 were 5. We used this information on the number of applicants per each combination of categories to weight our sample: by doing so, we were able to create a dataset containing information at the micro-level.
In order to answer our research questions, we perform a set of logit models as follows:

$$\textbf{M1: } Y = \beta_0 + \beta_1(Year) + \beta_2(Sex_i) + \beta_3(Age_i) + \beta_4(Citizenship_i) + \varepsilon_i$$

$$\textbf{M2: } Y = \beta_0 + \beta_1(Year\ X\ Sex_i) + \beta_2(Age_i) + \beta_3(Citizenship_i) + \varepsilon_i$$

$$\textbf{M3: } Y = \beta_0 + \beta_1(Year\ X\ Age_i) + \beta_2(Sex_i) + \beta_3(Citizenship_i) + \varepsilon_i$$

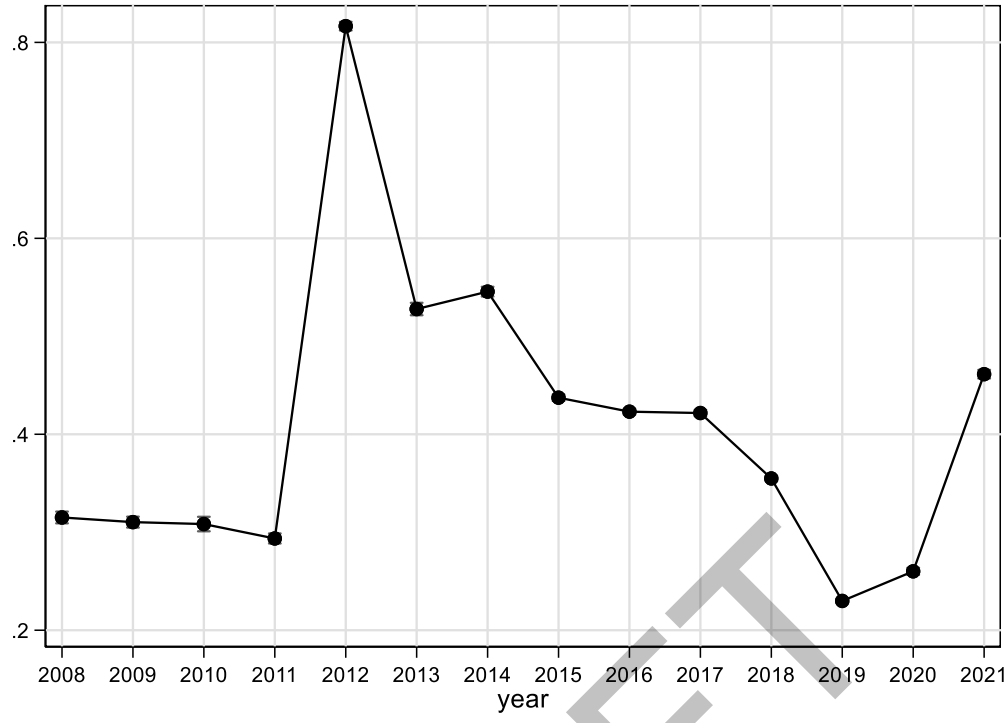
Model 1 estimates the association between year and rates of recognition of asylum applicants in Italy, net of the individual socio-demographic controls. With this model we are able to shed light on whether the "Immigration and security" decree has changed the acceptance rate of asylum requests in Italy. Model 2 and model 3 allows the conditionality on sex (model 2) and age (model 3) of the association between year and rates of recognition of asylum applicants, by including an interaction term between year and sex (model 2) and year and age (model 3). All models control for individual citizenship. Results are presented and commented in terms of predicted probabilities, since the interpretation of logit coefficients is not straightforward (Mood, 2010).

DRAFT

## 3. Main results

In figure 1 are shown the predicted probabilities of the likelihood of asylum recognition by year, as estimated in Model 1.

**Figure 1.** *Logit model on the likelihood of asylum recognition by year. Model 1. Predicted probabilities. The model controls for sex, age, and citizenship*.




The analysis of this figure permits evaluating the trend of recognition of asylum applicants, with particular attention to the years following the implementation of the "Immigration and security" decree in 2018. The results seem to underline the decree's important role in driving the acceptance rate of asylum recognition. Indeed, after an apex of recognition in 2012[1], when more than 80% of asylum requests were accepted, the trend started to assume a downward shape, and the recognition rate began to drop. Nevertheless, the most substantial drop (after that between 2012 and 2013) has been recorded from 2018 (the year of the application of the "Immigration and security" decree) to 2019, passing from the 35.5% to the 23.0% of acceptances. In this sense, the legislation has remarkably impacted our outcome of interest.

DRAFT

**Figure 2.** *Logit model on the likelihood of asylum recognition by year and sex. Model 2. Predicted probabilities. The model controls for age and citizenship.*

[1] The strong increase recorded in 2012 is due to the response of the Italian Government to the unrest in North Africa in 2011. On April 5th, 2011, recognizing the exceptional situation of North Africa, the Government adopted temporary measures of humanitarian protection in favor of refugees from North Africa ("*Emergenza Nord Africa*", ENA). The ENA provisions temporarily relaxed immigration policies: migrants who fled from Algeria, Egypt, Libya, Morocco and Tunisia to Italy between January 1st and April 5th, 2011 were automatically granted a temporary permit of stay for humanitarian reasons. On August 3rd, 2011, the duration of this regime was extended due to the persistent situation of instability in North Africa. Some specific emergency measures were lately applied also to migrants the Horn of Africa, Kenya, Sudan and Uganda (dalla Pellegrina et al., 2014).

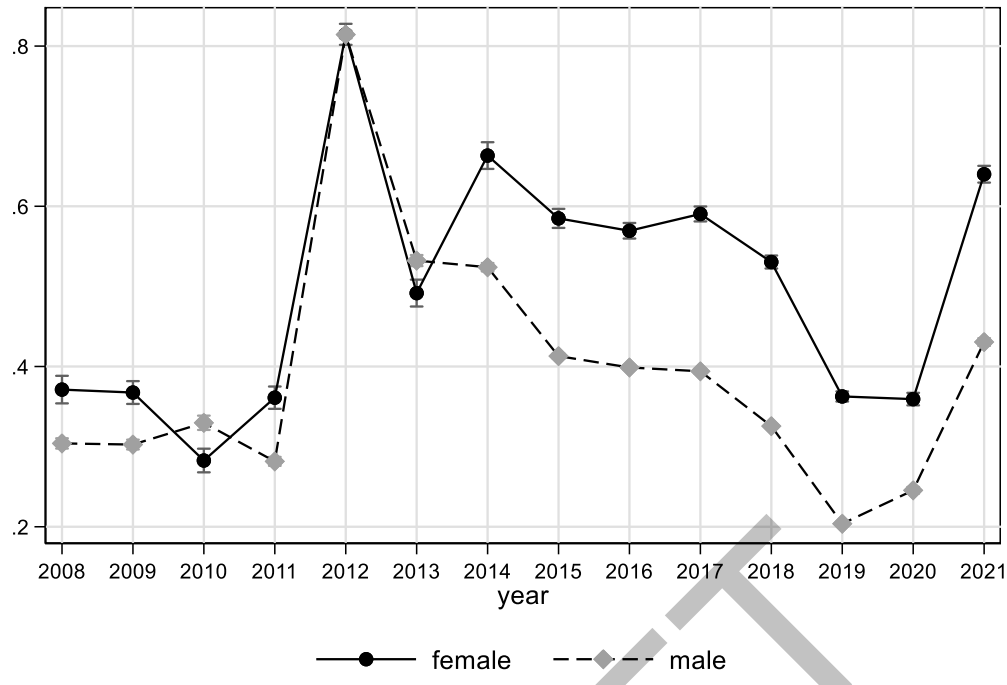
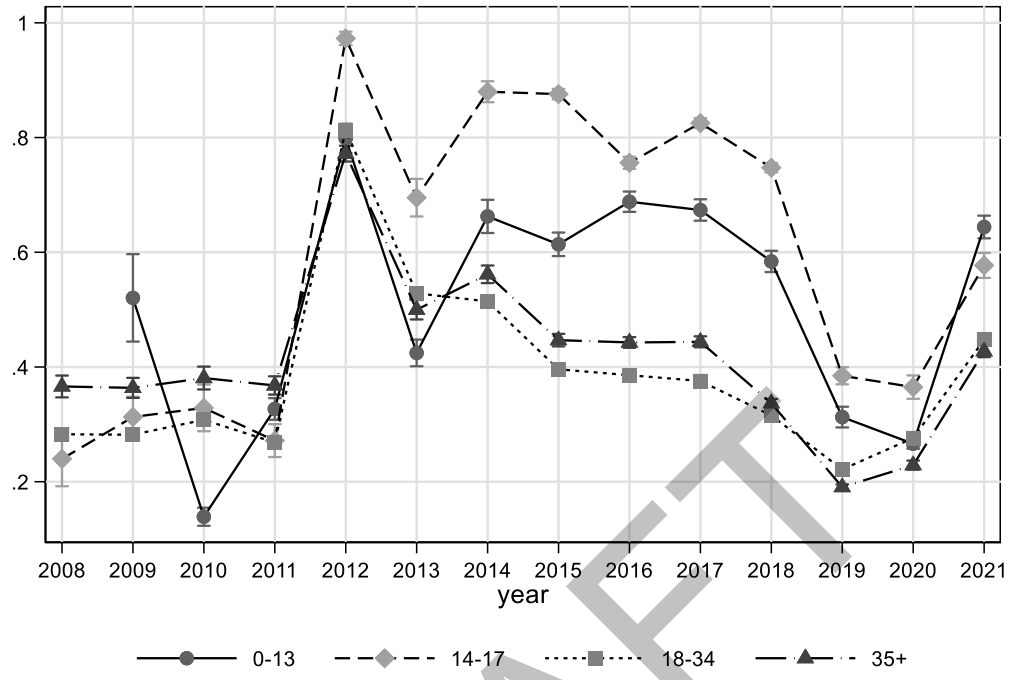



Figure 2 shows the predicted probabilities of asylum recognition by year and sex, as estimated in Model 2. This graph allows to evaluate if the association between year and asylum recognition was conditional on the applicant's sex. Results show that up to 2013, the rates of recognition of asylum applicants were very similar between men and women, with the highest difference between male and female applicants being recorded in 2011 (average marginal effect = .08). But from 2013 onwards, it emerges a gender gap in the asylum recognition, with female applicants more likely to have their request accepted, a result already observed in other European countries (Plümper, Neumayer, 2021). Such a higher likelihood is also maintained in the aftermath of the application of the decree, which lowers the acceptance at a similar pace both for men and for women.

Figure 3 shows the predicted probabilities of the likelihood of asylum recognition by year and age, as estimated in Model 3. Analyzing these predicted probabilities makes it possible to shed light on the interplay between year and applicant's age on asylum recognition. Differently from what emerged concerning the applicant's sex, the pattern of asylum acceptance by age does not seem to vary during the period considered in our analysis. Indeed, the highest likelihood of acceptance is observed for younger applicants, particularly those between 14 and 17 years old. This holds true from 2012 onwards, while before, the differences across different ages were less marked. On May 6th 2017, new legislation regarding the "Protection Measures for Unaccompanied Minors" (law n. 47/17, "*legge Zampa*") entered into force in Italy. This law has filled significant gaps in the protection of unaccompanied children, introducing important provisions, among other issues, on age assessment procedures. However, our results suggest that higher acceptance rates were observed before the law's implementation, which resulted from more than three years of advocacy efforts by Save the Children and other NGOs (Rozzi, 2017). In general, the older cohorts experience a lower likelihood of asylum request acceptance with respect to the younger ones, and this pattern is confirmed also in the aftermath of the "Immigration and security" decree of 2018.

DRAFT

**Figure 3.** *Logit model on the likelihood of asylum recognition by year and age. Model 3. Predicted probabilities. The model controls for sex, and citizenship.*




DRAFT

## 4. Conclusion

The "Immigration and security" decree has lowered the probability of asylum request acceptance in Italy. Although the pattern showed a downward trend from 2012, the drop after 2018, the decree's implementation year, was particularly remarkable, and the recognitions passed from 35.5% to 23.0% in 2019. Interestingly, from 2013 onwards, it emerged a gender gap in asylum recognition, with women having a higher likelihood of request acceptance than men; this gender-based difference has also been confirmed after the decree of 2018. Finally, our analysis showed that older cohorts have always experienced a lower acceptance rate of their requests than younger ones.

## References

Bronkhorst D. (1991) The 'realism' of a European asylum policy: A quantitative approach. Netherlands, *Quarterly on Human Rights*, 9, 142-58.

dalla Pellegrina L., Saraceno M., Suardi M. (2014) Migration policy: An assessment on the North Africa emergency provisions, *Procedia Economics and Finance*, 17, 156-164, doi: 10.1016/S2212-5671(14)00890-9.

De Petris, A. (2019) Pursuing Public Insecurity? The New Italian Decree on "Immigration and Security", *The Review of European Affairs*, 3(1), 53-81.

Holzer T., Schneider G., Widmer T. (2000) The impact of legislative deterrence measures on the number of asylum applications in Switzerland (1986–1995), *International Migration Review*, 34(4), 1182–1216.
Mood C. (2010) Logistic regression: why we cannot do what we think we can do, and what we can do about it, *European Sociological Review*, 26, 67–82.
Plümper T., Neumayer E. (2021) Human rights violations and the gender gap in asylum recognition rates, *Journal of European Public Policy*, 28(11), 1807-1826.
Rozzi E. (2017) The new Italian law on unaccompanied minors: a model for the EU? *EU Immigration and Asylum Law and Policy Blog* https://eumigrationlawblog.eu/
Thielemann E.R. (2003) Does policy matter? On governments' attempts to control unwanted migration, *IIIS Discussion Paper*, 9, 1–39.
Toshkov D.D. (2014) The dynamic relationship between asylum applications and recognition rates in Europe (1987–2010), *European Union Politics*, 15(2), 192–214.
Van Wolleghem P. G., Sicakkan H. G. (2022) Asylum seekers in the machinery of the state: administrative capacity vs. preferences. Recognition rates in EU member states. *European Union Politics*, 14651165221135113.

DRAFT

# Adaptive combinations of tail-risk forecasts

Alessandra Amendola[a], Vincenzo Candila[a], Antonio Naimoli[a], and Giuseppe Storti[a]

[a] Department of Economics and Statistics, University of Salerno, Italy, alamendola@unisa.it, vcandila@unisa.it, anaimoli@unisa.it, storti@unisa.it

**Abstract**

The continuous evolution of financial markets highlights how quantitative financial risk management has become a key tool in investment decisions, capital allocation, and regulation. Although several methods have been proposed to estimate the risk of an investment in capital markets, Value-at-Risk (VaR) and Expected Shortfall (ES) can be considered the standard measures of market risk, as they are used both for internal control of financial institutions and for regulatory purposes. In this direction, the choices of modelling and estimation methods for VaR and ES play a critical role. Nowadays, a variety of possibilities is available. For instance, there are models belonging to the class of parametric, semi-parametric, and non-parametric methods. Moreover, among the class of parametric models, there are several error distributions that could be considered. Also, some models allow for the use of variables mixed at different frequencies. To mitigate the impact of these sources of uncertainty, we propose a forecast combination strategy by adaptively weighting the pool of most accurate predictors based on the Model Confidence Set (MCS) results. The empirical analysis suggests that combinations of VaR and ES forecasts lead to higher predictive accuracy over a wide range of competitors.

***Keywords:*** Value-at-Risk, Expected Shortfall, Model Confidence Set, Volatility

DRAFT

## 1. Introduction

The past financial crisis has emphasized the importance for financial institutions to rely on reliable forecasts of Value-at-Risk (VaR) and Expected Shortfall (ES). In light of the Basel Capital Accords, VaR and ES can be considered the leading measures of tail risk forecasting.

The VaR, being a conditional quantile in the lower tail of a portfolio's return distribution, has been widely used as a measure of financial market risk for both regulatory and internal risk management purposes. It quantifies the maximum amount of loss for a portfolio of assets, under normal market conditions, over a given time period and at a certain confidence level. However, VaR has been subject to criticism because it fails to meet the requirements of a coherent risk metric (5), and at the same time has the limitation of not providing information on potential exceedances beyond the quantile.

Recently, the ES has received increasing attention as an alternative tail risk metric and is now recommended by the Basel Committee on Banking Supervision. The ES is a coherent risk metric (5) and provides the expected loss conditional on returns exceeding the VaR threshold (1).

However, despite its attractive properties, in contrast to VaR, the ES is not an "elicitable" measure, meaning that there is no loss function such that the ES is the solution that minimizes the expected loss. Consequently, this poses a challenge for ES estimation and backtesting (18). As a partial remedy to this problem, (15) define a set of joint loss functions, named $FZ$, for VaR and ES for which these two measures are jointly elicitable.

VaR and ES forecasts can be obtained through several approaches that are substantially attributable to parametric, non-parametric, and semi-parametric models (22). Parametric approaches are often based on GARCH models, which require the specification of the conditional distribution of returns and volatility dynamics. Semi-parametric approaches, such as quantile regression models (14; 27), include those based on methods that instead require specific assumptions about the dynamics of risk, but no assumptions about the distribution of returns. Finally, a popular example of non-parametric approach is the historical simulation (21).

The previous discussion highlights how the identification of an optimal forecasting model is subject to data, parameter and model uncertainty. In this context, a possible solution is to resort to combining forecasts. The idea of combining forecasts generated by different models dates back to (7). Since then, many articles have emphasized the advantages of combining forecasts of volatility (3; 2), VaR (8) or VaR and ES (28; 25) as a way to reduce the risk of selecting a single predictor, which may not always be optimal during the forecast period.

Combination has the greatest potential when the individual methods use different information or use information in different ways. Among the whole universe of potential VaR and ES forecasts, the chosen set of candidate models covers a wide range of frequently used parametric, semi-parametric and non-parametric techniques, as well as methods based on intraday data and mixed frequency variables.

Aiming to reduce the impact of model uncertainty in tail-risk forecasting, a novel tail-risk forecasting strategy is proposed. The proposed approach is based on a truncated mean in which only models that perform significantly better than others are considered in the combination. To select the set of best-performing models to be involved in the combination, the Model Confidence Set (MCS) by (20) is used. The proposed procedure is adaptive because the composition and dimension of the set of "best" models involved in the combined forecast varies over time and and does not require parameter estimation.

The performance of the proposed combination approach is compared to each individual model in the universe. For forecast evaluation, we use backtesting via the Unconditional Coverage (UC) backtest by (23), the Conditional Coverage (CC) backtest by (12) and the dynamic quantile (DQ) backtest of (14), together with the Regression-Based Expected Shortfall Backtesting (BD) of (9). Furthermore, we compare the forecasts with the MCS of (20) to determine the approach that produces the most accurate predictions.

The results on the S&P500 index indicate that the proposed combined predictors enter the set of superior models in the MCS evaluation, while all backtesting procedures are passed at the 1% significance level.

The rest of the paper is organized as follows. Section 2. illustrates the methodology adopted and the proposed combined predictors. Section 3. is devoted to the empirical application.

DRAFT

## 2. Methodology

In this work, three different combined predictors for VaR and ES are proposed. All the combined predictors are based on the models which, dynamically, enter the Set of Superior Models (SSM) according to the MCS procedure. We defined as *training* MCS the procedure used to find the best models in the so-called training or in-sample period. The loss used in the *training* MCS belongs to the class of $FZ$. More in detail, the loss is that of (24) labelled as $FZ0$. Formally:

$$FZ0(VaR_{i,t}(\tau), ES_{i,t}(\tau), r_{i,t}) = \frac{1}{\tau ES_t(\tau)} \mathbb{1}_{(r_{i,t} \leq VaR_{i,t}(\tau))}(r_{i,t} - VaR_{i,t}(\tau)) + \frac{VaR_{i,t}(\tau)}{ES_{i,t}(\tau)} + \log(-ES_{i,t}(\tau)) - 1, \quad (1)$$

where $\tau$ is the coverage level chosen, $r_{i,t}$ is the log-returns for day $i$ of the period $t$ and $\mathbb{1}_{(\cdot)}$ is an indicator function.

The chosen set of candidate models covers a wide range of frequently used parametric, semi-parametric and non-parametric techniques, as well as methods based on intraday data and mixed frequency variables. Table 1 reports the set of candidate models.

Table 1: Candidate models

| Model | Functional form | Err. Distr. |
|---|---|---|
| GARCH–N, GARCH–t (10) | $r_{i,t}\|\mathcal{F}_{i-1,t} = \sqrt{h_{i,t}}\eta_{i,t}$<br>$h_{i,t} = \omega + \alpha r_{i-1,t}^2 + \beta h_{i-1,t}$ | $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\eta_{i,t} \overset{i.i.d}{\sim} t_\nu$ |
| GJR–N, GJR–t (17) | $r_{i,t}\|\mathcal{F}_{i-1,t} = \sqrt{h_{i,t}}\eta_{i,t}$<br>$h_{i,t} = \omega + \left(\alpha + \gamma \mathbb{1}_{(r_{i-1,t}<0)}\right) r_{i-1,t}^2 + \beta h_{i-1,t}$ | $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\eta_{i,t} \overset{i.i.d}{\sim} t_\nu$ |
| GMIDAS–N, GMIDAS–t (13) | $r_{i,t}\|\mathcal{F}_{i-1,t} = \sqrt{\pi_t \times \xi_{i,t}}\eta_{i,t}$<br>$\xi_{i,t} = (1-\alpha-\beta-\gamma/2) + \left(\alpha + \gamma \cdot \mathbb{1}_{(r_{i-1,t}<0)}\right) \frac{r_{i-1,t}^2}{\pi_t} + \beta\xi_{i-1,t}$<br>$\pi_t = \exp\left\{m + \zeta \sum_{k=1}^{K} \delta_k(\omega) MV_{t-k}\right\}$ | $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\eta_{i,t} \overset{i.i.d}{\sim} t_\nu$ |
| R–GARCH–N, R–GARCH–t (19) | $r_{i,t}\|\mathcal{F}_{i-1,t} = \sqrt{h_{i,t}}\eta_{i,t}$<br>$h_{i,t} = const + \beta h_{i-1,t} + \alpha x_{i-1,t}$<br>$x_{i,t} = const_x + \delta h_{i,t} + \tau_1 \eta_{i,t} + \tau_2\left(\eta_{i,t}^2 - 1\right) + \sigma_u u_{i,t}$ | $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\eta_{i,t} \overset{i.i.d}{\sim} t_\nu$ |
| HS (21) | $VaR_{i,t}(\tau) = Q_{\boldsymbol{r}_{i,t}^w}(\tau)$<br>$\boldsymbol{r}_{i,t}^w = (r_{i-w,t}, r_{i-w+1,t}, \ldots, r_{i-1,t})$ | |
| CAViaR-SAV (14) | $VaR_{i,t}(\tau) = \beta_0 + \beta_1 VaR_{i-1,t}(\tau) + \beta_2 \|r_{i-1,t}\|$ | |
| CAViaR-AS (14) | $VaR_{i,t}(\tau) = \beta_0 + \beta_1 VaR_{i-1,t}(\tau) + (\beta_2 \mathbb{1}_{(r_{i-1,t}>0)} + \beta_3 \mathbb{1}_{(r_{i-1,t}<0)}) \|r_{i-1,t}\|$ | |
| CAViaR-IG (14) | $VaR_{i,t}(\tau) = -\sqrt{\beta_0 + \beta_1 VaR_{i-1,t}^2(\tau) + \beta_2 r_{i-1,t}^2}$ | |
| CAViaR-X (16) | $VaR_{i,t}(\tau) = \beta_0 + \beta_1 VaR_{i-1,t}(\tau) + \beta_2 x_{i-1,t}$ | |

DRAFT

In this work, we propose three different combined predictors: MCS-Comb, W-MCS-Comb and MW-MCS-Comb. In particular:

- MCS-Comb: equally weighting all the $VaR$'s and $ES$'s obtained from the models entering the SSM of the *training* MCS, using the (unweighted) $FZ0$;
- W-MCS-Comb: equally weighting all the $VaR$'s and $ES$'s obtained from the models entering the SSM of the *training* MCS, using the *weighted* $FZ0$;
- MW-MCS-Comb: weighting proportionally to the *cumulated* $FZLoss$ all the $VaR$'s and $ES$'s obtained from the models entering the SSM of the *training* MCS, using the (unweighted) $FZ0$.

The *weighted* $FZ0$ adopted in the *training* MCS weights differently remote and recent observations. According to (26), the idea is that recent observations should have more weight with respect to remote observations. We adopt the same *exponentially weighted* approach of the RiskMetrics model to weight differently the observations. The MW-MCS-Comb predictor instead weights differently the VaR's and ES's of models entering the SSM of the *training* MCS (using the (unweighted) $FZ0$). In particular, models with smaller cumulated $FZ0$ values have larger importance to define the MW-MCS-Comb predictor.

## 3. Empirical analysis

The empirical analysis uses daily data on S&P 500 index collected from the Oxford-Man Institute's Realized Library. The full sample covers the period from January 18, 2011 to February 11, 2021, for $T = 2525$ daily observations. High-frequency variables are the realized volatility (4) at 5 and 10 minutes and the realized kernel (6). Low-frequency variables are the Geopolitical Risk (GPR, 11) and the National Activity Index (NAI), taken as the first release from the ALFRED archive. GPR and NAI are observed monthly. Let $M$ be the set of candidate models, $T_{in}$ the length of the rolling period, $lstep$ the number of static one-step-ahead forecasts generated by each candidate model, $T$ the sample size, and $nstep = (T - T_{in})/lstep$ the number of steps employed in the algorithm. The algorithm for obtaining the combined predictors is as follows:

1. Estimate all the candidate models over the window including observations from the period $t = 1 + j$ to $t = T_{in} + j$. Conditionally on the estimated parameters, generate for the following $lstep$

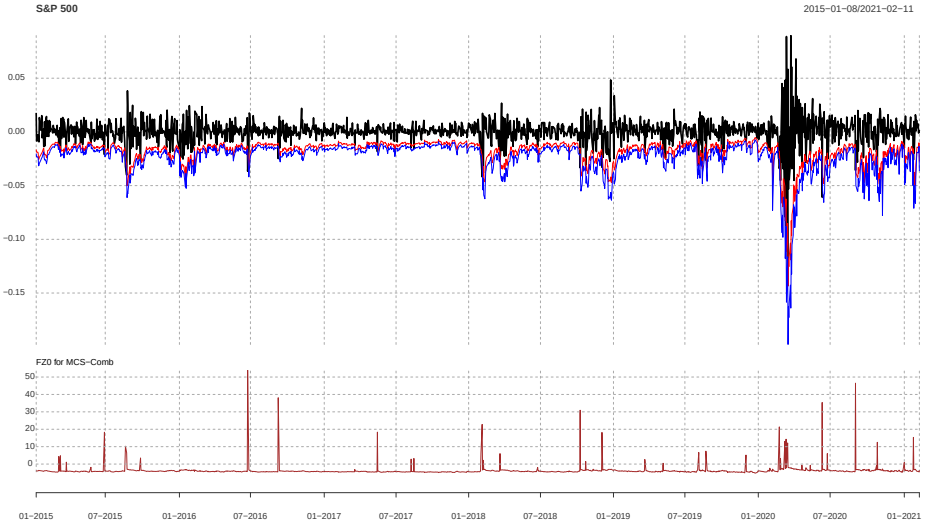



Figure 1: S&P 500 Log-returns (black line), VaR (red line), ES (blue line), and $FZ0$ values (bottom panel) of the MCS-Comb predictor

days both $VaR$ and $ES$ one-step ahead forecasts, with $i = 1, \cdots, M$.

2. Compute the *training* MCS over the training period going from $t = 1 + j$ to $t = T_{in} + j$.
3. Obtain the proposed combined predictors $VaR^{Comb}_{(T_{in}+j+1):(T_{in}+lstep+j)}$ and $ES^{Comb}_{(T_{in}+j+1):(T_{in}+lstep+j)}$.
4. Iterate steps 1, 2, and 3, with $j = \{0, lstep, 2lstep, \cdots, (nstep-1)lstep\}$.

We use $M = 25$ models, $T_{in} = 1000$, and $lstep = 25$. Globally, the out-of-sample period consists of 1525 observations from 8 January 2015 to 11 February 2021. The coverage level chosen is $\tau = 0.025$. Together with the three proposed combined predictors, we use two additional benchmarks: the equally weighted combination (EW-Comb) and the median combination (Median-Comb). The plot of the log-returns with the predicted VaR (red line) and ES (blue lines) obtained from the proposed MCS-Comb predictor for the out-of-sample period is in the top panel of Figure 1. The bottom panel of Figure 1 instead illustrates the $FZ0$ pattern of the MCS-Comb predictor, with the peaks observed when there are the VaR violations. Table 1 reports the violation rate (VR, in percentage) in the first column, the p-values of the UC, CC, and DQ tests for VaR in the columns from two to four, the p-values of the BD tests for VaR and ES in columns from five to eight, and the average of the $FZ0$ loss in the last column, for all the candidate models as well as the two combined benchmarks and the three proposed combined predictors. Shades of gray in the last column indicate that the model in the row enters the SSM of the MCS procedure at the significance level $\alpha = 0.25$. While many models (mainly some parametric and all the non-parametric specifications) do not perform well, all the proposed combined predictors pass the usual backtesting procedures at 1% significance level and, moreover, enter the SSM of the final MCS test.

DRAFT

## References

[1] Acerbi, C. and D. Tasche (2002). Expected shortfall: a natural coherent alternative to value at risk. *Economic notes 31*(2), 379–388.

[2] Amendola, A., M. Braione, V. Candila, and G. Storti (2020). A model confidence set approach to the combination of multivariate volatility forecasts. *International Journal of Forecasting 36*(3), 873–891.

[3] Amendola, A. and G. Storti (2015). Model uncertainty and forecast combination in high-dimensional multivariate volatility prediction. *Journal of Forecasting 34*(2), 83–91.

[4] Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association 96*(453), 42–55.

[5] Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999). Coherent measures of risk. *Mathematical finance 9*(3), 203–228.
[6] Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica 76*(6), 1481–1536.
[7] Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Journal of the operational research society 20*(4), 451–468.
[8] Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. *Econometrics and statistics 8*, 56–77.
[9] Bayer, S. and T. Dimitriadis (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics 20*(3), 437–471.
[10] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.
[11] Caldara, D. and M. Iacoviello (2022). Measuring geopolitical risk. *American Economic Review 112*(4), 1194–1225.
[12] Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, 841–862.
[13] Engle, R. F., E. Ghysels, and B. Sohn (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics 95*(3), 776–797.
[14] Engle, R. F. and S. Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*(4), 367–381.
[15] Fissler, T. and J. F. Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics 44*(4), 1680 – 1707.
[16] Gerlach, R. and C. Wang (2020). Semi-parametric dynamic asymmetric Laplace models for tail risk forecasting, incorporating realized measures. *International Journal of Forecasting 36*(2), 489–506.
[17] Glosten, L. R., R. Jagannanthan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance 48*(5), 1779–1801.
[18] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.
[19] Hansen, P. R., Z. Huang, and H. H. Shek (2012). Realized GARCH: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics 27*(6), 877–906.
[20] Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.
[21] Hendricks, D. (1996). Evaluation of value-at-risk models using historical data. *Economic policy review 2*(1), 39–69.
[22] Jorion, P. (1997). *Value at Risk*. Chicago: Irwin.
[23] Kupiec, P. H. (1995). *Techniques for verifying the accuracy of risk measurement models*, Volume 95. Division of Research and Statistics, Division of Monetary Affairs, Federal.
[24] Patton, A. J., J. F. Ziegel, and R. Chen (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics 211*(2), 388–413.
[25] Storti, G. and C. Wang (2023). Modeling uncertainty in financial tail risk: A forecast combination and weighted quantile approach. *Journal of Forecasting*.
[26] Taylor, J. W. (2008). Exponentially weighted information criteria for selecting among forecasting models. *International Journal of Forecasting 24*(3), 513–524.
[27] Taylor, J. W. (2019). Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution. *Journal of Business & Economic Statistics 37*(1), 121–133.
[28] Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting 36*(2), 428–441.

DRAFT

| | VR(%) | UC | CC | DQ | BD-1 | BD-2 | BD-3 | MCS |
|---|---|---|---|---|---|---|---|---|
| GARCH-N | 3.344 | 0.044 | 0.039 | 0.052 | 0.253 | 0.252 | 0.001 | -3.543 |
| GARCH-t | 3.279 | 0.063 | 0.047 | 0.063 | 0.623 | 0.627 | 0.248 | -3.599 |
| GJR-N | 3.213 | 0.087 | 0.134 | 0.235 | 0.206 | 0.202 | 0.142 | -3.597 |
| GJR-t | 2.885 | 0.347 | 0.256 | 0.230 | 0.205 | 0.260 | 0.530 | -3.643 |
| RGARCH-RVOL5-N | 3.344 | 0.044 | 0.039 | 0.011 | 1.000 | 1.000 | 1.000 | -3.566 |
| RGARCH-RVOL5-t | 3.410 | 0.031 | 0.032 | 0.007 | 0.422 | 0.310 | 0.857 | -3.598 |
| RGARCH-RVOL10-N | 3.803 | 0.002 | 0.005 | 0.002 | 0.144 | 1.000 | 0.802 | -3.51 |
| RGARCH-RVOL10-t | 4.393 | 0.000 | 0.000 | 0.000 | 0.256 | 0.171 | 0.838 | -3.532 |
| RGARCH-RK-N | 4.656 | 0.000 | 0.000 | 0.000 | 1.000 | 0.048 | 0.749 | -3.432 |
| RGARCH-RK-t | 5.180 | 0.000 | 0.000 | 0.000 | 0.075 | 0.059 | 0.735 | -3.445 |
| GM-N (GPR) | 4.066 | 0.000 | 0.001 | 0.000 | 0.024 | 0.024 | 0.047 | -3.032 |
| GM-t (GPR) | 3.213 | 0.087 | 0.134 | 0.000 | 0.007 | 0.006 | 0.576 | -3.285 |
| GM-N (NAI) | 3.672 | 0.006 | 0.004 | 0.001 | 0.030 | 0.030 | 0.166 | -2.677 |
| GM-t (NAI) | 3.016 | 0.211 | 0.214 | 0.041 | 0.172 | 0.004 | 0.413 | -3.424 |
| HS-25 | 5.770 | 0.000 | 0.000 | 0.000 | 0.024 | 0.016 | 0.000 | -2.844 |
| HS-50 | 4.459 | 0.000 | 0.000 | 0.000 | 0.033 | 0.037 | 0.000 | -3.068 |
| HS-100 | 3.279 | 0.063 | 0.000 | 0.000 | 0.196 | 0.248 | 0.055 | -3.186 |
| HS-250 | 3.672 | 0.006 | 0.001 | 0.000 | 0.241 | 0.194 | 0.126 | -3.262 |
| HS-500 | 3.934 | 0.001 | 0.000 | 0.000 | 0.178 | 0.167 | 0.028 | -3.203 |
| SAV | 3.148 | 0.119 | 0.275 | 0.155 | 0.812 | 0.815 | 1.000 | -3.614 |
| AS | 3.541 | 0.014 | 0.037 | 0.005 | 0.380 | 0.400 | 0.933 | -3.268 |
| IG | 2.885 | 0.347 | 0.531 | 0.248 | 1.000 | 0.897 | 1.000 | -3.633 |
| CAViaR-X-RVOL5 | 3.016 | 0.211 | 0.214 | 0.116 | 0.833 | 0.757 | 0.418 | -3.73 |
| CAViaR-X-RVOL10 | 3.279 | 0.063 | 0.109 | 0.069 | 1.000 | 0.793 | 0.314 | -3.715 |
| CAViaR-X-RK | 3.148 | 0.119 | 0.064 | 0.034 | 0.829 | 0.779 | 0.930 | -3.723 |
| EW-Comb | 2.951 | 0.273 | 0.081 | 0.032 | 0.708 | 0.657 | 0.484 | -3.65 |
| Median-Comb | 3.279 | 0.063 | 0.047 | 0.015 | 0.421 | 0.376 | 0.107 | -3.649 |
| MCS-Comb | 3.344 | 0.044 | 0.039 | 0.024 | 0.524 | 0.495 | 0.154 | -3.683 |
| W-MCS-Comb | 3.344 | 0.044 | 0.039 | 0.023 | 0.558 | 1.000 | 0.146 | -3.682 |
| MW-MCS-Comb | 3.344 | 0.044 | 0.039 | 0.024 | 0.514 | 0.495 | 0.124 | -3.684 |

**Notes**: Sample period: 2015-01-08 to 2021-02-11 (1525 observations). Column MCS represents the averages of the $FZ0$ loss (in the version of 24, reported in (1)). Shades of gray denote the inclusion in the Set of Superior Models of the MCS, at significance level $\alpha = 0.25$.

DRAFT

# Are Monetary Policy Announcements related to Volatility Jumps?

Giampero M. Gallo[a], Demetrio Lacava[b], and Edoardo Otranto[b]

[a]Italian Court of Audits (Corte dei conti – disclaimer), New York University in Florence, and CRENoS;
giampiero.gallo@nyu.edu
[b]University of Messina; dlacava@unime.it, eotranto@unime.it

**Abstract**

Central Banks interventions are frequent in response to exogenous events with direct implications on financial market volatility. In this paper, we introduce the Asymmetric Jump Multiplicative Error Model (AJM), which accounts for a specific jump component of volatility within an intradaily framework. Taking the Federal Reserve (Fed) as a reference, we propose a new model–based classification of monetary announcements based on their impact on the jump component of volatility. Focusing on a short window following each Fed's communication, we isolate the impact of monetary announcements from any contamination carried by relevant events that may occur within the same announcement day.

***Keywords:*** Financial markets, Realized volatility, Significant jumps, Monetary policy announcements, Multiplicative Error Model.

DRAFT

## 1. Introduction

The effectiveness of monetary policy communications crucially depends on a Central bank's credibility: "the extent to which the public believes that a shift in policy has taken place when, indeed, such a shift has actually occurred" (Cukierman, 1986) needs to be built through repeatedly consistent actions.
Monetary policy announcements have an essential role in driving financial markets expectations, particularly when financial turmoil or sudden changes in price dynamics require an active stance. There is a large consensus about the increase in volatility on announcement days: on those occasions, new information becomes available, either confirming prior beliefs or as a surprise. Either way, typically, an announcement has an impact on market activity, with increases in the number of trades some of which determine price changes or even jumps. In this paper we allow for a novel time–varying impact of announcements on volatility, by exploiting the granularity of intra daily data to derive consistent measures of volatility (see McAleer and Medeiros, 2008, for an exhaustive review), and identifying *significant* jumps (Andersen et al., 2007).

We suggest a new *composite* (Brownlees et al., 2012; Otranto, 2015) Multiplicative Error Model (MEM, Engle, 2002), which splits the dynamics of expected volatility at thirty–minute bins,[1] into a component for the continuous part of volatility and one for the discontinuous jump component, while accommodating time–of–day effects. Zooming into a particular time of the day around the release of a communication by the Federal Reserve (Fed), our model allows for the identification of the impact of

[1]The data are built starting from the tick–by–tick series of prices in the Trade and Quote (TAQ) database, which were cleaned according to the Brownlees and Gallo (2006) procedure and then sampled at five–minute intervals.

monetary policy announcements on volatility in the bin following the announcement itself. Our estimated results provide the basis for a new model–based classification procedure of monetary announcements according to their impact on volatility jumps right after the time of the announcement. Our analysis on some large–cap tickers shows a consistent classification approach across tickers and reveals useful information about the impact of monetary announcements on the volatility by market sector.

## 2. The model

The class of MEM (Engle, 2002; Engle and Gallo, 2006) specifies conditional volatility at bin $i$ of the day $t$ as the product of a time-varying positive conditional mean $\mu_{i,t}$ times an error term $\epsilon_{i,t}$ with a positive support. In this paper, we propose the Asymmetric Jump Multiplicative Error Model (AJM), where the conditional mean $\mu_{i,t}$ of the realized volatility $\widetilde{RV}_{i,t}$, adjusted to remove a time–of–day pattern, is the sum of $\varsigma_{i,t}$ (related to the continuous part) and $\kappa_{i,t}$ which reacts to the significant jumps:

$$
\begin{aligned}
&\widetilde{RV}_{i,t} = C_{i,t} + SJ_{i,t} \\
&C_{i,t} = \mu_{i,t}\epsilon_{i,t} \qquad\qquad \epsilon_{i,t}|\mathcal{F}_{i-1,t} \sim \Gamma(\vartheta, \tfrac{1}{\vartheta}) \\
&\mu_{i,t} = \varsigma_{i,t} + \kappa_{i,t} \\
&\varsigma_{i,t}, = [\omega + \alpha_1 C_{i-1,t} + \alpha_2 C_{i-2,t} + \beta\varsigma_{i-1,t} + \gamma I^{-}_{i-1,t} C_{i-1,t}] + \delta_1 |r_{i,t^\star}| + \delta_2 C_{i-1,t}, D_{i-1,t} \\
&\kappa_{i,t} = \varphi\mu_{i-1,t} + \psi SJ_{i-1,t}.
\end{aligned}
\tag{1}
$$

$\mathcal{F}_{i-1,t}$ is the information set at the previous bin (for the first bin of the day it is the last bin of the previous trading day); $C_{i,t}$ is the continuous volatility series and $SJ_{i,t}$ is the significant jump series.[2] In the specification for $\varsigma_{i,t}$ we adopt a standard AMEM(2,1) dynamics (in square brackets) with the asymmetric term connected to the sign of the most recent return through the indicator function $I^{-}_{i-1,t}$, augmented by two important sources of additional dynamics, a positive impact of the overnight return $r_{i,t^\star}$, which may capture news accumulation during market closing,[3] and a diminished impact of the first bin volatility on the second bin relative to the rest of the day (through a suitable dummy variable $D_{i,t}$). We keep a simple AR(1) specification for the jump component $\kappa_{i,t}$, driven by the observed significant jumps with an expected positive coefficient $\psi$.

Stationarity of both $\varsigma_{i,t}$ and $\kappa_{i,t}$ is required for the model to be stationary in covariance: relying on the identifiability results of (Engle and Lee, 1999), the continuous component is considered more persistent than the short–lived jumps, that is, $0 < \varphi < \beta < 1$ in Eq. (1). As for the required positiveness, while the usual GARCH constraints ($\omega, \alpha_1, \beta, \gamma > 0$) hold even in our framework, in the AJM(2,1) $\alpha_2$ can also assume negative value, with $\alpha_2 > -\alpha_1\beta$ ensuring $\varsigma > 0$ (Cipollini et al., 2020).

Among the distributions with positive support that can be specified for $\epsilon_{i,t}$, we opt for the Gamma distribution, because of its flexibility (values of $\vartheta$ ensure different shapes) and in view of its robustness (see Cipollini et al., 2013, for the equivalence between first order conditions and moment conditions in the univariate case). The Gamma generally depends only on the shape parameter $\vartheta$, so that it has a unit mean and a constant variance $(1/\vartheta)$: this makes the model very flexible with not only a time-varying conditional mean $E(\widetilde{RV}_{i,t}|\mathcal{F}_{i,t-1} = \mu_{i,t})$ but also a time varying conditional variance $Var(\widetilde{RV}_{i,t}|\mathcal{F}_{i,t-1} = \mu^2_{i,t}/\vartheta)$. The main implication of having a time varying volatility of volatility is that heteroskedasticity is implicit in the model, giving the opportunity of capturing possible structure in the innovations without resorting to any auxiliary regression.

Finally, Maximum Likelihood estimation involves Quasi Maximum Likelihood (QML) properties, ensuring consistency and asymptotic normality of the coefficient estimators (Engle, 2002), regardless the appropriateness of the selected distribution for the error term (Engle and Gallo, 2006). However, since $\vartheta$ is unknown, we resort to robust standard errors to shield against the actual shape of the distribution.

[2] $SJ_{i,t} = \mathcal{I}_{[J_{i,t} > \Phi_q]}(RV_{i,t} - BV_{i,t})$, were BV represents the realized bipower variation (Barndorff-Nielsen and Shephard, 2004, 2006) and $J_{i,t}$ is the jump statistics developed by Andersen et al. (2007).

[3] It is given by the difference between the opening log price of day $t$ and the closing log price of the previous day $t-1$, so that $t^\star$ denotes the time between $t-1$ and $t$. At the end of the first bin, such information is known.

DRAFT

## 2.1 Estimation results

Estimation results are reported in Table 1. coefficients are highly significant with a persistence (measured as $\alpha_1 + \alpha_2 + \beta + \gamma/2 + \delta_2/13$) estimated around $0.96$ across the considered assets. As expected, a large part of volatility comes from the ARCH terms, with $\hat{\alpha}_1$ ranging between $0.22$ and $0.27$ and $\hat{\alpha}_2$ between $-0.11$ and $-0.21$, given that, at an intraday level, volatility is very sensitive to news (with the first lag), but then contributes (with the second lag) to the absorption of news (see, for example Cipollini et al., 2020). This is also confirmed by the coefficients $\gamma$, which measure the impact of bad news (as represented by negative returns), are positive and significant at a $1\%$ level, with values between $0.01$ and $0.03$. The positiveness of $\delta_1$ represents evidence in favor of the *market opening effect*, i.e. a response of volatility at the beginning of each trading day between $0.008$ and $0.017$, due to whatever accumulation of news entails an overnight price movement. By the same token, the other bin–specific effect, marked by $\delta_2$, enters the model with the expected (and significant) negative sign, that is, the contribution to current bin volatility from the previous one is lower for bins after the first one. In such a case, the total effect is given by $\alpha_1 + \delta_2$, ranging between 0.18 and 0.23 across the considered series.

As for the jump component, the autoregressive coefficient $\varphi$ ranges between $0.017$ and $0.351$ across tickers, pointing out to a low persistence of the jump component - in line with the common view that volatility jumps are short-lived occurrences. The coefficient $\psi$ is positive, with expected jumps that are an increasing function of the identified significant jumps.

Table 1: Microsoft (MSFT), Goldman Sachs (GS), Johnson & Johnson (JNJ), 3M Co. (MMM). Estimation results with robust standard errors in parentheses. $SJ_{i,t}$ identified with $\Phi_{0.55}$. Sample period: January 03, 2010 – December 31,2021.

| | $\omega$ | $\alpha_1$ | $\alpha_2$ | $\beta$ | $\gamma$ | $\delta_1$ | $\delta_2$ | $\varphi$ | $\psi$ | $\vartheta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| MSFT | 0.1173 | 0.2589 | -0.2045 | 0.8926 | 0.0105 | 0.0081 | -0.0304 | 0.3509 | 0.2622 | 5.6748 |
| | (0.0358) | (0.0061) | (0.0128) | (0.0177) | (0.0032) | (0.0023) | (0.0086) | (0.0575) | (0.0151) | (0.0517) |
| GS | 0.2064 | 0.2463 | -0.1706 | 0.8763 | 0.0137 | 0.0137 | -0.0358 | 0.2040 | 0.2476 | 5.8705 |
| | (0.0399) | (0.0058) | (0.0118) | (0.0110) | (0.0023) | (0.0020) | (0.0072) | (0.0501) | (0.0154) | (0.0569) |
| JNJ | 0.1679 | 0.2442 | -0.1556 | 0.8723 | 0.0122 | 0.0154 | -0.0494 | 0.1141 | 0.2368 | 5.3593 |
| | (0.0246) | (0.0067) | (0.0105) | (0.0094) | (0.0023) | (0.0030) | (0.0077) | (0.0475) | (0.0153) | (0.0607) |
| MMM | 0.2215 | 0.2229 | -0.1081 | 0.8465 | 0.0273 | 0.0173 | -0.0457 | 0.0166 | 0.2579 | 5.7287 |
| | (0.0389) | (0.0080) | (0.0188) | (0.0118) | (0.0040) | (0.0026) | (0.0076) | (0.0685) | (0.0180) | (0.0681) |

To give a visual documentation of the relationship between Fed announcements[4] and significant volatility jumps, Figure 1 shows the evolution of the realized volatility, $\widetilde{RV}$ (grey line, left axis), together with the significant jump size ($SJ_{i,t}$) observed on announcement days (blue dots – both empty and solid – right axis scale) between December 14, 2016 and December 31, 2021 with a total of 44 announcements. On the basis of the empirical evidence in Figure 1, it is clear how the announcement effect is not constant over time: in particular, while some announcements have a limited impact on jumps, for some specific Fed's communications, significant jumps account for more than 20% of the overall level of volatility[5] (blue dots). This is the case, for example, of the announcement released on March 20, 2019, when the Federal Open Market Committee (FOMC) decided to maintain the federal funds rates (FFR) within the 2.25–2.50 percent bracket. Interestingly, a similar decision (with the FFR left constant at 0-0.25 percent) on September 16, 2020, is associated to a jump marking 55% of the total level of volatility. In synthesis, the same quantitative decision could translate into a different qualitative impact on jumps, pointing to the relevance of market expectations about the announcement rather than the decision per se.

[4]Data about the date, the time and the content of monetary policy announcements are available at https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm

[5]The share of volatility due to significant jumps can be computed as $\frac{SJ_{i,t}}{\widetilde{RV}_{i,t}} = 1 - \frac{C_{i,t}}{\widetilde{RV}_{i,t}}$.

DRAFT

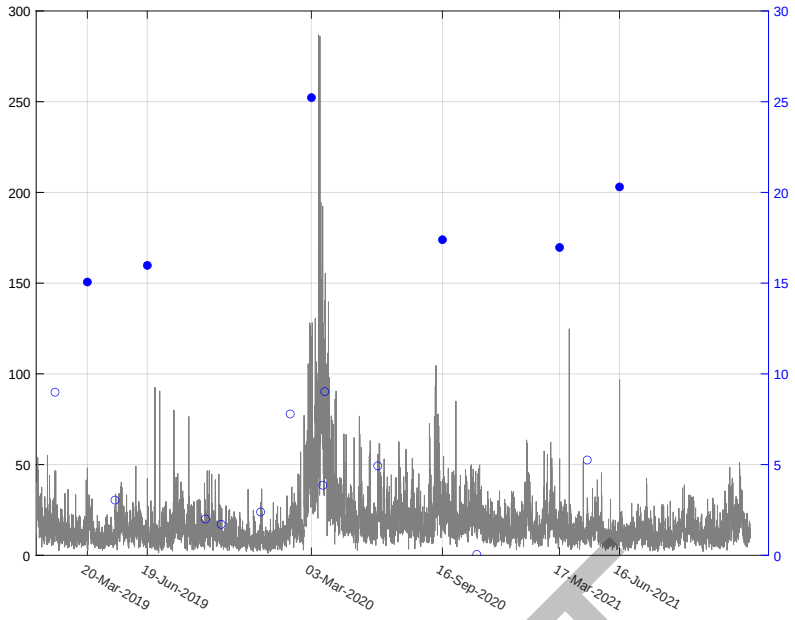



Figure 1: MSFT realized volatility (grey line, measured on the left axis) and jumps (blue dots both empty and solid, on the right axis) associated to monetary announcements (details in the text). Solid blue dots identify volatility jumps accounting for more than 20% of the overall level of volatility. Sample period: December 14, 2016 – December 31, 2021.

DRAFT

## 3. The classification method

Our model suggests a simple mechanism for a model-based classification of monetary announcements according to their impact on volatility jumps, that is, we distinguish expected jumps (as represented by $\kappa_{i,t}$ in Eq. 1) from what we call *jump surprise* or *unexpected* jumps. The latter is defined as the difference between what we expect and what we observe, $J^{surprise}_{\iota,\tau} = \kappa_{\iota,\tau} - SJ_{\iota,\tau}$ with $\tau = t$ on announcements days and $\iota$ indicating the first bin past the announcement time.

The proposed approach belongs to the class of time-point classification methods (Aghabozorgi et al., 2015): focusing on the $\tau$ announcement days, we suggest to classify Fed's announcements, according to whether at the bin $\iota$ within that day we detect

1. an Upward Spike (local maximum), if $\kappa_{\iota-1,\tau} < \kappa_{\iota,\tau} > \kappa_{\iota+1,\tau}$;
2. a Downward Spike (local minimum), if $\kappa_{\iota-1,\tau} > \kappa_{\iota,\tau} < \kappa_{\iota+1,\tau}$;
3. a Boost (an increase), if $\kappa_{\iota-1,\tau} < \kappa_{\iota,\tau}$;
4. a Drop (a reduction), if $\kappa_{\iota-1,\tau} > \kappa_{\iota,\tau}$,

for $\kappa$ (and, similarly, for $J^{surprise}$). Note that the identification of the Boost or the Drop bins of the announcement day is made excluding the corresponding spikes. Our classification method has the merit of being immediately applicable when an announcement is released and has the unique characteristic of giving information about the expected jumps.

The classification results are shown in Table 2. Focusing on the expected jumps $\kappa_{\iota,\tau}$: most of the announcements (between 29% and 34%) correspond to a Upward Spike, followed by Downward Spikes (26% – 24% of the cases), and Boost (20% – 26%). Results are quite homogeneous across assets, even if only few announcements belong to the same cluster (Downward Spike) for all the considered tickers.

Interestingly, some announcements affected only specific sectors, e.g. the decisions about the FFR on June 17, 2015, July 31, 2019 and July 28, 2021, which are classified as Upward Spike only for the financial sector. Conversely, some Fed communications had a different impact in different sectors, with announcements on April 27, 2016 and August 1, 2018 causing a Upward Spike for the industrial

Table 2: Classification results based on the restricted AJM. The total 104 Fed's monetary announcements are classified basing on their effects either on $\kappa_{\iota,\tau}$ or on the jump surprise, $J_{i,t}^{Surprise}$.

| | $\kappa_{\iota,\tau}$ | | | | $J_{\iota,\tau}^{surprise}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Upward Spike | Downward Spike | Boost | Drop | Upward Spike | Downward Spike | Boost | Drop |
| MSFT | 30 | 27 | 27 | 19 | 32 | 45 | 5 | 21 |
| GS | 30 | 27 | 30 | 17 | 34 | 39 | 11 | 20 |
| JNJ | 33 | 30 | 30 | 11 | 34 | 51 | 5 | 14 |
| MMM | 35 | 35 | 21 | 13 | 34 | 45 | 4 | 21 |

Table 3: Percentage adjusted Rand-index between: **Panel a** – pairs of assets basing either on $\kappa$ (lower triangular matrix) or $J^{Surprise}$ (upper triangular matrix); **Panel b** – $\kappa$ and $J^{Surprise}$, and $\kappa$ and $SJ$.

| **Panel a** | MSFT | GS | JNJ | MMM |
|---|---|---|---|---|
| MSFT | - | 57.67% | 55.06% | 55.73% |
| GS | 63.13% | - | 56.46% | 56.68% |
| JNJ | 61.67% | 61.31% | - | 56.42% |
| MMM | 61.24% | 60.51% | 59.65 % | - |
| **Panel b** | | | | |
| $\kappa$ vs $J^{Surprise}$ | 81.11% | 75.58% | 77.22% | 79.37% |
| $\kappa$ vs $SJ$ | 63.24% | 56.55% | 62.47% | 61.82% |

DRAFT

sectors (MMM) while they had the opposite effect (Downward Spike) on the financial and Information Technology (IT) sectors. As expected, the pharmaceutical (JNJ) sector is not integrated with the others, with around the 12% of announcements that have a different classification with respect to the other securities. This does not hold for the IT sector (MSFT), which shares the 27% of the classification with the financial sector – the percentage is 26% (31%) for announcements classified as Upward Spike (Downward, respectively). For an investor, the usefulness of this kind of information is twofold: on the one hand, knowing the degree of interconnection between sectors is important in designing diversification strategy; on the one other, predicting turning points of $\kappa_{\iota,\tau}$ is crucial for investment strategies based on momentum. When one turns to jump surprises, a higher degree of integration is encountered for the pharmaceutical sector with 37% classified as Drop only for JNJ. Similarly, the integration between the IT and financial sectors increases, with 35% of announcements having a common classification for MSFT and JPM – the percentage is 33% (62%) if the Upward Spike (Downward, respectively) is considered.

In evaluating the accuracy of our classification method, we rely on the adjusted Rand-index (Rand, 1971; Hubert and Arabie, 1985) computed between pairs of assets. Table 3 shows results deriving either from $\kappa$ or $J^{Surprise}$. Consistently, elements of the lower portion (expected jumps) are pairwise larger than the corresponding elements of the upper portion (surprise jumps), with indices that are above to 60%: given their size, a similar clustering among tickers is apparent, but an idiosyncratic component of reaction to announcements is present. This is confirmed by the fact that the index is higher than 75% when $\kappa$ is compared to $J^{surprise}$, while it 63% at most, when we compare $\kappa$ to $SJ$.

## 4. Concluding remarks

We introduce a novel MEM to decouple the continuous part of volatility from its jump component, geared toward a model–based classification of Central Bank's announcements based on their impact on volatility jumps. Conversely to the existing approaches, we reconstruct the dynamics of intradaily volatility by thirty-minute bins distinguishing a continuous part and a jump part components and accommodating some peculiarities related to the time–of–day effect. We allow for a time-varying announcement effect, with each Central Bank's communication having its own effect on volatility, with the possibility to classify announcements according to whether we detect a local maximum (minimum) or a simple increment

(reduction) of the considered series on announcement days.

The empirical application, based on some US tickers, reveals both commonalities and differences among different sectors of the market, with a high interconnection between the IT and financial sectors, with the pharmaceutical sector being less close. The adjusted Rand-index, reveals a good accuracy of our classification, which has the merit of being immediately applicable when an announcement is released.

The model could be further refined in at least two directions: a feedback effect could be accommodated within the dynamics of the continuous part making it depend on the lagged expected jump component; furthermore, in line with the logic of the Composite MEM, the two components enter the model additively, while an interesting comparison would be with a model with multiplicative components.

## References


Aghabozorgi, S., A. S. Shirkhorshidi, and T. Y. Wah (2015). Time-series clustering–a decade review. *Information systems 53*, 16–38.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2007). Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *Review of Economics and Statistics 89*, 701–720.

Barndorff-Nielsen, O. and N. Shephard (2004). Power and bipower variation with stochastic volatility and jumps (with discussion). *Journal of Financial Econometrics 2*, 1–48.

Barndorff-Nielsen, O. E. and N. Shephard (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics 4*, 1–30.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Bomfim, A. N. (2003). Pre-announcement effects, news effects, and volatility: monetary policy and the stock market. *Journal of Banking & Finance 27*(1), 133–151.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2012). Multiplicative error models. In L. Bauwens, C. Hafner, and S. Laurent (Eds.), *Volatility Models and Their Applications*, pp. 223–247. Wiley.

Brownlees, C. T. and G. M. Gallo (2006). Financial econometric analysis at ultra–high frequency: Data handling concerns. *Computational Statistics and Data Analysis 51*, 2232–2245.

Cipollini, F., R. F. Engle, and G. M. Gallo (2013). Semiparametric vector MEM. *Journal of Applied Econometrics 28*, 1067–1086.

Cipollini, F., G. M. Gallo, and A. Palandri (2020). Realized variance modeling: decoupling forecasting from estimation. *Journal of Financial Econometrics 18*(3), 532–555.

Cukierman, A. (1986). Central bank behavior and credibility: some recent theoretical developments. *Federal Reserve Bank of St. Louis Review 68*(5), 5–17.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics 17*, 425–446.

Engle, R. F. and G. M. Gallo (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics 131*, 3–27.

Engle, R. F. and G. J. Lee (1999). A permanent and transitory component model of stock return volatility. In R. F. Engle and H. White (Eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*, pp. 475–497. Oxford University Press, Oxford.

Hattori, M., A. Schrimpf, and V. Sushko (2016, April). The response of tail risk perceptions to unconventional monetary policy. *American Economic Journal: Macroeconomics 8*(2), 111–36.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*, 193–218.

McAleer, M. and M. Medeiros (2008). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics 147*, 104–119.

Otranto, E. (2015). Capturing the spillover effect with multiplicative error models. *Communications in Statistics-Theory and Methods 44*(15), 3173–3191.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association 66*(336), 846–850.

Wright, J. H. (2012). What does monetary policy do to long-term interest rates at the zero lower bound? *The Economic Journal 122*(564), F447–F466.


DRAFT

# Regularized Estimation and Prediction of the El Niño/Southern Oscillation Cycle

Tommaso Proietti[a] and Alessandro Giovannelli[b]

[a]Università di Roma "Tor Vergata" - Dipartimento di Economia e Finanza; Via Columbia, 2, 00133 Roma RM; tommaso.proietti@uniroma2.it

[b]Università dell'Aquila - Dipartimento di Ingegneria e scienze dell'informazione e matematica; Via Vetoio, Coppito - 67100 L'Aquila; alessando.giovannelli@univaq.it

**Abstract**

The El Niño/Southern Oscillation (ENSO) phenomenon is one of the most important sources of interannual climate variability. This paper focuses on the prediction of the sea surface temperatures in the four Niño regions, which represent the oceanic manifestation of ENSO. The series are the components of a larger dimensional dynamic system constituted by 15 time series that include the atmospheric component (sea level pressure in Darwin and Tahiti), trade and zonal winds anomalies, as well as series related to the ocean heat content. We propose a prediction method based on a novel regularized multivariate Durbin-Levinson algorithm, which performs the projection of the series into its past, avoiding the curse of dimensionality. The regularization concerns both the lag and the cross-sectional dimensions: we taper and threshold the partial canonical correlations computed on a mixture sample cross-covariance matrix that shrinks the traditional estimator towards a seemingly unrelated system. Using a rolling forecast experiment, we show that the cyclical properties of the series are essential for multi-step ahead forecasting and the use of cross–sectional information leads to significant forecasting gains with respect to traditional vector autoregressive modelling.

***Keywords:*** Regularization; Partial Canonical Correlations; Vector Autoregressive Models

DRAFT

# 3 Contributed Sessions

DRAFT

# Bayesian density estimation for modeling age-at-death distribution

Davide Agnoletto[a], Tommaso Rigon[b], and Bruno Scarpa[a]

[a]Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova, Italy
[b]Department of Economics, Management and Statistics, University of Milano–Bicocca, 20126 Milano, Italy

**Abstract**

Age-at-death distribution represents a fundamental quantity for the study of mortality since it provides an important tool to evaluate the longevity and lifespan variability in a population. In this paper, we analyze the set of such curves for the male population of Campania's municipalities region in 2020. By exploiting a Bayesian nonparametric mixture model, we produce an estimate of age-at-death distributions even in presence of small populations, simultaneously modeling multiple curves.

***Keywords:*** age-at-death distribution, Bayesian modeling, Campania.

DRAFT

## 1. Introduction

Modeling human mortality has challenged statisticians and demographers over the years, resulting in a wide range of proposed models. John Graunt and Edmund Halley developed the first life tables in the 17th century. Over time, this tool has been improved and become more mathematically rigorous, and it still represents an essential method for analyzing the behavior of the mortality phenomenon in the population. Modern life tables provide several quantities from which it is possible to derive the survival function, the hazard function, and the probability density function, which are the most diffused tools for the study of mortality. Each of these quantities is useful to analyze a specific aspect of mortality; thus, all three must be investigated to fully understand the phenomenon [3]. Historically, the literature focused on mortality rates. The first attempts to model mortality appeared in the 19th century [7; 10] and some generalizations have been proposed in the last century [14; 8]. These tools were developed within a frequentist framework and the estimation process can be problematic due to over-parametrization. For those reasons, some tools for face mortality modeling have also been developed in the Bayesian literature as well [4; 13].

More recently, other approaches focused on modeling the age-at-death distribution of a population instead of the mortality rates. The reason is that this quantity can be seen as a density function characterizing mortality. Based on this idea, [11] propose a six-parameters model for that quantity which mixes a half-normal distribution with a generalization of the skew-normal distribution while [2] offer a three-component mixture of a Dirac mass, a Gaussian distribution, and a skew-normal distribution to model and forecast dynamically the age-at-death distribution.

This paper aims to analyze the mortality phenomenon for the male population of the Campania region for the year 2020 for each municipality, focusing, on modeling the distribution of age-at-death. In particular, we are interested in detecting if the mortality phenomenon behaves similarly in some areas of the region and, eventually, in investigating the causes. To do so, we rely on the model proposed by [1], which is effective when dealing with curves related to small populations.

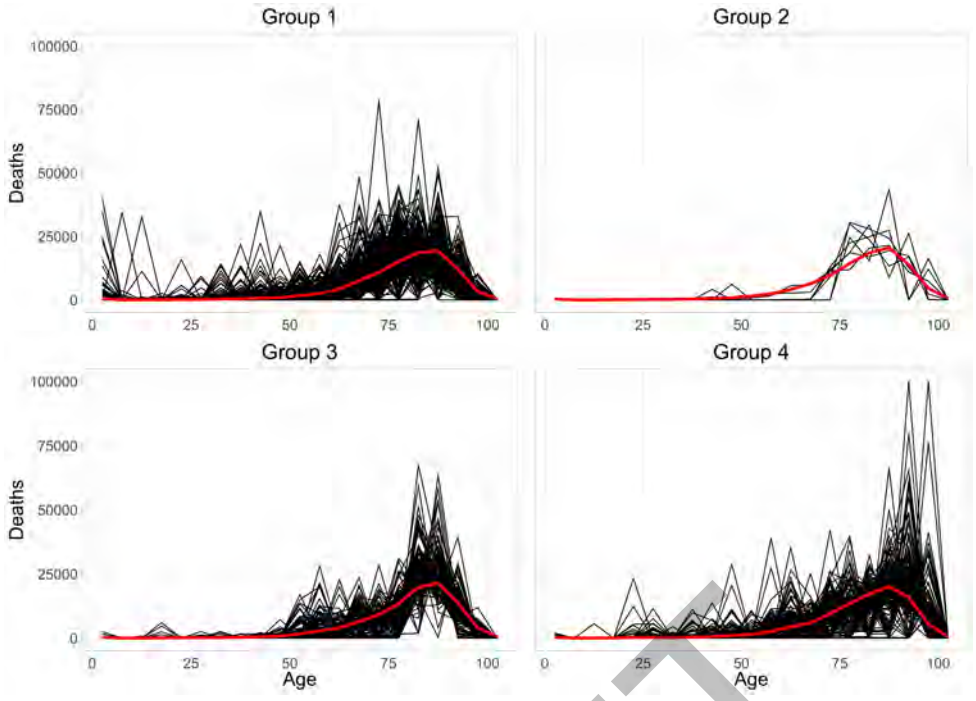



Figure 1: The age-at-death distributions $\boldsymbol{d}_j$ for each municipality computed using the classical life table technique allocated in the detected non-empty group $h = 1, \ldots, 4$ (in black), and the corresponding posterior expected quantity $\mathbb{E}(\boldsymbol{d}_j \mid G_j = h)$ (in red).

DRAFT

## 2. A Bayesian nonparametric mixture model

Let $d_{x,j}$ the number of deaths at age classes $[x, x+5)$ for population $j$, with $j = 1, \ldots, J$ and $x = 0, 5, \ldots, 95, 100$, where the last age class is $[100, +\infty)$, and let $\boldsymbol{d}_j = (d_{0,j}, \ldots, d_{100,j})$ represent the age-at-death distribution for population $j$.

Such a distribution is classically computed using life tables, that provide the key conversion from a set of observed age-specific death rates to a set of age-specific probabilities of dying. In this way, an age standardization of all the quantities is operated in order to eliminate the differences in age composition in separate populations. The computation of the tables is usually made with a decrement process starting from a standard fictitious population of size $10^5$ [12]. Despite their versatility and popularity, the life-table approach has some important drawbacks when facing the simultaneous modeling of small populations. For instance, consider the case of modeling the age-at-death distribution for $J$ municipalities. Firstly, in that situation, the life table provides an estimate for each municipality separately, although some similarities could occur. Secondly, when the resident population in a municipality is very small, so is the corresponding number of deaths. As a consequence, the shape of the age-at-death distribution can be extremely jagged and irregular. However, it is reasonable to think that such an unusual pattern is not due to some peculiarity of the mortality phenomenon but instead to the fact that each death can strongly impact the shape of the distribution in a small municipality. These two main problems also persist in the widely used model-based solutions mentioned in Sect. 1.

Based on these considerations, [1] provide a Bayesian nonparametric mixture model allowing for borrowing of information and simultaneous modeling of different populations. Assuming that each age-at-death distribution is calculated using a classical fixed initial population $n$ [12] such that $\sum_{x \in \{0,5,\ldots,95,100\}} d_{x,j} = n$ for all $j = 1, \ldots, J$, they consider $\boldsymbol{d}_j$ as the outcome of $n$ realizations from a multinomial random variable and model the set of $J$ "raw" age-at-death distributions (i.e., the age-at-death distribution computed with the life-table method) as a mixture of $H$ latent of components. We write

$$\begin{aligned} \boldsymbol{d}_j \mid H, G_j = h, \boldsymbol{\pi}_h &\sim \text{Multinomial}(n; \boldsymbol{\pi}_h), \quad \text{independently for } j = 1, \ldots, J, \\ G_j \mid H, \boldsymbol{w}_H &\sim \text{Multinomial}(1; \boldsymbol{w}_H), \quad \text{independently for } j = 1, \ldots, J, \end{aligned}$$

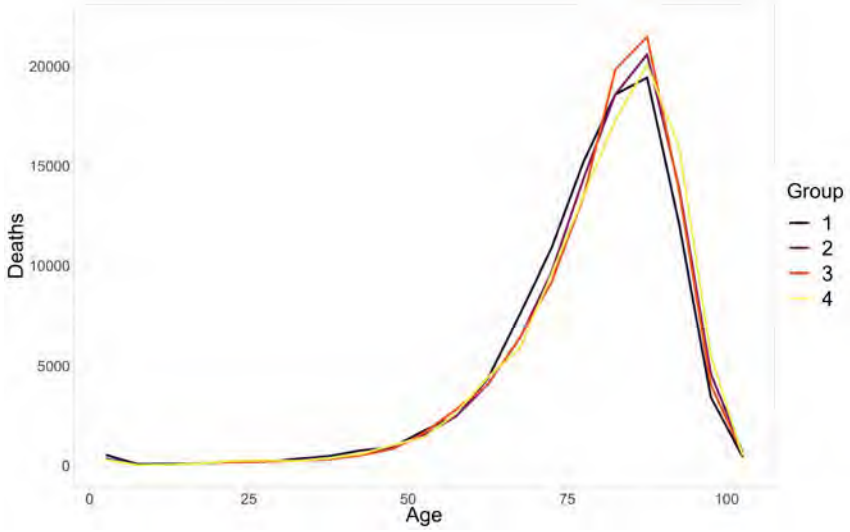



Figure 2: Expected age-at-death distribution for each group. The labeling of the groups is chosen based on the posterior median age-at-death, in increasing order. The first group contains the distributions with a higher number of deaths in adult and young-adult age classes. The raw distributions in the second cluster do not present particular irregularities. The third has its peak in the $[80, 90)$ age class and while the fourth group collects the curve with a high number of deaths which are 90 or older.

DRAFT

where $G_j$ is the latent allocation variable of observation $\boldsymbol{d}^j$ and $\boldsymbol{w}_H = (w_1, \ldots, w_H)$. Each element $G_j \in \{1, \ldots, H\}$, $j = 1, \ldots, J$, take one of the $H$ group labels as its realization. We are interested in inferring the number of non-empty components and the probabilities of dying at each age class $\boldsymbol{\pi}_h = (\pi_{0h}, \ldots, \pi_{100h})$, for all $h$ such that $\sum_{j=1}^{J} \mathbb{1}(G_j = h) > 0$. The allocation of each raw curve to a mixture component depends only on its shape, and clustering arises in a natural way. In order to guarantee flexibility, a Dirichlet process prior is set to express the prior knowledge about the probability of dying at each age-class [9; 6], with the induced distribution

$$\boldsymbol{\pi}_h \mid \alpha, P_0 \sim \text{Dir}_{21}\left(\alpha P_0^{(0)}, \ldots, \alpha P_0^{(100)}\right),$$

for all $h = 1, \ldots, H$. The base measure $P_0$ represents the *a-priori* knowledge about how the mortality phenomenon behaves in relation to age, and $P_0^{(x)}$ denotes the death probability assigned to the base measure to age class $[x, x+5)$, $x \in \{0, 5, \ldots, 100\}$. The precision parameter $\alpha$ regulates the shrinkage of induced distribution towards the base measure. Hence for each identified cluster, the model does borrowing of information between the base measure and the group of age-at-death distributions having similar shapes among the $J$ curves. Coherently with the Bayesian mixture models literature, Dirichlet prior distribution for the mixture weights is specified, as

$$\boldsymbol{w}_H \mid H \sim \text{Dir}_H\left(\frac{1}{H}, \ldots, \frac{1}{H}\right).$$

Finally, we specify a beta-negative-binomial prior on the number of mixture components $H$. The estimation algorithm is provided in [5]. For further technical details, see [1].

## 3. Campania 2020 male population data

This study focuses on the mortality phenomenon for the male population of the Campania region in 2020. The daily deaths for each municipality are freely available from the Italian National Institute of Statistics (ISTAT) database and, after aggregation, the raw curve for each municipality is computed following [12]. The municipalities under examination are 550 and the associated distributions present very different shapes according to the size of their respective population.

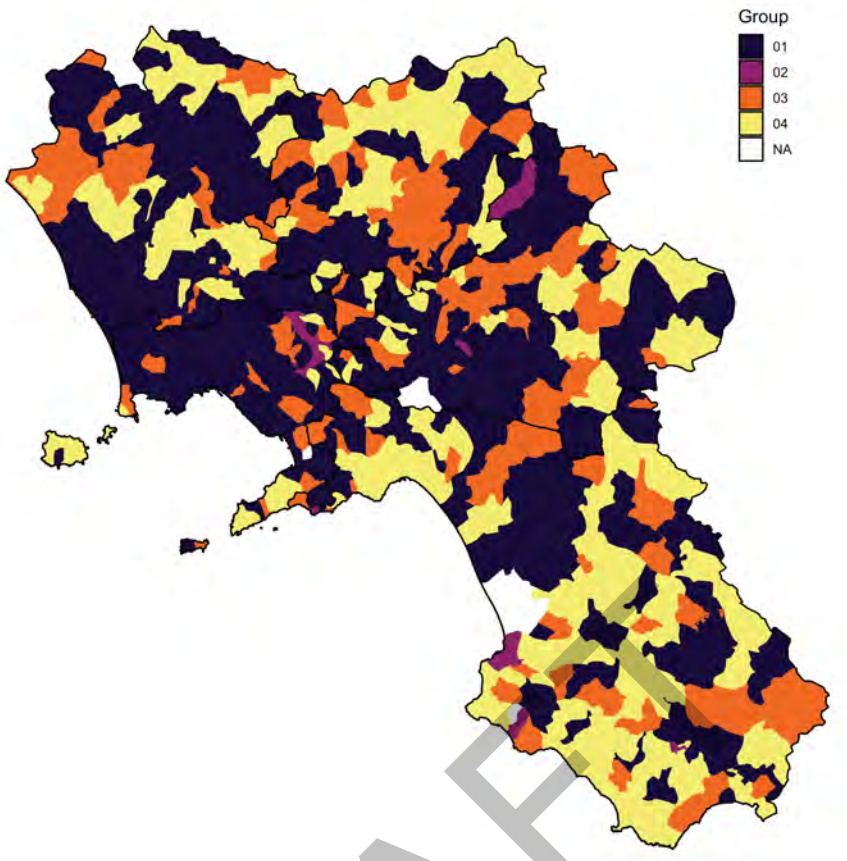



Figure 3: The geographical position of each municipality. The color represents the detected group $G_j$, $j = 1, \ldots, J$, in which each municipality is allocated.

Data are analyzed by applying the model proposed in [1]. In particular, we consider as baseline measure the curve of the entire Italian male population for the year 2020. Adopting a Bayesian methodology allows one to make inference on any functional of the posterior distribution. Uncertainty measures are naturally obtained from the sampling procedure.

The model detects four groups of curves, each of that is characterized by the shape of the curves allocated in, as is shown in Fig. 1. These features of the expected age-at-death distributions for each group are shown in (Fig. 2). An interesting result is obtained by observing the geographical position of the municipalities in relation to each group (Fig. 3). It is evident how the first group has a massive geographical characterization, indeed, it collects the majority of the municipalities of Naples hinterland. This area constitutes the so-called "Terra dei Fuochi," ninety municipalities between the provinces of Naples and Caserta affected by illegal dumping and wild abandonment of urban and special waste, often associated with its burning. Waste burning has caused concern because of the fumes and the pollutants poured onto farmland that can put the local population's health at risk. Several studies [15] identify a causal relationship between the presence of waste in this area and the formation of cancers in the population. This can be the reason for such anticipation of mortality in that geographical area together with the outbreak of COVID-19 pandemic that affected the region.

This analysis shows the potentiality of the Bayesian density estimation models for dealing with the age-at-death distribution of small populations. The borrowing of information that the model implements allows for estimated distributions that are a reliable representation of how mortality behaves in a set of populations of different sizes.

DRAFT

## References


[1] Agnoletto, D., Rigon, T., Scarpa, B.: Estimation of mortality curves using a Dirichlet process prior. Technical report (2023).

[2] Aliverti, E., Mazzuco, S., Scarpa, B.: Dynamic modeling of mortality via mixtures of skewed distribution functions. Journal of the Royal Statistical Society: Series A (Statistics in Society), **185**, 1030–1048 (2022).

[3] Basellini, U., Camarda, C. G.: Modelling and forecasting adult age-at-death distributions. Population studies, **73**(1), 119–138 (2019).

[4] Dellaportas, P., Smith, A. F. M., Stavropoulos, P.: Bayesian analysis of mortality data. Journal of the Royal Statistical Society: Series A (Statistics in Society) **164**(2): 275–291 (2001).

[5] Frühwirth-Schnatter, S., Malsiner-Walli, G., & Grün, B.: Generalized mixtures of finite mixtures and telescoping sampling. Bayesian Analysis, **16**(4), 1279–1307 (2021).

[6] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin: Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC (2013).

[7] Gompertz, B: On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London, **115**, 513–585 (1825).

[8] Heligman, L., Pollard., J. H.: The age pattern of mortality. Journal of the Institute of Actuaries, **107**(1), 49–80 (1980).

[9] Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G.: Bayesian nonparametrics, vol. 28. Cambridge University Press (2010).

[10] Makeham, W.: On the Law of mortality and construction of annuity tables. The Assurance Magazine and Journal of the Institute of Actuaries, **8**, 301–310 (1860).

[11] Mazzuco, S., Scarpa, B., Zanotto, L.: A mortality model based on a mixture distribution function. Population Studies, **72**(2), 191-200 (2018).

[12] Preston, S., Heuveline, P. and Guillot, M.: Demography, measuring and modeling population processes. Blackwell Publishers, Malden, MA (2001).

[13] Sharrow, D. J., Clark, S. J., Collinson, M. A., Kahn, K., Tollman, S. M.: The age pattern of increases in mortality affected by HIV: Bayesian fit of the Heligman-Pollard model to data from the Agincourt HDSS field site in rural northeast South Africa. Demographic Research **29**, 1039–1096 (2013).

[14] Siler, W.: Parameters of mortality in human populations with widely varying life spans. Statistics in Medicine, **2**, 373–380 (1983).

[15] Triassi, M., Alfano, R., Illario, M., Nardone, A., Caporale, O., Montuori, P.: Environmental pollution from illegal waste disposal and health effects: A review on the "Triangle of Death". International journal of environmental research and public health, **12**(2), 1216–1236 (2015).


DRAFT

# Bayesian Mixing Distribution Estimation in the Gaussian-smoothed 1-Wasserstein Distance

Catia Scricciolo[a]

[a]Dipartimento di Scienze Economiche, Università di Verona, Via Cantarane 24, 37129 Verona;
catia.scricciolo@univr.it

**Abstract**

We consider the problem of nonparametric mixing distribution estimation for discrete exponential family models. It has been recently shown that, under the Gaussian-smoothed optimal transport (GOT) distance, *i.e.*, the 1-Wasserstein distance between the Gaussian-convolved distributions, the accuracy of the nonparametric maximum likelihood estimator (NPMLE) is improved to a polynomial rate from the sub-polynomial rate with respect to the standard 1-Wasserstein distance. The focus of this work is on studying the problem taking a Bayesian nonparametric approach. We provide conditions under which the Bayes' estimator of the true mixing distribution converges at least as fast as the NPMLE in the GOT distance.

***Keywords:*** mixture models, nonparametric MLE, Poisson distribution, rates of convergence.

DRAFT

## 1. Introduction

We consider the mixing distribution estimation problem. Let $\{p(\cdot;\theta), \theta \in \Theta \subseteq \mathbb{R}\}$ be a known parametric family of probability "densities" (including the possibility of discrete atoms) with respect to a dominating measure on $\mathscr{X} \subseteq \mathbb{R}$ and let $X^{(n)} := (X_1, \ldots, X_n)$ be $n$ independent and identically distributed (i.i.d.) observations drawn from the mixture

$$f_{G_0}(x) := \int_\Theta p(x;\theta)\,\mathrm{d}G_0(\theta), \quad x \in \mathscr{X},$$

where the mixing distribution $G_0$ of $\theta$ is unknown. The goal is to estimate $G_0$ using the data $X^{(n)}$. We consider a one-parameter discrete exponential family, see, *e.g.*, Efron (2022) (1),

$$p(x;\theta) = g(\theta)w(x)\theta^x, \quad x \in \mathbb{N}_0 := \{0, 1, 2, \ldots\}, \tag{1}$$

with $w(x) > 0$ for all $x \in \mathbb{N}_0$ and $\theta \in [0, \theta_*]$ for $\theta_* < \theta_r$, where $\theta_*$ is a known fixed constant, $\theta_r \in (0, \infty]$ is the radius of convergence of the power series $\theta \mapsto \sum_{x=0}^\infty w(x)\theta^x$ and $g(\cdot)$ is an analytic function in a neighborhood of zero. Clearly,

$$\frac{1}{g(\theta)} = \sum_{x=0}^\infty w(x)\theta^x < \infty \ \text{ for } \theta \in [0, \theta_*]$$

and

$$\int_0^{\theta_*} g(\theta)\theta^x\,\mathrm{d}G_0(\theta) < \infty \ \text{ for every } x \in \mathbb{N}_0.$$

### Examples

Model (1) includes the Poisson and negative binomial distributions:

- Poisson distribution, $\mathrm{Poi}(\theta)$, $\theta \in [0,\infty)$, with

$$g(\theta) = e^{-\theta} \quad \text{and} \quad w(x) = \frac{1}{x!};$$

- negative binomial distribution, $\mathrm{NBin}(k, 1-\theta)$, $k \in \mathbb{N}$ and $\theta \in [0,1]$, with

$$g(\theta) = (1-\theta)^k \quad \text{and} \quad w(x) = \binom{x+k-1}{x}.$$

□

For model (1) the mixture takes the form

$$f_{G_0}(x) := \int_0^{\theta_*} g(\theta) w(x) \theta^x \, \mathrm{d}G_0(\theta), \quad x \in \mathbb{N}_0. \tag{2}$$

Estimation of $G_0$ has been investigated using, among others, the maximum likelihood approach. The nonparametric maximum likelihood estimator (NPMLE), proposed in Kiefer and Wolfowitz (1956) (3), is defined as

$$\widehat{G}^{\mathrm{ML}} := \operatorname*{argmax}_{G \,\mathrm{on}\, [0,\theta_*]} \prod_{i=1}^{n} f_G(X_i). \tag{3}$$

It has been recently shown that the minimax lower bound on the rate relative to the 1-Wasserstein distance is sub-polynomial. To state it, we preliminarily recall the definition of the 1-Wasserstein distance. For any pair $G_1, G_2 \in \mathscr{P}_1(\mathbb{R})$, the latter being the set of all probability measures on $\mathbb{R}$ with finite first moment $m_1(G) := \int_{\mathbb{R}} u \, \mathrm{d}G(u) < \infty$, the 1-Wasserstein distance is defined as

$$W_1(G_1, G_2) := \inf_{\gamma \in \Gamma(G_1, G_2)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \gamma(\mathrm{d}x, \mathrm{d}y),$$

where $\Gamma(G_1, G_2)$ is the set of all couplings with marginal distributions $G_1$ and $G_2$. Theorem 2.1 of Han *et al.* (2021) (2) states that the lower bound for nonparametric estimation of the mixing distribution in model (2) with respect to the 1-Wasserstein distance is sub-polynomial,

$$\inf_{\widehat{G}} \sup_{G \,\mathrm{on}\, [0,\theta_*]} \mathsf{E} W_1(\widehat{G}, G) \geq \frac{c}{\log n}, \tag{4}$$

where the infimum is taken over all estimators of distributions $G$ supported on $[0, \theta_*]$ and the constant $c \equiv c(\theta_*)$ only depends on $\theta_*$. Such slow rate is achieved by the NPMLE $\widehat{G}^{\mathrm{ML}}$. If, in fact, there exists $C \geq 1$ such that

$$\frac{1}{w(x)} \leq C^x, \quad x \in \mathbb{N},$$

then, for some $C' \equiv C'(\theta_*, C) > 0$,

$$\sup_{G \,\mathrm{on}\, [0,\theta_*]} \mathsf{E} W_1(\widehat{G}^{\mathrm{ML}}, G) \leq \frac{C'}{\log n},$$

see Theorem 2.2, part (a), of Han *et al.* (2021) (2). By adopting, instead, the *Gaussian-smoothed optimal transport* (GOT) distance $W_1^\sigma$ defined, for any pair $G_1, G_2 \in \mathscr{P}_1(\mathbb{R})$, as the 1-Wasserstein distance between their convolutions with the centered Gaussian distribution of variance $\sigma^2 > 0$, denoted by $\mathscr{N}_\sigma$ with density $\varphi_\sigma$,

$$W_1^\sigma(G_1, G_2) := W_1(G_1 * \mathscr{N}_\sigma, G_2 * \mathscr{N}_\sigma),$$

DRAFT

under some conditions on $w(\cdot)$, the rate of convergence of the NPMLE $\widehat{G}^{\mathrm{ML}}$ is improved to a *polynomial* rate,

$$\sup_{G \text{ on } [0,\theta_*]} \mathsf{E} W_1^{\sigma}(\widehat{G}^{\mathrm{ML}}, G) \leq C'' n^{-\eta}, \tag{5}$$

for constants $C \equiv C(\sigma, \theta_*, w)$ and $\eta \equiv \eta(\theta_*, w)$ only depending on $\{\sigma, \theta_*, w\}$ and $\{\theta_*, w\}$, respectively, see Theorem 3.1 of Han *et al.* (2021) (2). A motivation for using the Gaussian smoothing when studying the mixing distribution estimation problem is provided in Remark 3.2 of Han *et al.* (2021) (2). The aim of this note is to show that, by adopting a Bayesian nonparametric approach to the problem of estimating the true mixing distribution $G_0$ with respect to the GOT distance, also the Bayes' estimator can attain a polynomial rate. In Section 2, we state a sufficient condition on the frequentist asymptotic behaviour of the posterior distribution corresponding to a prior law on the mixing distribution so that the entailed Bayes' estimator of $G_0$, the posterior expected distribution, attains a polynomial rate. This condition is illustrated for Poisson mixtures. Final remarks are exposed in Section 3.

## 2. Main Result

Let $X^{(n)}$ be a sample of $n$ i.i.d. observations drawn from the true data generating probability measure $P_0$, with probability mass function $f_{G_0}$ as in (2). We want to estimate $G_0$ taking a Bayesian nonparametric approach. Let $\Pi_n$ be a prior law on the set of probability measures on $[0, \theta_*]$ and let $\Pi_n(\cdot \mid X^{(n)})$ be the posterior distribution

$$\Pi_n(B \mid X^{(n)}) = \frac{\int_B \prod_{i=1}^n f_G(X_i)\, \mathrm{d}\Pi_n(G)}{\int \prod_{j=1}^n f_{G'}(X_j) \mathrm{d}\Pi_n(G')},$$

where

$$f_G(\cdot) = \int_0^{\theta_*} p(\cdot; \theta)\, \mathrm{d}G(\theta)$$

is the model probability mass function. Let

$$\widehat{G}^{\mathrm{B}} := \mathsf{E}[G \mid X^{(n)}]$$

be the Bayes' estimator of $G_0$, the posterior expectation of $G$. Our goal is to show that, under the GOT distance, also the speed of convergence of the Bayes' estimator $\widehat{G}^{\mathrm{B}}$, as that of the NPMLE $\widehat{G}^{\mathrm{ML}}$, is accelerated to a polynomial rate, from the sub-polynomial rate relative to the standard 1-Wasserstein distance. To the aim, we state the following assumption, considered also in Han *et al.* (2021) (2), on the tail behaviour of $\{1/w(x)\}_{x \in \mathbb{N}}$, which describes either a polynomial or an exponential decay rate for $\{w(x)\}_{x \in \mathbb{N}}$.

### Assumption A

There exist constants $c_i, C_i > 0$, $i = 1, 2, 3$, such that either

$$c_1 c_2^x \leq \frac{1}{w(x)} \leq C_1 C_2^x \quad \text{for every } x \in \mathbb{N} \tag{6}$$

or

$$c_1 c_2^x x^{c_3 x} \leq \frac{1}{w(x)} \leq C_1 C_2^x x^{C_3 x} \quad \text{for every } x \in \mathbb{N}. \tag{7}$$

**Theorem 1.** *Suppose that Assumption A holds. If for $\tau > 0$ and sufficiently large $M > 0$,*

$$\mathsf{E}\Pi_n(d(f_G, f_{G_0}) > M n^{-1/2} (\log n)^{\tau} \mid X^{(n)}) \to 0 \tag{8}$$

*exponentially fast, where $d$ can be either the Hellinger or the $L^1$-distance, then, for positive constants $C_0 = C_0(\sigma, \theta_*, w, G_0)$ and $\eta_0 = \eta_0(\theta_*, w, G_0) < 1/2$,*

$$\mathsf{E} W_1^{\sigma}(\widehat{G}^{\mathrm{B}}, G_0) \leq C_0 n^{-\eta_0}. \tag{9}$$

DRAFT

*Proof.* Set $\varepsilon_n := C_0 n^{-\eta_0}$, since $W_1$ is convex with respect to each argument, we have

$$\begin{aligned} W_1^\sigma(\widehat{G}^{\mathrm{B}}, G_0) &\leq \int W_1^\sigma(G, G_0)\, \mathrm{d}\Pi_n(G \mid X^{(n)}) \\ &\leq \varepsilon_n + 2\theta_* \times \Pi_n(W_1^\sigma(G, G_0) > \varepsilon_n \mid X^{(n)}), \end{aligned}$$

where the last line follows from

$$W_1^\sigma(G, G_0) \leq W_1(G, G_0) = \|F_G - F_{G_0}\|_1 \leq 2\theta_*,$$

with $F_G$ and $F_{G_0}$ being the distribution functions of $G$ and $G_0$, respectively. It follows that

$$\mathsf{E} W_1^\sigma(\widehat{G}^{\mathrm{B}}, G_0) \leq \varepsilon_n + 2\theta_* \times \mathsf{E}\Pi_n(W_1^\sigma(G, G_0) > \varepsilon_n \mid X^{(n)}).$$

We now study $\mathsf{E}\Pi_n(W_1^\sigma(G, G_0) > \varepsilon_n \mid X^{(n)})$. Set the position $\ell_\sigma := \ell * \varphi_\sigma$, from Step 3 in Theorem 3.1 of Han *et al.* (2021) (2) with $\widehat{Q}$ and $Q$ replaced by $G$ and $G_0$, respectively, by the dual representation of $W_1$ for probability measures with bounded support,

$$W_1(G_1, G_2) = \sup_{\ell \in 1\text{-Lip}} \int_0^{\theta_*} \ell(\mathrm{d}G_1 - \mathrm{d}G_2),$$

where the supremum is taken over all Lipschitz functions $\ell : [0, \theta_*] \to \mathbb{R}$ with constant equal to 1, we have

$$\begin{aligned} W_1^\sigma(G, G_0) \leq \sup_{\ell \in 1\text{-Lip}: \ell(0)=0} &\int_0^{\theta_*} \Big\{ \ell_\sigma(\theta) - \ell_\sigma(0) - \sum_{x=0}^{2k} b_x p(x; \theta) \Big\} (\mathrm{d}G - \mathrm{d}G_0)(\theta) \\ &+ \int_0^{\theta_*} \sum_{x=0}^{2k} b_x p(x; \theta) (\mathrm{d}G - \mathrm{d}G_0)(\theta) \\ =: I + II, \end{aligned}$$

where, by Step 1 of the same theorem, for any $\sigma > 0$, integer $k \geq 2$ and $\ell \in 1$-Lip on $[-\theta_*, \theta_*]$ with $\ell(0) = 0$, there exist $C_4 \equiv C_4(\theta_*, \sigma) > 0$ and $\{b_x \in \mathbb{R}, x = 0, \ldots, 2k\}$ so that

$$I \leq C_4 \Big\{ (\theta_*^{-1} \sigma \sqrt{ek})^{-k} + \sum_{x \geq k+1} w(x) \theta_*^x \Big\},$$

while, for $\|b\|_\infty := \max_{0 \leq x \leq 2k} |b_x|$,

$$II \leq \left| \sum_{x=0}^{2k} b_x [(f_G - f_{G_0})(x)] \right| \leq \|b\|_\infty \times \|f_G - f_{G_0}\|_1 \leq \|b\|_\infty \times d_{\mathrm{H}}(f_G, f_{G_0}),$$

where $d_{\mathrm{H}}(f_G, f_{G_0}) := \|\sqrt{f_G} - \sqrt{f_{G_0}}\|_2$ is the Hellinger distance. Hence,

$$W_1^\sigma(G, G_0) \leq C_4 (\theta_*^{-1} \sigma \sqrt{ek})^{-k} + \sum_{x \geq k+1} w(x) \theta_*^x + \|b\|_\infty \times d(f_G, f_{G_0}). \tag{10}$$

By the upper bound on $\|b\|_\infty$ given in Step 2 of Theorem 3.1 of Han *et al.* (2021) (2), for any $R \in (\theta_*, \theta_r)$ and a suitable constant $C_9 > 0$, for any $k \in \mathbb{N}$,

$$W_1^\sigma(G, G_0) \leq C \times \begin{cases} (\theta_*/R)^k + C_9^{2k} d(f_G, f_{G_0}), & \text{under condition (6)}, \\ k^{-c_3 k} + k^{2C_3 k} d(f_G, f_{G_0}), & \text{under condition (7)}, \end{cases} \tag{11}$$

where the constant $C > 0$ may be different depending on the condition in force. Since, by assumption,

$$d(f_G, f_{G_0}) \leq M n^{-1/2} (\log n)^\tau$$

on a set of posterior probability tending to one, if, for some $0 < \delta < 1/2$, we choose $C_9^{2k} = O(n^\delta)$ under condition (6) or $k^{2C_3 k} = O(n^\delta)$ under condition (7), then $W_1^\sigma(G, G_0)$ is bounded above by $\varepsilon_n$ and the assertion follows.

□

DRAFT

Theorem 1 shows that, under the GOT distance, the Bayes' estimator $\widehat{G}^{\mathrm{B}}$ for $G_0$, like the NPMLE $\widehat{G}^{\mathrm{ML}}$, can attain a polynomial speed of convergence, thus improving the slower sub-polynomial rate relative to the standard 1-Wasserstein distance. The key step in the proof is inequality (10), which, besides allowing to reduce the problem of deriving an upper bound on the speed of $\widehat{G}^{\mathrm{B}}$ to that of assessing the $L^1$-norm posterior contraction rate at $f_{G_0}$, can be directly used to derive the rate of convergence for the NPMLE, as hereafter illustrated in the case of Poisson mixtures.

### Example: Poisson mixtures

For the Poisson mixture model

$$f_G(x) = \int_0^{\theta_*} \frac{e^{-\theta}\theta^x}{x!}\,\mathrm{d}G(\theta), \quad x \in \mathbb{N}_0,$$

the NPMLE in (3) has a unique solution with at most $n$ atoms, see Simar (1976) (5). It has been recently established that

$$\sup_{G\in\mathscr{G}} \mathsf{E} d_{\mathrm{H}}^2(f_{\widehat{G}^{\mathrm{ML}}}, f_G) \leq Cn^{-1} \times \begin{cases} \dfrac{\log n}{\log\log n}, & \mathscr{G} = \{G : \operatorname{supp}(G) \text{ is compact}\}, \\ \log n, & \mathscr{G} = \{G : \int e^{cu}\,\mathrm{d}G(u) \leq 1\}, \end{cases}$$

where both upper bounds are minimax-optimal with exact logarithmic factors, see Polyanskiy and Wu (2020) (4). Then, by the inequalities in (11), we immediately get the upper bound in (5). Besides, under suitable conditions on the prior, the posterior $L^1$-norm contraction rate is, up to a logarithmic factor, almost parametric so that Theorem 1 goes through to the corresponding Bayes' estimator $\widehat{G}^{\mathrm{B}}$, which has polynomial rate with respect to the GOT distance.

## 3. Final Remarks

Theorem 1 shows that, under the GOT distance, also the Bayes' estimator, as the NPMLE, for the mixing distribution of discrete exponential family mixtures can attain a polynomial rate, improving the slower sub-polynomial rate under the standard 1-Wasserstein distance. We believe the actual rate is, up to a logarithmic factor, parametric, but a proof is still elusive to us at the moment. The problem with the above proof comes from the term $C_9^{2k}$ or $k^{2C_3k}$, depending on tail behaviour of $\{w(x)\}_{x\in\mathbb{N}}$, which arises when bounding $\|b\|_\infty$ and slows down the overall rate of the product $\|b\|_\infty \times d(f_G, f_{G_0})$.

DRAFT

## References


[1] Efron, B.: Exponential Families in Theory and Practice. Institute of Mathematical Statistics Textbooks. Cambridge University Press (2022)

[2] Han, F., Miao, Z., Shen, Y.: Nonparametric mixture MLEs under Gaussian-smoothed optimal transport distance. *Arxiv preprint arXiv:2112.02421* (2021)

[3] Kiefer, J., Wolfowitz, J.: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. The Annals of Mathematical Statistics **27**(4), 887 – 906 (1956)

[4] Polyanskiy, Y., Wu, Y.: Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *Arxiv preprint arXiv:2008.08244* (2020)

[5] Simar, L.: Maximum likelihood estimation of a compound Poisson process. The Annals of Statistics **4**(6), 1200 – 1209 (1976)

# Bayesian nonparametric estimation of heterogeneous intrinsic dimension via product partition models

Francesco Denti[a], Antonio Di Noia[b,c], and Antonietta Mira[c,d]

[a]Università Cattolica del Sacro Cuore, Milan, Italy
[b]ETH Zurich, Zurich, Switzerland
[c]Università della Svizzera italiana, Lugano, Switzerland
[d]Università dell'Insubria, Varese, Italy

DRAFT

**Abstract**

The intrinsic dimension (id) of a dataset conveys essential information regarding the complexity of the underlying data-generating process. In particular, it describes the dimensionality of the latent manifold on which the data-generating probability distribution has support. Complex datasets may be characterized by multiple manifolds having different ids. To properly estimate these heterogeneous ids, a recent modeling approach uses finite scale mixtures of Pareto distributions aided by a homogeneity-inducing term in the likelihood. In this contribution, we explore a different modeling perspective, estimating Pareto's scale mixtures via spatial product partition models. We present the general idea and introduce Spider, our Bayesian nonparametric approach. Finally, we showcase some encouraging preliminary results.

***Keywords:*** Intrinsic dimension, Hidalgo, spatial dependence, product partition models.

## 1. Introduction

The estimation of the intrinsic dimension (id) of a dataset is an essential step for many data analyses involving any dimensionality reduction steps. By reliably knowing the dimension of the latent manifold embedding the data, dimensionality reduction can be efficiently carried out, avoiding any loss of information. The literature on dimensionality reduction is vast, and a numerous methods address the id estimation problem from many different perspectives: see, for example, [4] for a review. Here, we focus on likelihood-based id estimation methods, which were pioneered by [10].

Recently, [7] introduced the `TWO-NN` model, an id estimator based on the distributional properties of the ratios of distances between each data point and its first and second nearest neighbors (NNs). In particular, under the hypothesis that the data are generated from a locally homogeneous Poisson point process, the ratios are Pareto distributed with unitary shape parameter and scale parameters equal to the id, respectively. Modeling extensions have been proposed, for example, by [6], who considered ratios of NNs of generic order, and [11], who explore the id estimation employing discrete metrics. As the majority of the proposals in the literature, these methods are devoted to estimating a single, homogeneous id for the entire dataset (sometimes obtained as the ensemble of point-wise id estimates).

The Bayesian heterogeneous id algorithm (Hidalgo) [1] extends this framework, allowing for multiple manifolds in the same dataset, each characterized by its own id. Hidalgo segments the data via a

model-based clustering approach, modeling the ratios of NN distances as a mixture of Pareto distributions. Hidalgo postulates an additional term in the likelihood to overcome the difficulties of clustering solely based on the scale parameters of highly overlapping Pareto distributions. This extra term, obtained by modeling the adjacency matrix of the data, is employed to introduce spatial information in the model estimation. The rationale for its usage is that points close to each other should be more likely to be sampled from the same manifold and, consequently, should contribute to estimating the id value of that same manifold.

Nonetheless, the local homogeneity term is mainly motivated by computational convenience rather than by realistic modeling assumptions. In this contribution, we propose to modify Hidalgo to obtain a more coherent inferential framework using product partition models (PPM) [9; 14]. In particular, we adopt the spatial extension of PPM introduced in [13] to estimate a clustering solution that is spatially informed. Our article proceeds as follows. In the next section, we briefly describe Hidalgo, discuss its structure, and then delineate our modeling proposal. Next, we show some preliminary results on simulated data in Section 3. Finally, we discuss potential future directions in Section 4.

## 2. Model specification

Consider a dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top$ where each observation $\boldsymbol{x}_i$ is embedded in an ambient space of dimension $D$ but lies on a manifold of id $d$. In other words, $\boldsymbol{X}$ has $D$ columns but each row $\boldsymbol{x}_i$ is generated from a $d$-dimensional probability density function. Let $r_{i,1}$ and $r_{i,2}$ be the distances between the $i$-th point and its first and second NNs, respectively. In [7], it is proved that $\mu_i = r_{i,2}/r_{i,1} \sim Pareto(1, d)$, for $i = 1, \dots, n$. This result is the foundation of the TWO-NN id estimator. One can generalize this framework, supposing that the data are actually generated from a mixture of densities supported on manifolds with different ids. Then, assuming independence between $\mu_i$'s the joint distribution of the random vector of ratios $\boldsymbol{\mu}$ is given by a mixture of Pareto distributions:

$$f(\boldsymbol{\mu}|\boldsymbol{p}, \boldsymbol{d}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k d_k \mu_i^{-d_k - 1}, \tag{1}$$

where $\boldsymbol{d}$ is a vector of different ids associated with the $K$ manifolds and $\boldsymbol{p}$ is a vector of mixture weights.

To make the estimation of the model in Equation (1) feasible, we can introduce the latent membership labels $\boldsymbol{z} = (z_1, \dots, z_n)$, where $z_i \in \{1, \dots, K\}$ for $i = 1 \dots, n$ and $z_i = k$ implies that the $i$-th observation is generated by the $k$-th mixture density. It is clear that multiple observations can be assigned to the same mixture component. In doing so, we can identify a partition of the $n$ data points $\rho_n = \{S_k\}_{k=1}^K$, where $S_j = \{i : z_i = k\}$, $\cup_{k=1}^K S_k = \{1, \dots, n\}$, and $S_i \cap S_j = \emptyset$ for $i \neq j$. Ideally, each $S_k$ would contain points belonging to the same latent manifold. We can rewrite model (1) as

$$f(\boldsymbol{\mu}|\boldsymbol{z}, \boldsymbol{d}) = \prod_{i=1}^{n} d_{z_i} \mu_i^{-d_{z_i} - 1}, \qquad f(\boldsymbol{z}|\boldsymbol{p}) = \prod_{i,k} p_k^{\mathbb{1}\{z_i = k\}}. \tag{2}$$

Identifying the different mixture components is complicated since the kernel Pareto densities, despite having different scale parameters, overlap to a great extent.

### 2.1 Hidalgo

To overcome the identification problem, [1] considered an additional source of information by jointly modeling the vector of ratios $\boldsymbol{\mu}$ and the $n \times n$ binary proximity matrix $\mathcal{N}^{(q)}$ computed from $\boldsymbol{X}$, where $\mathcal{N}_{ij}^{(q)} = \mathbb{1}\{j \in \Omega_i^q\}$ and $\Omega_i^q$ is the set of first $q$ NNs of unit $i$. Then, we assume $f(\mathcal{N}_{ij}^{(q)} = 1 | z_i = z_j) \propto \zeta$ and $f(\mathcal{N}_{ij}^{(q)} = 1 | z_i \neq z_j) \propto 1 - \zeta$, where $\zeta > 0.5$, i.e., points assigned to the same manifold have more chances to be neighbors. This modeling choice introduces local homogeneity. The resulting likelihood becomes

$$\mathcal{L}(\boldsymbol{\mu}|\boldsymbol{d}, \boldsymbol{z}, \zeta) = \prod_{i=1}^{n} d_{z_i} \mu_i^{-d_{z_i} - 1} \cdot f\left(\mathcal{N}_i^{(q)} | \boldsymbol{z}, \zeta\right). \tag{3}$$

DRAFT

The Hidalgo model is completed by conjugate prior specifications for the parameters $\boldsymbol{d}$ and $\boldsymbol{p}$. Inference is carried out via Gibbs sampler (see the R package `intRinsic` [5] for an example of implementation).

Adding the homogeneity-inducing term in the likelihood is appealing since it introduces a spatial dependence in the model a posteriori. This addition greatly improves the identification of manifolds without compromising too much the feasibility of the posterior computations. The modeling assumption in (3) is over the neighborhood structure, *given the membership labels*. However, a more natural and coherent option that motivates our proposal is represented by the converse: modeling the distribution of membership labels informed by the spatial structure, e.g., by specifying $f(\boldsymbol{z}|\mathcal{N}^{(q)})$.

## 2.2 Spatial PMM for ID estimation using NNs ratios

We propose a novel method to fit a mixture model for heterogeneous id estimation that incorporates spatial information using product partition models, which we now briefly introduce. First, recall that there is a bijective map $g : \mathcal{P}^n \to \{1, \dots, K_n\}^n$ where $\mathcal{P}^n$ is the space of partitions of $n$ data points and $K_n$ is the number of distinct clusters which is inferred from the data. In other words, $g$ maps the partition of the observations $\rho_n$ to the set of membership labels so that $\boldsymbol{z} = g(\rho_n)$. Following [9; 14], we specify a product partition prior distribution on $\rho_n = \{S_1, \dots, S_{K_n}\}$. Notice that we can avoid the ex-ante specification of a fixed number of clusters $K$ since we are working directly on the space of partitions. Formally, we define the PPM prior as

$$f(\rho_n) \propto \prod_{k=1}^{K_n} C(S_k), \tag{4}$$

where $C(\cdot)$ is a generic, non-negative *cohesion* function measuring how likely it is that elements of the cluster $S_k$ are, a priori, grouped together. In other words, a well-defined cohesion is expected to assign higher probability to partitions that display homogeneous and well-separated clusters. A well-known example of cohesion is $C(S_k) = (|S_k| - 1)!M$, which implies the same clustering behavior induced by a Dirichlet Process [8] with concentration parameter $M$. This modeling framework is characterized by extreme flexibility since the cohesion can be tailored according to the specific application setting. For example, it can be a function of exogenous covariates: see, for example, the PPMx models [12]. Here, we follow [13] and specify a location-aware cohesion function. Let $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$ denote the $n$ spatial coordinates of the data points in $\boldsymbol{X}$, and let $\boldsymbol{s}_k^\star = \{\boldsymbol{s}_i : i \in S_k\}$. The cohesion function for a spatial PPM becomes

$$f(\rho_n) \propto \prod_{k=1}^{K_n} C(S_k, \boldsymbol{s}_k^\star).$$

In particular, we consider the following cohesion:

$$C(S_k, \boldsymbol{s}_k^\star) = \frac{M\Gamma(|S_k|)}{\Gamma(\alpha\mathcal{D}_k)\mathbb{1}\{\mathcal{D}_k \geq 1\} + \mathcal{D}_k\mathbb{1}\{\mathcal{D}_k < 1\}}\mathbb{1}\{|S_k| > 1\} + M\,\mathbb{1}\{|S_k| = 1\}, \tag{5}$$

where $\mathcal{D}_k = \sum_{i \in S_k} d(\boldsymbol{s_i}, \bar{\boldsymbol{s}}_k)$ and $d(\cdot)$ denotes the Euclidean distance. The quantity $\mathcal{D}_k$ measures the overall dispersion of the members of the $k$-th cluster around its centroid $\bar{\boldsymbol{s}}_k$, defined as $\bar{\boldsymbol{s}}_k = |S_k|^{-1}(\sum_{j \in S_k} s_{j,1}, \dots, \sum_{j \in S_k} s_{j,D})^\top$. In the previous formula, $D$ is the number of recorded coordinates (equal to the ambient dimension). Joining the likelihood of ratios of NN distances and the new prior, we obtain

$$\begin{aligned} f(\boldsymbol{\mu}|\boldsymbol{z}, \boldsymbol{d}) &= \prod_{i=1}^{n} d_{z_i}\mu_i^{-d_{z_i}-1}, \quad \boldsymbol{z} = g(\rho_n), \\ p(\rho_n) &\propto \prod_{k=1}^{K_n} C(S_k, \boldsymbol{s}_k^\star), \quad d_k \sim Gamma(a_0, b_0). \end{aligned} \tag{6}$$

We call the model defined in (6) the **S**patial **P**PM for **id** **e**stimation using NNs **r**atios (**Spider**).

DRAFT

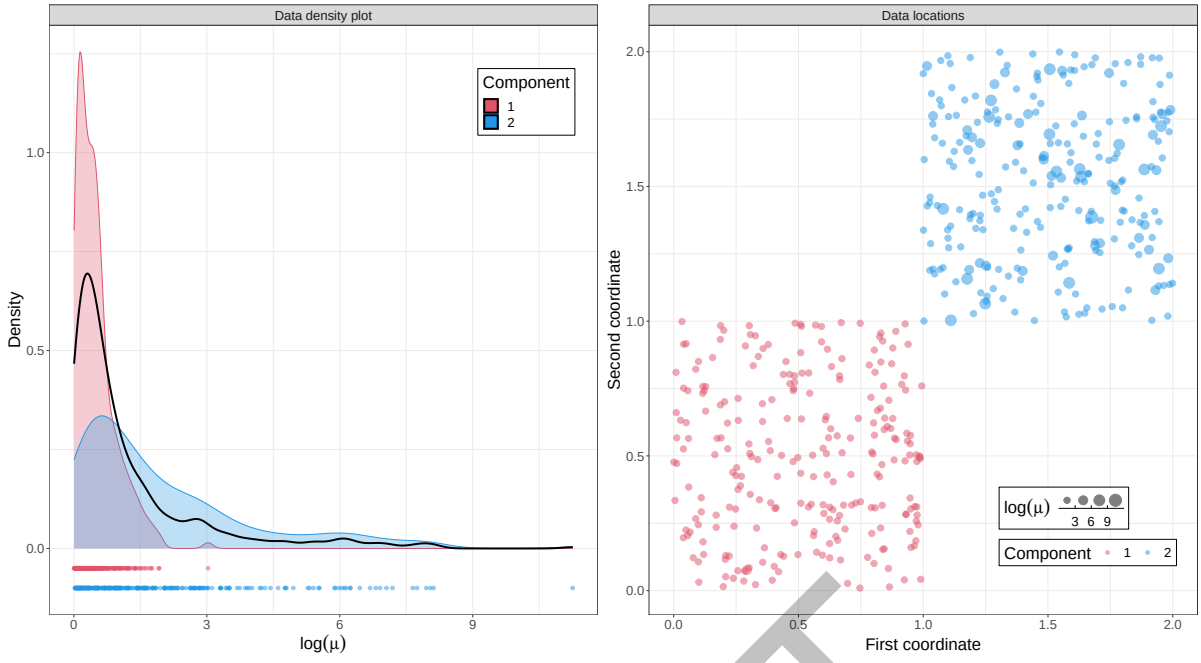



Figure 1: The left panel shows the densities and dots representing the (transformed) simulated data from the two mixture components (in red and blue), and the observed joint density (in black). The right panel shows the spatial coordinates of the data in $D = 2$ dimensions. The two simulated groups of ratios correspond to data well separated in space.

DRAFT

## 3. Applications

We show how Spider works with the help of a simple simulation scenario. We consider a mixture of two Pareto distributions, generating 500 observations each with distribution

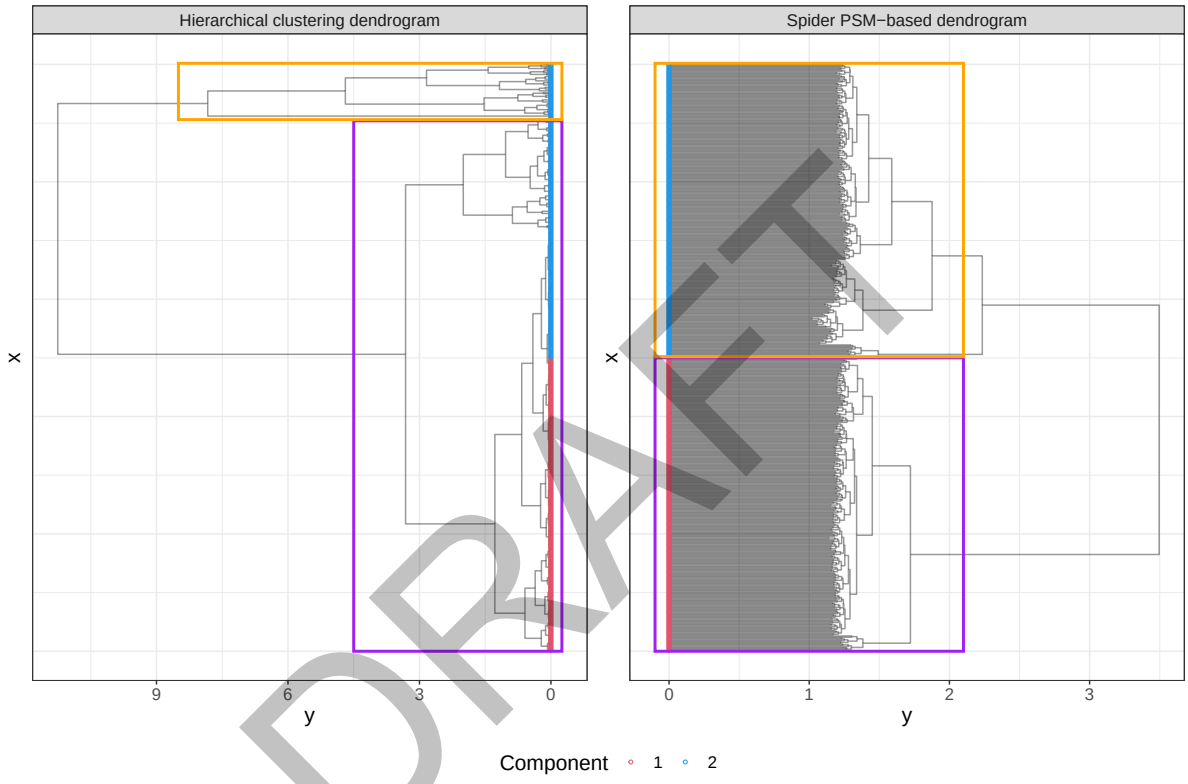
$$\mathcal{L}(\mu_i) = \frac{1}{2} Pareto(1, 0.5) + \frac{1}{2} Pareto(1, 2). \tag{7}$$

The two distributions can originate from a dataset where the observations lie on manifolds with ids 0.5 and 2, respectively. We suppose the two manifolds are embedded in an ambient space of $D = 2$ and well-separated in space. We highlight their coordinates, generated uniformly at random, in the right panel of Figure 1. Despite the simplicity of this distribution, recovering the original clusters is highly challenging. The left panel of Figure 1 shows the true data-generating densities (after log-transforming the data to enhance the visual representation) in blue and red. Once the data are collected, what we observe is only the density function in black. This makes standard clustering algorithms dramatically fail to classify the observations correctly. The left part of Table 1 and the left panel of Figure 2 display the classification results of using a simple hierarchical clustering algorithm on the $\boldsymbol{\mu}$, thresholding the resulting dendrogram to find 2 clusters. The estimated partition is clearly off, as expected.

We then estimate Spider, employing tailored marginal Bayesian nonparametric MCMC sampling techniques. Given the Monte-Carlo chains of cluster allocation parameters, we compute the posterior similarity matrix $PSM = (v_{i,j})_{i,j=1}^{n}$, where each entry corresponds to the percentage of times the MCMC allocated observations $i$ and $j$ in the same cluster. We can construct a dendrogram from the dissimilarity matrix $1 - PSM$, and estimate the partition after thresholding, as we did before. The resulting partition recovers the ground truth very well (see the right part of Table 1 and the right panel in Figure 2).

| | hclust | | Spider | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| Component 1 | 250 | 205 | 248 | 0 |
| Component 2 | 0 | 45 | 2 | 250 |
| $\hat{d}_k$ | 1.281 | 0.181 | 2.152 | 0.529 |

Table 1: Confusions matrixes comparing the ground truth (by rows) and the clustering allocation via hierarchical clustering (columns 1 and 2) and via Spider (columns 3 and 4). The last row contains the estimated id within each cluster (the true id values are 2 and 0.5).




DRAFT

Figure 2: The figure displays the dendrograms constructed using a simple hierarchical clustering approach, using only the ratios $\boldsymbol{\mu}$ to compute the distances (left panel), and the dendrogram resulting from the PSM when applying Spider (right panel).

## 4. Conclusions

In this contribution, we presented some promising results concerning the application of spatial PPM to effectively identify the components of a scale mixture of Pareto distributions when spatial information is available and consequently estimate a reliable id-based partition of the data. Many interesting research directions are worth exploring. First, one could devise a novel cohesion function to use when the ambient dimension $D$ is large to ensure the efficiency of the estimating algorithm while preserving the location information. Moreover, one can tackle the introduction of spatially informed mixtures from other perspectives. For example, one could consider Hidden Markov Random Fields to specify the distribution of the membership labels [2; 3].

## References


[1] Allegra, M., Facco, E., Denti, F., Laio, A., and Mira, A. (2020). Data segmentation based on the local intrinsic dimension. *Scientific Reports*, 10(1):1–27.

[2] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

[3] Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48:259–302.

[4] Campadelli, P., Casiraghi, E., Ceruti, C., and Rozza, A. (2015). Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering*, 2015.

[5] Denti, F. (2021). intRinsic: an R package for model-based estimation of the intrinsic dimension of a dataset.

[6] Denti, F., Doimo, D., Laio, A., and Mira, A. (2021). Distributional Results for Model-Based Intrinsic Dimension Estimators. *ArXiv Preprint*.

[7] Facco, E., D'Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1).

[8] Ferguson, T. S. (1973). A Beyesian Analysis of Some non-parametric problems. *The Annals of Statistics*, 1(2):209–230.

[9] Hartigan, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods*, 19(8):2745–2756.

[10] Levina, E. and Bickel, P. J. (2005). Maximum Likelihood Estimation of Intrinsic Dimension. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press.

[11] Macocco, I., Glielmo, A., Grilli, J., and Laio, A. (2022). Intrinsic dimension estimation for discrete metrics.

[12] Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

[13] Page, G. L. and Quintana, F. A. (2016). Spatial product partition models. *Bayesian Analysis*, 11(1):265–298.

[14] Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(2):557–574.


DRAFT

# Bayesian nonparametric multiple change point detection for time series of compositional data

Edoardo Marchionni[a] and Riccardo Corradin[b]

[a]Dipartimento di Matematica, Politecnico di Milano, Italy; edoardo.marchionni@gmail.com
[b]School of Mathematical Sciences, University of Nottingham, UK;
riccardo.corradin@nottingham.ac.uk

**Abstract**

We perform change-point detection of time series of compositional data from a Bayesian nonparametric perspective. In particular, we build a model to infer change points relying on product partition models approach. We consider a combinatorial prior and an underlying diffusion process for simplex-supported time-dependent realizations. This extends one of the main approaches of Bayesian nonparametric change point detection to compositional time series, avoiding mapping the data to spaces of real numbers. An application to Italian Covid-19 data will be further investigated.

***Keywords:*** Bayesian nonparametric, compositional data, time series, change point detection

DRAFT

## 1. Introduction

Compositonal time series are of great interest in applications and can arise in different fields, such as biology, chemistry, ecology, epidemiology and finance. The models to analyse this type of datum are targeted with respect to the purpose of the analysis. In our case, our aim is to perform (multiple) change point detection, that is, to identify the time instants when the underlying generating diffusion process changes its governing parameters and a new regime occurs. In particular, the objects of inference are both the number of change points and their location.

We rely on a popular approach introduced for the first time in (2; 3): product partition models. We assume that observations belong to the different regimes by means of the latent ordered composition induced by the unique values of their parameters, i.e.

$$\forall i \in A_j \quad \boldsymbol{\theta}_{t_i} = \boldsymbol{\theta}^*_j,$$

where the sets $A_1, \ldots, A_k$ are sets of indices representing the different regimes, $\boldsymbol{\theta}_{t_i}$ is the parameter corresponding to the observation at time $t_i$ and $\boldsymbol{\theta}^*_j$ is the unique value of the latent parameter shared by observations in regime $j$. In particular, these sets are nonempty and represent a partition of $[\![n]\!] = \{1, \ldots, n\}$, where $n$ is the number of observations, subject to a monotone increasing constraint, namely:

1. $A_l \cap A_m = \emptyset$ for $1 \leq l, m \leq n$;
2. $\cup_{i=1}^{k} A_i = [\![n]\!]$;
3. if $i \in A_l$ and $h \in A_m$, with if $l > m$, then $i > h$.

We indicate with $\rho_n$ the vector storing such sets, i.e. $\rho_n = (A_1, \ldots, A_k)$. The observations in different groups, conditioned to the regime-specific parameter $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_k$, are assumed independent. Moreover, the block-specific parameters conditioned to the particular partition are distributed a priori independently,

meaning that the behaviour of the time series in some specific regime does not depend on the change-regime process. Finally, a prior distribution is assumed on the partition $\rho_n$ of the form

$$\mathcal{L}(\rho_n) = K \prod_{j=1}^{k} c(A_j),$$

where $K$ is a normalization constant and $c(\cdot)$ are some cohesion functions.

Despite the authors referring to the above model as *parametric* product partition model, since we drive the change-point detection on the underlying parameters of the generating process, this suggests that we are in a Bayesian nonparametric setting. It was indeed pointed out by Quintana and Igleasias in (18) the equivalence between product partition models with a particular cohesion functions and infinite mixture models driven by the Dirichlet process (5; 6).

In this framework, in 2014 Martínez and Mena (13) proposed a multiple change-point detection product partition model for univariate time series using as cohesion function the exchangeable partition probability function arising from Pitman-Yor process (17) restricted to the suitable partition space (see Section 2). Later, Corradin et al. (4) considered a multivariate extension of such a modelling strategy. Among the different applications presented, a compositional time series of proportions of new daily cases of Covid-19 in Italy per three different macro-areas was analyzed. To be able to perform change-points detection with the developed model a log-ratio transformation was applied to proportions, in order to obtain an $\mathbb{R}^d$-supported time series. In this paper, we propose a similar model to those previous two, with the aim of avoiding mapping data to spaces of real numbers, while performing multiple change-point detection in compositional time series directly on simplex spaces.

## 2. Proposed model

Our data structure is a time series of compositional data: we have for $t \in \mathbb{R}_+$ an underlying random vector $\boldsymbol{Y}_t \in \Delta_{d-1} \subset \mathbb{R}_+^d$ *a.s.*, distributed as some continuous diffusion process, where $\Delta_{d-1} = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^d : x_i \geq 0, \sum_{i=0}^{d} x_i = 1\right\}$ denotes the $(d-1)$-dimensional simplex . The number of realizations observed are finite, hence we can index them by means of a subscript $i \in \{1, \ldots, n\}$, getting finally as observed data $\boldsymbol{y}_{t_1}, \ldots, \boldsymbol{y}_{t_n}$.

As anticipated, the main purpose of this paper is to adapt the model developed in (13; 4) to compositional time series, avoiding transformations of the data. With this in mind, the most natural choice is to select a diffusion process supported on the simplex. Restricting ourselves to homogeneous diffusions that are Dirichlet distributed when stationarity is reached, we chose the Wright-Fisher diffusion (7; 8; 12; 20). This process is governed by a vector of parameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d) \in \mathbb{R}_+^d$ that corresponds to the parameter of the Dirichlet distribution reached at stationarity. The choice was driven first and foremost by the analytical tractability of the transition densities, hence, in the case of the Wright-Fisher continuous process a spectral expansion for arbitrary dimensions is available (10). We will rely on this expansion in the sampling scheme.

Following the above argumentation, we assume the diffusive process generating our data is made of an unknown number $k$ of Wright-Fisher regimes, which are identified by unique values $\boldsymbol{\omega}_1^*, \ldots, \boldsymbol{\omega}_k^*$ of the generating parameters $\boldsymbol{\omega}_{t_1}, \ldots, \boldsymbol{\omega}_{t_n}$ associated to each datum. All the above-specified hypothesis underlying product partition models are assumed: observations belonging to different regimes are independent and given the partition, the group-specific parameters are independent and identically distributed according to some prior distribution $P_0$.

As mentioned before, we consider as prior distribution for the latent partition the exchangeable partition probability function (EPPF) arising from Pitman-Yor process (17). Note that for our purposes, the EPPF cannot be straightforwardly used, due to the fact that the class of admissible partitions in our problem are only the ones that preserve the order. In (16), the author re-weights this EPPF with respect to this new combinatorial class, naming this new function *exchangeable random order distribution*. We

DRAFT

report its expression

$$p_n^{(k)}(n_1,\dots,n_k) = \frac{n!}{k!}\frac{\prod_{j=1}^{k-1}(\theta+j\sigma)}{(\theta+1)_{(n-1)}}\prod_{j=1}^{k}\frac{(1-\sigma)_{n_j-1}}{n_j!},$$

for $\sigma \in [0,1)$ and $\theta \in (-\sigma,\infty)$, where $k$ is the number of blocks and $n_j$ is the number of observations in block $j$ and where $(\cdot)_{(\cdot)}$ is the Pochhammer symbol. Furthermore, in order to relax the model specification, we set an hyperprior structure on the parameters $(\theta,\sigma)$ of the restricted Pitman-Yor EPPF.

Summing up, we can compactly write our model as

$$\{\boldsymbol{Y}_{t_i} : i \in A_j\}|\boldsymbol{\omega}_j^*, \rho_n \overset{ind}{\sim} \text{Wright-Fisher}(\boldsymbol{\omega}_j^*) \tag{1}$$

$$\boldsymbol{\omega}_j^* \overset{iid}{\sim} \mathrm{P}_0 \tag{2}$$

$$\rho_n|\theta,\sigma \sim \mathcal{L}(\rho_n|\,\sigma,\theta) \tag{3}$$

$$(\sigma,\theta) \sim \pi(\sigma,\theta). \tag{4}$$

Our underlying diffusion process is Markovian (we refer to (1; 9) for an exhaustive treatise on diffusion processes). Hence, conditioning to the block allocation, the likelihood of each datum depends on the previous one and on the group-specific parameter, except for the very first one and for any changing point. Let denote $n_j$ the cardinality of the block $A_j$, and $m_j$ the index of the last point of cluster $j$, where $m_1 = 0$. The integrated likelihood for each cluster can be expressed as the following product

$$\mathcal{L}(\{\boldsymbol{Y}_{t_i} : i \in A_j\}|\,\boldsymbol{\omega}_j^*, \rho_n) = p\left(\boldsymbol{y}_{t_{m_j+1}}|\boldsymbol{\omega}_j^*\right)\prod_{i=m_j+2}^{m_j+n_j} p\left(t_i - t_{i-1}, \boldsymbol{y}_{t_i}|\boldsymbol{y}_{t_{i-1}}, \boldsymbol{\omega}_j^*\right), \tag{5}$$

where $p(\cdot|\boldsymbol{\omega}_j^*)$ represents the stationary density of the Wright-Fisher diffusion with parameter $\boldsymbol{\omega}_j^*$ , i.e. the Dirichlet distribution, and $p(\cdot,\cdot|\cdot,\boldsymbol{\omega}_j^*)$ is the transition density of the same process. The integrated likelihood of all our data, conditioned to the partition, is just the product of all the cluster integrated likelihood (5):

$$\mathcal{L}(\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n|\boldsymbol{\omega}_1^*,\dots,\boldsymbol{\omega}_k^*,\rho_n) = \prod_{j=1}^{k}\mathcal{L}(\{\boldsymbol{Y}_{t_i} : i \in A_j\}|\,\boldsymbol{\omega}_j^*,\rho_n), \tag{6}$$

where $k$ is the number of blocks in the conditioning partition $\rho_n$.

For what concerns the prior on the block parameters $\boldsymbol{\omega}_j^*$, we assume each component is distributed a priori as a gamma distribution of parameters $\alpha$ and $\beta$:

$$\omega_{j,\ell}^* \overset{iid}{\sim} \Gamma(\alpha,\beta), \qquad \ell = 1,\dots,d$$

whereas as hyperpriors for $\sigma$ and $\theta$, in accordance with (13), we set the following hierarchical structure

$$\sigma \sim \text{beta}(\alpha_\sigma,\beta_\sigma)$$
$$\theta|\sigma \sim s\Gamma(-\sigma;\alpha_\theta,\beta_\theta),$$

where $s\Gamma$ denotes the shifted-gamma distribution. In particular, a random variable $\xi$ is $s\Gamma(-\sigma,\alpha,\beta)$ distributed, if $\xi+\sigma$ is gamma distributed with parameters $\alpha$ and $\beta$.

We underline that in (13) and (4) another likelihood form was derived. In particular, in both papers it is argued that, considering we want to perform inference on the latent partition of the data and on the number of blocks, we do not need explicitly to keep track of the values block specific parameters assume. Authors hence integrate out from the block integrated likelihoods the dependence on $\boldsymbol{\omega}_j^*$, getting finally each block-specific *marginal* integrated likelihood. In the case of the two papers this was rather convenient, since closed-forms of those marginal integrated likelihoods can be found. On the contrary, this is not the case of our model, due to the lack of conjugacy between our prior on the parameters and our block-specific likelihood on data.

DRAFT

## 2.1 Sampling algorithm

Our aim is to perform inference on the latent partition of the data and on the number of regimes. As a consequence, the target distribution is

$$\mathcal{L}\left(\rho_n|\text{rest}\right) \propto \mathcal{L}\left(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n|\boldsymbol{\omega}_1^*,\ldots,\boldsymbol{\omega}_k^*,\rho_n\right)\mathcal{L}\left(\rho_n|\sigma,\theta\right)\prod_{j=0}^{k} P_0\left(\boldsymbol{\omega}_j^*\right), \tag{7}$$

that is, the posterior distribution of our latent partition. In particular, the support of this distribution is, for all $k \in \{1,\ldots,n\}$, all vectors $(n_1,\ldots,n_k) \in \mathbb{N}^k$ such that $\sum_{j=1}^{k} n_j = n$, where $n$ is the number of observed data. Due to the lack of conjugacy, the need to resort to sampling algorithms is obvious. We may think to a general fashion Markov chain Monte Carlo with target distribution the distribution (7) and with state space the above specified support. However, since the the cardinality of this space is growing rapidly as $n$ increases, it is virtually impossible to obtain a representative sample of the posterior. Therefore, we will rely on a different Markov chain Monte Carlo method that is an adjustment to our framework of the one presented by Martínez and Mena in (13). This algorithm relies on a *split & merge* procedure.

At each iteration, we randomly choose according to some probability, except for the boundary cases in which the choice is naturally driven, either to perform a merge or to perform a split; in both cases, we propose a new value for $k$ and a new value for $\rho_n$. Note that, heuristically speaking, those two moves exhaust all the possible moves, due to the ordering proviso of the data, unlike in clustering problems. Those proposed split-merge moves are accepted or rejected according to a Metropolis-Hastings step. Finally, to help the algorithm fasten convergence, at the end of each iteration we perform an additional reshuffle step, in which we reshuffle the dimension of two existing contiguous blocks, accepting this new setting still according to a Metropolis-Hastings step.

The idea underlying our algorithm will be substantially the same as in (13), under a few suitable modifications. In (13) the Metropolis-Hastings steps are driven exploiting the marginal integrated likelihood, but in our case we lack a closed form of transition densities and hence of the marginal integrated likelihood. For this reason, once we propose a new cluster-structure, we introduce a Monte Carlo integration step, in which we estimate the marginal integrated likelihood for the clusters of the new proposed setting. As anticipated, we rely on the spectral expansion available in literature (10) for the transition densities of the Wright-Fisher process.

DRAFT

# 3. Simulated data and algorithm calibration

We use model (1)-(4) with simulated with the purpose of testing it. Moreover, our sampling scheme needs the specification of the level of truncation of the series expansion of the transition densities (10) as well as the number of nodes in the MC integration of the integrated likelihood (6) to approximate the *marginal* integrated likelihood. We investigate how the choice of these two values influences the performance of the sampling algorithm.

Data are approximately simulated from a Wright-Fisher process following the approach presented in (11). We consider 150 observations in $\Delta_2$ belonging to three different regimes of 50 observations each $\rho_{150} = (\{1,\ldots,50\},\{51,\ldots,100\},\{101,\ldots,150\})$. In particular, we set $\boldsymbol{\omega}_1^* = (0.2, 0.5, 1)$, $\boldsymbol{\omega}_2^* = (1, 0.5, 0.5)$ and $\boldsymbol{\omega}_3^* = (0.3, 0.1, 0.2)$ as parameters of the three Wright-Fisher regimes. Moreover, we assume $\alpha = 2$ and $\beta = 2$ as hyperparameters of the priors ($\omega_{j,\ell}^* \overset{iid}{\sim} \Gamma(2,2)$ for $\ell = 1,\ldots,d$). Concerning the hyperpriors, we set $\sigma \sim \text{beta}\,(1,1)$ and $\theta|\sigma \sim s\Gamma(-\sigma;1,1)$.

The point estimate of the posterior cluster allocation is obtained by minimizing the expected value of the posterior loss. We choose the variation of information (14) as loss function and we exploit the optimization algorithm described in (19). Two different summaries are computed in order to assess the quality of the posterior estimate of the partition. From one side, we evaluate how much the posterior partition is different from the true one, using as metric the normalized variation of information (14). From the other side, we inspect the capability of the MCMC chain to produce a good sample of the posterior.

This is done by evaluating the effective sample size of the entropy of the partition that represents the state of the chain at each iteration.

Our simulation study consists in running an MCMC for a grid of the two above-mentioned hyeperparameters, which are the truncation level of the series expansion of the transition densities and the number of integration nodes in the MC integration step. In particular, we truncate the series at the first, second or third term and we consider a number of integration nodes ranging in $\{1, 20, 100, 250, 500\}$. We run 5000 iterations, where 1000 are considered burning-in iterations, on the corresponding grid expansion of the hyperparameters, setting an equal probability either to perform a split or a merge at each iteration. We average the reported results over 5 replications.

Table 1: Table showing the posterior summaries of the simulations on the grid of hyperparameters averaged over 5 replications

| Truncation level | Integration nodes | VI | ESS - entropy |
|---|---|---|---|
| 1 | 1 | 0.438 | 1.787 |
| | 20 | 0.047 | 15.491 |
| | 100 | 0.065 | 96.583 |
| | 250 | 0.068 | 114.859 |
| | 500 | 0.060 | 84.836 |
| 2 | 1 | 0.350 | 1.739 |
| | 20 | 0.049 | 19.280 |
| | 100 | 0.050 | 110.822 |
| | 250 | 0.116 | 110.070 |
| | 500 | 0.052 | 163.858 |
| 3 | 1 | 0.445 | 2.147 |
| | 20 | 0.126 | 16.559 |
| | 100 | 0.050 | 108.452 |
| | 250 | 0.057 | 130.853 |
| | 500 | 0.050 | 148.889 |

DRAFT

Posterior summaries are reported in table 1. We observe an increasing trend of precision of the estimate of the partition that is independent from the level of truncation of the series. Indeed, fixing a value of the number of nodes in MC integration, we appreciate that the normalized variation of information between the estimated partition and the true one gets comparable values with respect to the level of truncation of the series. This is not surprising since the series expansion in (10) presents a multiplicative term that decays exponentially, hence adding an extra-term does not introduce much gain in term of the approximation precision. For what concerns the effective sample size of the partition sample, we report a similar behaviour. Independently from the level of truncation of the series, there is an increasing trend where values for the same number of integration nodes have the same order of magnitude in most cases. We conclude that our model in general behaves as expected on simulated data. The posterior estimate of the partition gets fast precise and the MCMC *split & merge* sampling scheme explores efficaciously the space of possible orderings. In this sense, no advising of nonconvergent behave of the chain was detected.

As further investigation, the model will be applied to proportions of daily cases of Covid-19 by region areas in Italy, as done in (4). In particular, Italian regions are labelled as “north”, “centre” and “south/island”, according to the official grouping of the Italian *Istituto Nazionale di Statistica*, and we build the compositional time series of the proportion of daily new cases in those macro-areas, ending up with a time series in $\Delta_2$. The purpose of this study is double. From one side, the detected change points by our model are compared with the true events that may have had an effect on the regime, enabling to test the capability of the model to perfom change point detection on real data. From the other side, we can compare this approach to the most used approach of transforming the compositional data to spaces of real numbers. For instance, in (4), authors rely on a log-transformation of the data for performing

change point detection on a similar time series.

## References


[1] Baldi, P.: Stochastic calculus, An Introduction Through Theory and Exercises. Springer Cham (1992)

[2] Barry, D., Hartigan, J. A.: Product Partition Models for Change Point Problems. Ann. Stat. (1992) doi: 10.1214/aos/1176348521

[3] Barry, D., Hartigan, J. A.: A Bayesian Analysis for Change Point Problems. J. Am. Stat. Assoc. (1993) doi: 10.2307/2290726

[4] Corradin, R., Danese, L., Ongaro, A.: Bayesian nonparametric change point detection for multivariate time series with missing observations. Int. J. Approx. Reason. (2022) doi: 10.1016/j.ijar.2021.12.019.

[5] Ferguson, T. S.: A Bayesian Analysis of Some Nonparametric Problems. Ann. Statist. (1973) doi: 10.1214/aos/1176342360

[6] Ferguson, T. S.: Prior Distributions on Spaces of Probability Measures. Ann. Statist. (1974) doi: 10.1214/aos/1176342752

[7] Fisher, R. A.: XXI - On the Dominance Ratio. Proc. R. Soc. Edinb. (1923) doi: 10.1017/S0370164600023993

[8] Fisher, R. A.: The evolution of dominance. Biol. Rev. **6**, 345 – 368 (1931)

[9] Gardiner, C.: Stochastic Methods, A Handbook for the Natural and Social Sciences. Springer Berlin, Heidelberg (2009)

[10] Griffiths, R. C.: A Transition Density Expansion for a Multi-Allele Diffusion Model. Adv. Appl. Probab. (1979) doi: 10.2307/1426842

[11] Griffiths, R. C., Li, W. H.: Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. Theor Popul Biol. (1983) doi: 10.1016/0040-5809(83)90003-5

[12] Karlin, S., Taylor, H. M.: A Second Course in Stochastic Processes. Academic Press, New York. (1981)

[13] Martínez, A. F., Ramsés, H. M.: On a Nonparametric Change Point Detection Model in Markovian Regimes. Bayesian Anal. (2014) doi: 10.1214/14-BA878

[14] Meilă, M.: Comparing clusterings - an information based distance. J. Multivar. Anal. (2007) doi: 10.1016/j.jmva.2006.11.013.

[15] Müller, P., Quintana, F. A., Jara, A., Hanson, T.: Bayesian Nonparametric Data Analysis. Springer Cham (2015)

[16] Pitman, J.: Combinatorial Stochastic Processes. Springer Berlin, Heidelberg (2006)

[17] Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. (1997) doi: 10.1214/aop/1024404422

[18] Quintana, F., Iglesias, P.: Bayesian clustering and product partition models. J. R. Stat. Soc. Series B (2003) doi: 10.1111/1467-9868.00402.

[19] Wade, S., Ghahramani, Z.: Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). Bayesian Anal. (2018) doi: 10.1214/17-BA1073

[20] Wright, S.: Evolution in Mendelian Populations. Genetics. (1931) doi: 10.1093/genetics/16.2.97



We acknowledge Prof. A. Guglielmi for her insightful comments and suggestions on early draft of this manuscript.



DRAFT

# Galton-Watson process: a non parametric prior for the offspring distribution

Massimo Cannas[a], Michele Guindani[b], and Nicola Piras[a]

[a]University of Cagliari; massimo.cannas@unica.it, nicola.piras97@unica.it
[b]University of California at Los Angeles; mguindani@ucla.edu

**Abstract**

In this article we propose a non parametric prior for the probabilities of the Galton-Watson process based on the Dirichlet Process. After recalling the main properties of the Galton-Watson process and presenting the estimation methods already present in the literature, such as maximum likelihood and Bayesian conjugate analysis, we define the new prior by pointing out how it is more general than the Dirichlet prior used in the conjugate analysis, which is a special case of our extension. Finally, we show the results of a simulation study illustrating how our analysis leads to a more accurate classification of the process.

***Keywords:*** Galton-Watson process, Dirichlet process, Bayesian inference, offspring distribution.

DRAFT

## 1. Background

The Galton-Watson process is a probabilistic model originally proposed to study genealogy problems, with recent applications to nuclear fission, queuing models, viral phenomena, social networks, neuroscience, and disease spread. It assumes individuals in the population reproduce independently, each individual giving rise to a random number $X$ of descendants according to the *offspring distribution*:

$$P(X = j) = \pi_j \quad j \in \mathbb{S}(X)$$

$$\sum_{j \in \mathbb{S}(X)} \pi_j = 1$$

The offspring distribution holds for all generations, i.e, $\pi_j$ does not depend on the generation number. Given $Z_0$ original ancestors (we always assume that $Z_0 = 1$) the evolution of the population can be described by $\{Z_i \quad i = 0, 1, \dots\}$, where $Z_i$ denotes the total number of individuals in the $i$-th generation. The assumptions above imply that $Z_0, Z_1, \dots$ is a Markov chain, meaning that if we know the size of the $n$-th generation then the probability law of later generations does not depend on the sizes of generations preceding the $n$-th. In particular note that if $Z_i = 0$ for some $i$ then $Z_r = 0 \quad \forall r \geq i$. In case $Z_i = 0$ we say that the population is *extinct* at time $i$.

The extinction probabilities, that is, the probability that the population becomes extinct after $r$ generations given that each individual has $0, 1, \dots$ descendants with probabilities $p_0, p_1, \dots$ according to the offspring distribution, is often the main feature of interest of the process; see Harris (1). In fact, the model was first studied by Francis Galton to analyze the extinction of family names in England at the end of XIX century and later by A.J.Lotka, who replied the analysis using the US census (2).

**Extinction probability** The overall extinction probability is defined as:

$$q = P(Z_i = 0 \quad \text{for some} \quad i)$$

The following results leads to a simple classification of the process.

**Theorem 1.** *The extinction probability $q$ is the smallest root of the equation $f(s) = s$ where $f(s) = \sum_{j \in \mathbb{S}(X)} \pi_j s^j \quad |s| \leq 1$ is the probability generating function of the offspring distribution.*

The theorem highlights the close connection between the extinction probability and the reproduction mean, defined as

$$m = \sum_{j \in \mathbb{S}(X)} j\pi_j$$

Indeed, by simply computing $m$, we can realize whether or not the extinction of the process is certain, as the reproduction mean leads to a simple classification: critical and subcritical processes lead to extinction with probability 1 while supercritical processes lead to extinction with probability $q < 1$.

## 1.1 Inference for Galton Watson process

Inference for GW process naturally focuses on the offspring distribution $\pi$ and its average $m$. The minimum amount of information required to make inference is the vector of population sizes, $\{Z_i \quad i = 0, 1, \dots \ N\}$. We have a more detailed information if we can observe for each generation the number of individuals having $j$ descendants, for all $j = 0, 1, \dots$, that is, if we observe the vector $\mathbf{Z}_c = \{Z_{ij} \quad i = 0, 1, \dots \quad j = 0, 1, \dots, k\} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_N\}$ where $Z_{ij}$ is the number of individuals in the $i$-th generation having exactly $j$ descendants and $\mathbf{Z}_i = (Z_{i0}, \dots, Z_{ik})'$.
Therefore

$$Z_i = \sum_{j=0}^{k} Z_{ij} \quad \text{and} \quad Z_{i+1} = \sum_{j=0}^{k} jZ_{ij}$$

so we have

$$p(\mathbf{Z}_1, \dots, \mathbf{Z}_n | \pi) = \prod_{i=0}^{n} p(\mathbf{Z}_i | Z_i, \pi)$$

where $p(\mathbf{Z}_i | Z_i, \pi) \sim Multinomial(Z_i, \pi)$. The likelihood function is:

$$L(\pi | \mathbf{Z}_c) \propto \prod_{i=0}^{n} \prod_{j=0}^{k} \pi_j^{Z_{ij}} \propto \prod_{j=0}^{k} \pi_j^{Z_{nj}}$$

with $Z_{nj} = \sum_{i=0}^{n} Z_{ij}$ the total number of individuals giving rise to exactly $j$ descendants.

**MLE estimator** The maximum-likelihood estimators of $\pi$ and $m$ are (1):

1) $\hat{\pi}_j = \dfrac{Z_{nj}}{Y_n}$
2) $\hat{m} = \frac{Z_1 + \dots + Z_n}{Z_0 + \dots + Z_{n-1}}$

where $Y_n = \sum_{j=0}^{k} Z_{nj} = \sum_{i=0}^{n} Z_i$ is the number of individuals up to the $n$-th generation.
These estimators are consistent conditional on the non extinction of the process. Notice that $\hat{m}$ is the

DRAFT

total number of children divided by the total number of parents and it is the same whether we observe only the total number of individuals in each generation or the detailed information given by $\mathbf{Z}_c$, i.e. $\hat{m}(\mathbf{Z}_c) = m(\hat{\pi}) = \sum_{j=0}^{k} j\hat{\pi}_j$.

**Bayesian analysis** A Bayesian conjugate analysis for the Galton-Watson process was proposed by Mendoza and Guterrez-Pena (4). These authors considered as prior distribution:

$$\pi \sim Dirichlet(\alpha_0, \ldots, \alpha_k)$$

leading to the posterior distribution $\pi|\mathbf{Z}_c \sim Dirichlet(\beta_0, \ldots, \beta_k)$

where $\beta_j = \alpha_j + Z_{nj}$. If we write $\beta = \sum_{j=0}^{k} \beta_j$, then the posterior distribution of $\pi_j$ is

$$\pi_j|\mathbf{Z}_c \sim Beta(\beta_j, \beta - \beta_j) \equiv Beta\left(\alpha_j + \sum_i Z_{ij}, \sum_{i\neq j} \alpha_i + \sum_{ij} Z_{ij} - \sum_i Z_{ij}\right) \quad (1)$$

so that $E(\pi_j|\mathbf{Z}_c) = \frac{\beta_j}{\beta}$ $\quad Var(\pi_j|\mathbf{Z}_c) = \frac{\beta_j(\beta-\beta_j)}{\beta^2(\beta+1)}$ $\quad Cov(\pi_j, \pi_\ell|\mathbf{Z}_c) = -\frac{\beta_j\beta_\ell}{\beta^2(\beta+1)}$

Notice that the prior requires knowledge of the true support size, $k$, which is assumed finite. An appropriate choice of $\{\alpha_0, \cdots, \alpha_k\}$ can ensure neutrality with respect to either the $\pi_j$'s or the reproduction mean, $m$. The latter is usually the inferential target as it allows the classification of the process. The first two posterior moments of $m$ can be obtained from those of $\pi$, exploiting $m$ is a linear combination of the entries of the vector $\pi$:

$$E(m|\mathbf{Z}_c) = \mathbf{h}'\mu = m(\mu) \quad Var(m|\mathbf{Z}_c) = \mathbf{h}'\Sigma\mathbf{h} = \sigma^2(\mu)/(\beta+1)$$

with $\mathbf{h} = (1, \ldots, k)'$, $\mu$ and $\Sigma$ the posterior mean vector and the variance-covariance matrix of $\pi$. The distribution of $m$ is not available. However, conditional on the non-extinction of the process, a normal approximations is valid for the posterior of $m$. In general, sampling techniques can be useful to approximate its distribution.

DRAFT

## 2. A DP prior on $\pi$

In this section we introduce a flexible prior for inference on $\pi$ based on the Dirichlet Process of Ferguson (3). We recall that $G$ is a $DP(a, G_0)$ if, for any partition $B_0, B_1, \cdots B_k$ of the real line, the vector $G(B_0), G(B_1), \cdots G(B_k)$ follows a Dirichlet distribution of parameter $(aG_0(B_0), \ldots, aG_0(B_k))$. Here $a > 0$ is a concentration parameter regulating confidence on the centering measure $G_0$. An important property of the Dirichlet Process (DP in the following) is that its conditional distribution given a sample from the process is still a DP, i.e., $G|x \sim DP\left(n + a, \frac{1}{n+a}\sum_{i=1}^{n} \delta_{x_i} + \frac{a}{n+a}G_0\right)$ where $n$ is the sample size, $a$ is the prior number of observations, and $\delta_x(\cdot)$ counts the number of $x$ in the set $\{\cdot\}$. The posterior centering measure is thus a weighted average of the prior measure $G_0$ and the empirical distribution function.

We now turn back to the problem of assigning a flexible prior on the set of all possible offspring distributions of a Galton-Watson process. We can induce the prior on $\pi$ by setting

$$\pi_j \sim G(A_j) \quad (2)$$

where $G \sim DP(a, G_0)$. In this context, $G_0$ is a distribution on the positive integers reflecting our prior guess on the offspring distribution; $a$ is a scalar reflecting the degree of confidence on $G_0$ and $A_0, A_1, \cdots A_k$ is any partition such that $A_j$ contains only the j-th natural number. It follows that, a priori

$$\pi_j \sim Beta(aG_0(j), a(1 - G_0(j))$$

and, a posteriori

$$\pi_j|x \sim Beta\left(\frac{n}{n+a}\sum_{i=1}^{n}\frac{1}{n}\delta_{x_i}(i) + \frac{a}{n+a}G_0(A_j), 1 - \frac{n}{n+a}\sum_{i=1}^{n}\frac{1}{n}\delta_{x_i}(A_j) - \frac{a}{n+a}G_0(A_j)\right) \quad (3)$$

The $DP$ prior above generalizes the classic Dirichlet prior proposed by Guterrez-Pena (4) by allowing any offspring distribution $G_0$ with arbitrary degree of confidence $a$. In particular, comparison of posterior distributions in equations 1 and 3, shows that the distribution induced by the $DP$ prior is the same of the classic Dirichlet prior when $G_0$ is a discrete distribution on $\{0, \cdots, k\}$ and $a$ is equal to $\sum_i \alpha_i$.

For posterior inference on $\pi_j$ an important role is played by the posterior average $E(\pi_j|x)$:

$$E(\pi_j|x) = E(G(A_j)|x) = \frac{n}{n+a}\sum_{i=1}^{n}\frac{1}{n}\delta_{x_i}(A_j) + \frac{a}{n+a}G_0(A_j)$$

which is a weighted mean of the prior offspring probability $G_0(A_j) = Prob(G_0 = j)$ and the empirical proportion of type $j$ individuals.

## 3. Simulation

In this section we show the results of a small simulation study comparing the estimates of $\pi$ and $m$ obtained with maximum likelihood and conjugate Bayesian analysis with those obtained with our proposal. For the classic Bayesian estimator we used the Jeffreys' prior on the true support. For the flexible non parametric estimator we use a low value of the concentration parameter $\alpha$ and a discrete uniform distribution on the true support as centering measure $G_0$. In particular we consider two different estimates for $m$ using the DP prior: the first (bayes.dp) obtained from the posterior means of $\pi$ from eq.3 and the second (bayes.dp.mc) obtained as a Monte Carlo estimate by sampling from the a posteriori DP.

DRAFT

Table 1: Subcritical process - average and standard error over 100 MC replications

| [a] | $m$ | $p_0$ | $p_1$ | $p_2$ |
|---|---|---|---|---|
| mle | 0.836 (0.17) | 0.46 | 0.37 | 0.10 |
| bayes.dir | 1.006 (0.01) | 0.40 | 0.31 | 0.16 |
| bayes.dp | 0.941 (0.06) | 0.42 | 0.33 | 0.14 |
| bayes.dp.mc | 0.928 (0.14) | 0.42 | 0.33 | 0.14 |

[a] mle is maximum likelihood estimate; bayes.dir uses Jeffreys prior $(1/2, 1/2, 1/2, 1/2)$; bayes.dp uses the $DP(\alpha, G_0)$ prior with $\alpha = 1$ and $G_0$ discrete uniform on $\{0, 1, 2, 3\}$

In Table 1 we consider a subcritical process with probability vector ($\pi = (0.4, 0.3, 0.2, 0.1)$ leading to $m$=1. We replicated the generation process 100 times and we report average and standard error (in parenthesis for $m$) of estimates over $S = 100$ replications. The simulation shows that the bayes.dp prior performs better than mle and bayes.dir, leading to correct classification of the process.

In Table 2 we consider a supercritical process with probability vector ($\pi = (0.3, 0.3, 0.3, 0.1)$ leading to $m = 1.2$. In this situation the bayes.dp and the bayes.dir perform similarly (and both perform better than mle).

Table 2: Supercritical process - average and standard error over 100 MC replications

| [a] | $m$ | $p_0$ | $p_1$ | $p_2$ |
|---|---|---|---|---|
| mle | 0.954 (0.25) | 0.40 | 0.29 | 0.24 |
| bayes.dir | 1.089 (0.10) | 0.36 | 0.28 | 0.25 |
| bayes.dp | 1.038 (0.15) | 0.37 | 0.28 | 0.25 |
| bayes.dp.mc | 1.052 (0.16) | 0.37 | 0.28 | 0.25 |

[a] mle is maximum likelihood estimate; bayes.dir uses Jeffreys prior $(1/2, 1/2, 1/2, 1/2)$; bayes.dp uses the $DP(\alpha, G_0)$ prior with $\alpha = 1$ and $G_0$ discrete uniform on $\{0, 1, 2, 3\}$

Recall that the classic Bayesian prior bayes.dir is replied by the proposed DP prior bayes.dp when $a = \sum \alpha_i$. Setting $a = 1 < \sum \alpha_i$ leads to better estimate of $m$ when the process is subcritical and prior sample size can heavily influence the posterior mean. Instead, for supercritical processes generating, on average, more individuals the results are very similar. Finally, preliminary results show that bayes.dp is useful also when the offspring distribution has infinite support or the support is finite but its size is unknown.

**Acknowledgments** Research partially supported by the project UniCA/FdS 2020 by Fondazione di Sardegna

## Bibliography

[1] Harris, T.E.: The theory of branching processes. Springer-Verlag, Berlin (1963)
[2] Lotka, A.J.: The extinction of families. J. Wash. Acad. Sci, **21**, 377-380 (1931)
[3] Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems. Ann. Stat. **1**, 209–230 (1973)
[4] Mendoza, M., Gutierrez-Pena, E.: Bayesian conjugate analysis of the Galton-Watson process. Test **9**, 149–171 (2000)

DRAFT

# Hierarchical processes in survival analysis

Riccardo Cogo[a], Federico Camerlenghi[a], and Tommaso Rigon[a]

[a]Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano; r.cogo1@campus.unimib.it, federico.camerlenghi@unimib.it, tommaso.rigon@unimib.it

**Abstract**

Hierarchical processes are popular tools in Bayesian nonparametrics because of their central role within the partial exchangeable framework, which is a natural assumption for modeling different, though related, groups of observations. In this paper, we focus on neutral to the right (NTR) processes, a class of random distributions widely used as nonparametric priors in survival analysis thanks to their conjugate behavior in presence of censored data. After a brief introduction to NTR processes, we focus on their hierarchical extension by leveraging their characterization as functionals of completely random measures. We conclude by showing an example of a hierarchical NTR process, called hierarchical Beta-Stacy.

***Keywords:*** Bayesian nonparametrics, Beta-Stacy process, completely random measures, hierarchical priors, neutral to the right processes, partial exchangeability

DRAFT

## 1. Introduction

In this work, we consider partially exchangeable data, i.e., observations coming from different (though similar) populations, and we define a general class of processes that can be used as nonparametric priors in order to carry out full Bayesian inference under this framework. Let us consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a measurable space $(\mathbb{X}, \mathcal{X})$, where $\mathbb{X}$ is a Polish space with its Borel $\sigma$-algebra $\mathcal{X}$. Let us denote by $\mathbb{P}_{\mathbb{X}}$ the space of all probability measures over $(\mathbb{X}, \mathcal{X})$, and by $\mathbb{P}_{\mathbb{X}}^d$ the corresponding $d$-dimensional product space. Suppose we are provided with $d$ groups of $\mathbb{X}$-valued observations, more precisely we denote by $X_{i,j}$ the $i$th observation of group $j$, for $j = 1, \ldots, d$, where $N_j$ is the number of observations in group $j$. Each observation $X_{i,j}$ is a $\mathbb{X}$-valued random element defined on the common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and the infinite sequences $\mathbf{X}_j = (X_{i,j})_{i \geq 1}$ for $j = 1, \ldots, d$ are assumed to be partially exchangeable, i.e., $(\mathbf{X}_1, \ldots, \mathbf{X}_d) \stackrel{\mathrm{d}}{=} (\pi_1 \mathbf{X}_1, \ldots, \pi_d \mathbf{X}_d)$ for each set $\pi_1, \ldots, \pi_d$ of finite permutations on the natural numbers, where $\pi_j \mathbf{X}_j = (X_{\pi_j(i),j})_{i \geq 1}$.

Thanks to the de Finetti representation theorem, the partial exchangeability of the sequences $\mathbf{X}_j$'s is equivalent to the existence of the so-called *de Finetti measure*, that is, a probability measure $Q_d$ over $\mathbb{P}_{\mathbb{X}}^d$ such that

$$\mathbb{P}\left[\bigcap_{j=1}^{d}\bigcap_{i=1}^{N_j}\{X_{i,j} \in A_{i,j}\}\right] = \int_{\mathbb{P}_{\mathbb{X}}^d}\prod_{j=1}^{d}\prod_{i=1}^{N_j} p_i(A_{i,j}) Q_d(\mathrm{d}p_1, .., \mathrm{d}p_d),$$

for any $(N_1, \ldots, N_d) \in \mathbb{N}^d$ and for any collection of Borel sets $A_{i,j} \in \mathcal{X}$, as $j = 1, \ldots, d$ and $i = 1, \ldots, N_j$. The de Finetti measure $Q_d$ works as a prior distribution.

The definition and investigation of a suitable de Finetti measure $Q_d$ in order to induce dependence across the $d$ groups of observations is a well-studied topic, and *hierarchical processes* are highly popular Bayesian nonparametric models in this context. More precisely, the general structure of a hierarchical process is

$$\begin{aligned}(\tilde{p}_1, \ldots, \tilde{p}_d) \mid \tilde{p}_0 &\overset{\text{iid}}{\sim} \tilde{\mathcal{L}}_0, \\ \tilde{p}_0 &\sim \mathcal{L}_0,\end{aligned} \tag{1}$$

where $\tilde{\mathcal{L}}_0$ is the probability distribution of each random probability measure $\tilde{p}_i$ such that $\mathbb{E}_{\tilde{\mathcal{L}}_0}[\tilde{p}_i \mid \tilde{p}_0] = \int p\tilde{\mathcal{L}}_0(\mathrm{d}p) = \tilde{p}_0$, whereas $\mathcal{L}_0$ is such that $\mathbb{E}_{\mathcal{L}_0}[\tilde{p}_0] = \int p\mathcal{L}_0(\mathrm{d}p) = P_0$ for some fixed non-atomic probability measure $P_0$. The vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$ in (1) defines a prior for the probability distributions of $d$ partially exchangeable samples, with dependence across samples being induced by the measure $\tilde{p}_0$. A notable example is the hierarchical Dirichlet process of (13). Other instances of hierarchical models were introduced in the context of survival analysis: the work of (1) presents a class of multivariate mixtures whose distribution acts as a prior for the vector of sample-specific baseline hazard rates, whereas (11) extends (4) and introduces a nonparametric model for the survival functions of $d \geq 2$ groups of survival times, modeling the dependence structure via Lévy copulas. We refer to (12) for a complete review of dependent structures in Bayesian nonparametrics.
This manuscript aims to introduce a hierarchical generalization of the popular neutral to the right processes, as introduced in (3), to the partially exchangeable framework. Note that this model is particularly suitable in applications with a high number of groups ($d \gg 1$) and a low number of observations per group, i.e., in contexts in which the borrowing of information can be particularly appreciable.

The paper is organized as follows. In Sect. 2, we recall the definition of neutral to the right processes, focusing on their characterization via completely random measures and their relevance in survival analysis. In Sect. 3, we define a hierarchical extension of neutral to the right processes based on a $d$-dimensional generalization of completely random measures. Finally, in Sect. 3.1, we provide an example of hierarchical neutral to the right processes called hierarchical Beta-Stacy.

DRAFT

## 2. Neutral to the right processes

Completely random measures (CRMs), introduced by (8), are tractable probabilistic tools that allow the construction of several nonparametric priors (9). Let us denote by $M_{\mathbb{X}}$ the space of boundedly finite measures on $(\mathbb{X}, \mathcal{X})$, i.e., a measure $\mu$ belongs to $M_{\mathbb{X}}$ if and only if $\mu(A) < \infty$ for any bounded set $A \in \mathcal{X}$. Moreover, let us denote by $\mathcal{M}_{\mathbb{X}}$ the corresponding Borel $\sigma$-algebra.

**Definition 1** (CRM)**.** *A measurable map $\tilde{\mu} : (\Omega, \mathcal{A}) \to (M_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ is a completely random measure (CRM) if the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_k)$ are mutually independent for any finite choice of mutually disjoint events $A_1, \ldots, A_k \in \mathcal{X}$, for any $k \geq 1$.*

Kingman (8) proved that each CRM $\tilde{\mu}$ admits the following decomposition

$$\tilde{\mu} = \mu_0 + \sum_{i \geq 1} V_i \delta_{x_i} + \sum_{k \geq 1} J_k \delta_{X_k},$$

where $\mu_0$ is a non-random measure, $(x_i)_{i \geq 1}$ are fixed elements of $\mathbb{X}$, $(V_i)_{i \geq 1}$ are positive and mutually independent random variables. In addition, $\tilde{N} = \sum_{k \geq 1} \delta_{(J_k, X_k)}$ is a Poisson process over $\mathbb{R}^+ \times \mathbb{X}$, therefore both $J_k$'s and $X_k$'s are random variables. In other words, each CRM is the sum of three components: a non-random measure $\mu_0$, a sum of random jumps at fixed locations, and a sum of random jumps at random locations (*atoms*). In this paper, we focus on CRMs $\tilde{\mu}$ composed by a sequence of positive random jumps $(J_k)_{k \geq 1}$ and a sequence of random atoms $(X_k)_{k \geq 1}$, so that $\tilde{\mu} = \sum_{k \geq 1} J_k \delta_{X_k}$. Moreover, since a CRM is a transformation of a Poisson process, its distribution can be easily characterized by means of its Laplace functional. Indeed, the Laplace functional equals

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}\right] = \exp\left(-\int_{\mathbb{R}^+ \times \mathbb{X}} \left(1 - e^{-sf(x)}\right) \nu(\mathrm{d}s, \mathrm{d}x)\right), \tag{2}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f| \mathrm{d}\tilde{\mu} < \infty$ almost surely. In (2) the measure $\nu$ is called *Lévy intensity* of the CRM $\tilde{\mu}$.
We assume the Lévy intensity $\nu$ can be written as $\nu(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s)\alpha(\mathrm{d}x)$, where $\alpha$ is a measure on $(\mathbb{X}, \mathcal{X})$ and $\rho$ is a *transition kernel* on $\mathbb{X} \times \mathcal{B}(\mathbb{R}^+)$, i.e., the function $x \mapsto \rho_x(A)$ is $\mathcal{X}$-measurable for any $A \in \mathcal{B}(\mathbb{R}^+)$ and $\rho_x$ is a measure on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ for any $x \in \mathbb{X}$. The Lévy intensity characterizes the CRM since it contains all the information about the distributions of its jumps and locations. We will write $\tilde{\mu} \sim \mathrm{CRM}(\nu)$ to denote the distribution of a CRM $\tilde{\mu}$ having Laplace functional as in (2). Note that if the jumps and the atoms of the CRM are independent, then its Lévy intensity can be written as $\nu(\mathrm{d}s, \mathrm{d}x) = \rho(\mathrm{d}s)\alpha(\mathrm{d}x)$, where $\rho$ does not depend on $x \in \mathbb{X}$.

Neutral to the right (NTR) processes, introduced by Doksum (3), are popular nonparametric priors in survival analysis. We now recall the definition of an NTR process, in which $\mathbb{X} = \mathbb{R}^+$ since the observations are survival times.

**Definition 2** (NTR)**.** *A random probability measure $\tilde{p}$ on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ is neutral to the right (NTR) if for any finite partition $0 \leq t_1 < t_2 < \ldots < t_k < \infty$ of $\mathbb{R}^+$, for any $k \geq 1$, the random variables*

$$\tilde{F}(t_1), \frac{\tilde{F}(t_2) - \tilde{F}(t_1)}{1 - \tilde{F}(t_1)}, \ldots, \frac{\tilde{F}(t_k) - \tilde{F}(t_{k-1})}{1 - \tilde{F}(t_{k-1})},$$

*are independent, where $\tilde{F}(t) = \tilde{p}((0, t])$ for any $t > 0$.*

Importantly, NTR processes are tied to completely random measures by the following result.

**Theorem 1** (Doksum)**.** *A random probability measure $\tilde{p}$ is NTR if and only if there exists a CRM $\tilde{\mu}$ on $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ such that*

$$\mathbb{P}\left[\lim_{t \to \infty} \tilde{\mu}((0, t]) = \infty\right] = 1$$

*and the corresponding random distribution function $\tilde{F}(\cdot) = \tilde{p}((0, \cdot])$ satisfies*

$$\{\tilde{F}(t) : t > 0\} \stackrel{\mathrm{d}}{=} \{1 - e^{-\tilde{\mu}((0,t])} : t > 0\}.$$

*We will write $\tilde{p} \sim \mathrm{NTR}(\tilde{\mu})$.*

Theorem 1 characterizes an NTR measure $\tilde{p} \sim \mathrm{NTR}(\tilde{\mu})$ in terms of the Lévy intensity $\nu$ of the associated CRM $\tilde{\mu}$. Moreover, NTR processes enjoy an appealing conjugacy property when used as nonparametric priors for exchangeable data, even in the presence of censored data. In particular, it can be proved (3; 5) that an NTR prior for possibly censored exchangeable data leads to an NTR posterior. For a complete discussion of the subject, see for example (6; 9).

## 2.1 NTR priors for survival analysis

Let us consider a set of dependent survival times $T_1, \ldots, T_n$ and a set of right censoring times $C_1, \ldots, C_n$. The observed values in survival analysis are the elements of the sequence $\{(X_i, \Delta_i) : i = 1, \ldots, n\}$, where $X_i = \min(T_i, C_i)$ and $\Delta_i = \mathbb{1}_{(0, C_i]}(T_i)$ for any $i = 1, \ldots, n$. In other words, the variable $X_i$ represents the $i$th observed time, while the variable $\Delta_i$ tells us whether the $i$th observation is censored or not. Let us define $Y(x) = \sum_{i=1}^n \mathbb{1}_{[x, \infty)}(X_i)$ the *at risk process*, which counts the number of subjects still alive at time $x$. Moreover, let us also consider the distinct observations $X_1^*, \ldots, X_k^*$, where $k \leq n$, and define $\Delta_i^* = \max_{j : X_j = X_i^*} \Delta_j$. Finally, consider the counting processes $n_i = \sum_{j=1}^n \mathbb{1}_{(X_j = X_i, \delta_j = 1)}$ and $n_i^c = \sum_{j=1}^n \mathbb{1}_{(X_j = X_i, \delta_j = 0)}$. The general result that characterizes the posterior distribution of an NTR prior can be found in (5), and it is recalled here for convenience.

**Theorem 2** (Ferguson and Phadia)**.** *Let $(X_n)_{n \geq 1}$ be an exchangeable sequence of random variables in $\mathbb{R}^+$ such that*

$$\begin{aligned} X_1, \ldots, X_n \mid \tilde{p} &\stackrel{iid}{\sim} \tilde{p}, \\ \tilde{p} &\sim \mathrm{NTR}(\tilde{\mu}), \end{aligned}$$

DRAFT

*where* $\tilde{\mu} \sim \text{CRM}(\nu)$ *and* $\nu(\mathrm{d}s, \mathrm{d}x) = \rho_x(\mathrm{d}s)\alpha(\mathrm{d}x)$*. Then, the law of* $\tilde{\mu} \mid (X_i, \Delta_i)_{i=1}^n$ *coincides with the law of the random measure*

$$\tilde{\mu}^* = \tilde{\mu}_c^* + \sum_{i:\Delta_i^*=1} J_i \delta_{X_i^*},$$

*where* $\tilde{\mu}_c^* \sim \text{CRM}(\nu^*)$ *and* $\nu^*(\mathrm{d}s, \mathrm{d}x) = e^{-Y(x)s}\rho_x(\mathrm{d}s)\alpha(\mathrm{d}x)$*. The* $J_i$*'s are independent random variables, independent from* $\tilde{\mu}_c^*$*, having density on* $\mathbb{R}^+$ *proportional to* $(1-e^{-s})^{n_i} e^{-s(\overline{n}_{i+1}+\tilde{n}_i^c)}\rho_{X_i^*}(\mathrm{d}s)$*, where* $\overline{n}_r = \sum_{i=r}^k n_i$*,* $\tilde{n}_r^c = \sum_{i=r}^k n_i^c$ *and* $\overline{n}_{k+1} = 0$*.*

Theorem 2 provides a general characterization of the posterior of an NTR prior observing possibly censored data, and it holds true in the exchangeable framework. Examples are discussed in (7; 9; 14).

## 3. Hierarchical NTR processes

In many applications concerning survival analysis the exchangeability assumption is often inadequate. In such a setting, the characterization of NTR processes given in Theorem 1 can be extended by defining a hierarchical version of completely random measures that we introduce below. Let us consider the notation and the framework introduced in Sect. 1, 2. Following (1; 10), let $\tilde{\mu}_0 \sim \text{CRM}(\nu_0)$ be a CRM having Lévy intensity $\nu_0 = \rho_x(s)\mathrm{d}s\alpha_0(\mathrm{d}x)$. Exploiting Sect. 2. let us write $\tilde{\mu}_0$ in terms of the points $(\tilde{h}_{0,k}, \tilde{x}_k)_{k\geq 1}$ of a marked Poisson point process on $\mathbb{R}^+ \times \mathbb{X}$ as follows:

$$\tilde{\mu}_0 = \sum_{k\geq 1} \tilde{h}_{0,k}\delta_{\tilde{x}_k}. \tag{3}$$

Let $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ be a vector of CRMs, for $d \geq 1$, which are independent conditional on $\tilde{\mu}_0$, i.e., each CRM of the vector can be conditionally represented as

$$\tilde{\mu}_j \mid \tilde{\mu}_0 \stackrel{\text{d}}{=} \sum_{k\geq 1} \tilde{h}_{j,k}\delta_{\tilde{x}_k}, \tag{4}$$

for any $j = 1, \ldots, d$, where $\tilde{x}_k$'s are the same atoms of $\tilde{\mu}_0$ and $\tilde{h}_{j,k}$ represents the $k$th non-negative jump of the $j$th random measure $\tilde{\mu}_j$. Moreover, conditionally on $\tilde{\mu}_0$, the random variables $(\tilde{h}_{j,k})_{j\geq 1, k\geq 1}$ are assumed to be independent.

**Definition 3** (hCRM)**.** *The vector characterized by* (3)–(4) *is termed a vector of hierarchical Completely Random Measures (hCRMs).*

In this hierarchical approach, the baseline measure $\tilde{\mu}_0$ induces dependence between the groups of partially exchangeable observations. Moreover, each group is modeled by a different element of the hCRM vector $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$. Exploiting this hierarchical extension of completely random measures, we are now able to extend the definition of neutral to the right processes as follows.

**Definition 4** (hNTR)**.** *Let us consider a* $d$*-dimensional vector of hCRMs* $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ *as defined in Definition 3, with* $\mathbb{X} = \mathbb{R}^+$*. The vector of random distributions* $(\tilde{F}_1, \ldots, \tilde{F}_d)$ *such that*

$$\tilde{F}_j(t) = \tilde{p}_j((0,t]) = 1 - e^{-\tilde{\mu}_j((0,t])},$$

*for any* $j = 1, \ldots, d$ *and* $t > 0$*, is called a vector of hierarchical Neutral To the Right (hNTR) processes.*

The class of hierarchical processes in Definition 4 is a natural extension of NTR processes to the partially exchangeable framework. Therefore, they are a natural family of nonparametric priors for partially exchangeable survival times. Note that Definition 4 implies that given a vector of hCRMs $(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$, the corresponding $d$-dimensional vector of hNTR processes is $(\tilde{F}_1, \ldots, \tilde{F}_d)$ such that $\tilde{p}_j \sim \text{NTR}(\tilde{\mu}_j)$, where $\tilde{F}_j(\cdot) = \tilde{p}_j((0,\cdot])$ for each $j \in \{1, \ldots, d\}$. Therefore, following Definition 3, a vector of hNTR processes is completely specified choosing a base measure $\tilde{\mu}_0$ and the law of the conditional jumps $\tilde{h}_{j,k} \sim f_j(h \mid \tilde{h}_{0,k}, \tilde{x}_k, b_j)$ for each $j = 1, \ldots, d$, where $b_j$ is an additional parameter that can be defined in order to specify the $j$th measure of the hCRM vector.

DRAFT

## 3.1 The hierarchical Beta-Stacy process

The Beta-Stacy process is an NTR process introduced by Walker and Muliere (14) as nonparametric prior for survival times in the exchangeable setting.

**Definition 5** (Beta-Stacy)**.** *Let $\alpha$ be a probability measure on $\mathbb{R}^+$ which is absolutely continuous w.r.t. the Lebesgue measure, and let $c : \mathbb{R}^+ \to \mathbb{R}^+$ be a piecewise continuous function. The random probability distribution $\tilde{F}$ is a Beta-Stacy process with parameters $c$ and $\alpha$ if $\tilde{F}(t) = \tilde{p}((0,t])$ for each time $t > 0$ and $\tilde{p} \sim \text{NTR}(\tilde{\mu})$, where*

$$\tilde{\mu} \sim \text{CRM}(\nu), \qquad \nu(\mathrm{d}s, \mathrm{d}x) = \frac{e^{-sc(x)\alpha((x,\infty))}}{1-e^{-s}} c(x)\mathrm{d}s\alpha(\mathrm{d}x).$$

*In such a case we say that the CRM $\tilde{\mu}$ is a log-Beta measure with parameters $c$ and $\alpha$ and we write $\tilde{F} \sim \text{Beta-Stacy}(c, \alpha)$ and $\tilde{\mu} \sim \text{log-Beta}(c, \alpha)$.*

It is known from (14) that in a model for exchangeable survival times, the Beta-Stacy process is a conjugate nonparametric prior, i.e., assuming a Beta-Stacy prior the posterior is a Beta-Stacy process with updated parameters. Let us now consider the hierarchical model defined in Sect. 3. A possible extension of the Beta-Stacy process to the partially exchangeable framework can be obtained by choosing a log-Beta CRM as the base measure $\tilde{\mu}_0$ in (3) and specifying dependent jumps distributed as transformations of Beta variables in (4). In particular, let us define

$$\begin{aligned} \tilde{\mu}_0 &\sim \text{log-Beta}(c, \alpha), \\ 1 - e^{-\tilde{h}_{j,k}} \mid \tilde{\mu}_0 &\sim \text{Beta}(c_j(\tilde{x}_k)F_j(\tilde{h}_{0,k}), c_j(\tilde{x}_k)(1 - F_j(\tilde{h}_{0,k}))), \end{aligned}$$

where $\tilde{\mu}_0 = \sum_{k\geq 1} \tilde{h}_{0,k}\delta_{\tilde{x}_k}$ is the usual representation of the CRM $\tilde{\mu}_0$ as functional of a Poisson process, $c_j : \mathbb{R}^+ \to \mathbb{R}^+$ is a function referring to group $j \in \{1, \ldots, d\}$, while $F_j$ is c.d.f. which should encapsulate our prior opinion on the hazard function for the $j$th group. This is a noteworthy example of an hNTR process, which we will refer to as a *hierarchical Beta-Stacy process*. The specification of a particular prior of this family of hierarchical processes should allow a precise characterization of its posterior distribution as well as simulation studies, even without conjugacy property that might be lost in the partially exchangeable framework. Our aim in (2) is to study the whole class of hNTR processes, characterizing the posterior distribution with a particular focus on the hierarchical Beta-Stacy process; we are also interested to develop marginal and conditional algorithms to address posterior inference. We think that the model proposed here, that we plan to investigate in (2), can be used to solve a large variety of survival problems.

DRAFT

## References


[1] Camerlenghi, F., Lijoi, A., Pruenster, I.: Survival analysis via hierarchically dependent mixture hazards. Ann. Stat. **49**(2), 863–884 (2021).

[2] Cogo, R., Camerlenghi, F., Rigon, T.: Hierarchical neutral to the right priors. In preparation.

[3] Doksum, K.: Tailfree and neutral random probabilities and their posterior distributions. Ann. Probab., **2**, 183–201 (1974).

[4] Epifani, I., Lijoi, A.: Nonparametric priors for vectors of survival functions. Stat. Sin., **20**(4), 1455–1484 (2010).

[5] Ferguson, T.S., Phadia, E.G.: Bayesian nonparametric estimation based on censored data. Ann. Stat. **7**, 163–186 (1979).

[6] Ghosal, S., Van der Vaart, A.: Fundamentals of nonparametric Bayesian inference. Cambridge University Press (2017).

[7] Hjort, N.L.: Nonparametric Bayes estimators based on beta processes in models for life history data. Ann. Stat. **18**(3), 1259–1294 (1990).

[8] Kingman, J.: Completely random measures. Pacific J. Math., **21**, 59–78 (1967).

[9] Lijoi, A., Prunster, I.: Models beyond the Dirichlet process. In: Bayesian nonparametrics, Cambridge University Press (2010).
[10] Masoero, L., Camerlenghi, F., Favaro, S., Broderick, T.: Posterior representations of hierarchical completely random measures in trait allocation models. BNP@NeurIPS (2018).
[11] Palacio A.R., Leisen, F.: Bayesian nonparametric estimation of survival functions with multiple-samples information. Electron. J. Stat. **12**, 1330–1357 (2018).
[12] Quintana, F.A., Müller, P., Jara, A., MacEachern, S.N.: The Dependent Dirichlet Process and Related Models. Stat. Sci. **37**(1), 24–41 (2022).
[13] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet processes. J. Amer. Statist. Assoc. **101**, 1566–1581 (2006).
[14] Walker, S., Muliere, P.: Beta-Stacy processes and a generalization of the Pólya-urn scheme. Ann. Stat. **25**(4), 1762–1780 (1997).

DRAFT

# Empowering futures adopting a spatial convergence of opinions: a Real-Time Spatial Delphi approach

Yuri Calleo[a], Simone Di Zio[b], and Francesco Pilla[a]

[a] School of Architecture, Planning and Environmental Policy, University College Dublin, Ireland;
yuri.calleo@ucdconnect.ie,francesco.pilla@ucd.ie
[b] Department of Legal and Social Sciences, University "G. d'Annunzio", Chieti-Pescara;
s.dizio@unich.it

DRAFT

**Abstract**

The Delphi method is a structured communication approach used to gather and process expert judgments in order to make predictions or estimates about a specific issue. In the Futures Studies (FS) context, the Delphi method is combined with different methods, however, one of the main combinations is with the scenario method, forming the Delphi-based scenarios (DBS). Scenarios are hypothetical descriptions of the futures, adopted in different disciplines in order to prevent possible future outcomes. Nevertheless, one of the main weaknesses of DBS is the lack of spatial perspectives, where most of the time quantitative models are often used to develop spatial scenarios. To overcome this challenge, in this paper, we illustrate an enhancement of the Real-Time Spatial Delphi, through a novel web-based open platform useful to obtain, with innovative tools, a spatial convergence of opinions among a cohort of experts. We apply the method, developing spatial scenarios in the climate change context, for the city of Dublin in 2050.

**Keywords:** Real-Time Spatial Delphi, Futures Studies, spatial convergence

## 1. Introduction and Theoretical Framework

From the first application by the RAND Corporation (Linstone and Turoff, 1975) in the 1950s, the Delphi method is widely known as a structured technique for gathering and aggregating the opinions of a group of experts on a specific issue or problem. It is often used when there is a lack of data (e.g., historical data) or when the experts have diverse opinions on a complex subject. The Delphi method is generally conducted in multiple iterative rounds, where experts with high competencies, provide anonymous judgments to a series of questions (Baker et al., 2006). The judgments are carefully analysed generally with a statistical summary (e.g., mean, median, interquartile range (IQR), etc.) and the results are used to create a set of questions for the next rounds, where experts can revise their opinions based on that information. In this case, the process is repeated until there is a consensus among the experts (but is not limited to), or when we have a stability condition (Dajani et al., 1979; von der Gracht, 2012). The Delphi method is widely adopted in different disciplines and is characterized by four main features, such as anonymity, iteration, controlled feedback, and statistical group responses (Rowe and Wright, 1999).

In the Futures Studies (FS) context, scenarios are representations of possible future states and outcomes, useful to help explore and understand the potential impacts of different events, trends, and decisions that may play out over time (von der Gracht and Darkow, 2010). Scenarios can take various forms, including narrative-style descriptions, mathematical models, and visual representations (Varho and Tapio, 2013). The aim of developing scenarios is to provide a holistic and detailed picture of potential futures, considering the interactions between various factors and their effects on society, technology, economics, and the environment (Bishop et al., 2007; Hines et al., 2017). Among the many combinations

of the scenario method, one of the widest used is the Delphi method, forming the so-called Delphi-based scenarios (DBS). From what emerged, DBS constitute an interesting method to explore and empower different futures, however, one of the main weaknesses is the lack of spatial references. The conventional Delphi technique consists of multiple cyclical stages, but the surveys commonly use open-ended, closed-ended, and scaling questions which are not suitable for forecasting future territorial distributions in the form of points or polygons. To overcome this challenge, in the literature, two main methods are developed, namely the Spatial Delphi (Di Zio and Pacinelli, 2011) and the Real-Time Spatial Delphi (Di Zio et al., 2017), however, from the first applications, these two methods are still unexplored.

To overcome this challenge, we adopted Real-Time Geo-Spatial Consensus System (http://www.rtgscs.com/), a novel web-based open platform, to develop spatial scenarios for the climate change context in the city of Dublin, Ireland.

## 2. Materials and Methods

In this paper, we follow a modified version of the DBS, suitable to our context, starting from the Strategic Foresight approach proposed by Bishop et al., (2007): Framing, Scanning, Forecasting, Visioning, Planning and Acting integrating it with the novel platform.

The first step sees careful desk research in order to develop guidelines, acquire spatial data information and understand the territorial framework. This study is conducted in the climate change context, in the city of Dublin, where a Coastal City Living Lab, part of the SCORE H2020 EU Project, is located. Dublin faces a significant threat from coastal flooding in the coming years due to insufficient defences and a lack of public support for upgrades, such as higher flood walls. For our study, we first determine the territorial boundaries, to then select a time horizon (in this case, 2050) that is long enough to be relevant for decision-making but not too far to make accurate predictions. To define a list of key drivers, we did not involve experts in a workshop, however, we extract the main drivers from the project proposal, where experts have already defined the possible future risks. Starting from that, we have 6 hazards that may affect the climate future of Dublin in 2050, specifically: coastal flooding, land flooding, landslides, heatwave, storm surge, and coastal erosion.

From these drivers, we can now obtain the questions for our panel: $RQ1$: Thinking about 2050, what area will be most at risk of flooding? $RQ2$: Thinking about 2050, what area will be most at risk of erosion? $RQ3$: Thinking about 2050, what area will be most affected by extreme events? Once the questions have been validated by the research team (in terms of clarity), we upload them into the platform. In this study, for the selection of a suitable sample of experts we choose to include two types of experts: 1) Internal experts: part of the SCORE H2020 project, and 2) External experts: with a strong background and expertise in the topics emerged, but not part of the project. We select academics with a high level of experience, stakeholders, and representatives from companies, local authorities, and other agencies, ensuring the diversity of expertise in the panel. We reached out to 12 internal and 50 external experts by sending an introductory email about the research and requesting their participation. Out of these, 26 experts agreed to participate in the study, including 6 internal and 20 external experts. Once the experts accepted to participate in the study, we sent each panellist a registration form, including relevant attachments such as spatial data information, territorial framework, guidelines, etc.

To pursue the aim of this paper, we adopt Real-Time Geo-Spatial Consensus System (RT-GSCS, www.rtgscs.com), a novel web-based open platform adopting Real-Time Spatial Delphi (Di Zio et al., 2017) to obtain a spatial convergence of opinions among panellists in a spatial context (Fig. 1). Once all of them have registered to the platform we can start the survey. We ask the panellists to select the question and answer by placing a point on the map and electing the plausibility of the answer from 1-5. From this point, an automatic circle appears on the map and based on others' judgments it moves, shrinks, or enlarges in real-time. The experts can at any time, motivate their answers with specific comments.

Following the logic expressed by Di Zio and Pacinelli, in 2011, spatial convergence is pursued by adopting a geometric element represented by a circle $C$, the smallest one among all the possible circles, including the 50% of $N$ judgments (analogue of the IQR of the traditional Delphi). If, we assume that $N = n_1, n_2, n_3, \ldots, n_j$ is the number of the experts judgments on a question (points on the map), we want to find a minimum area $A_i$ of a circle $C_i$ covering half of those points: $A_i \supseteq T_{(N/2)}$ , where $T_{(N/2)}$ denotes a set containing 50% of the $N$ points. We take into account 50% of the points because some of the points could be placed in some locations not suitable for our research objectives. Since there are infinite circles like this, to limit the search we impose the constraint that $C_i$ must have its centre in one of the $N$ points.

DRAFT

Therefore, for each question we find a vector $A = A_1, A_2, A_3, \ldots, A_N$ where $A_i$ is the area of a circle containing 50% of the $N$ points and centred in point $n_i$. Then, $min(A)$ corresponds to the geo-consensus.

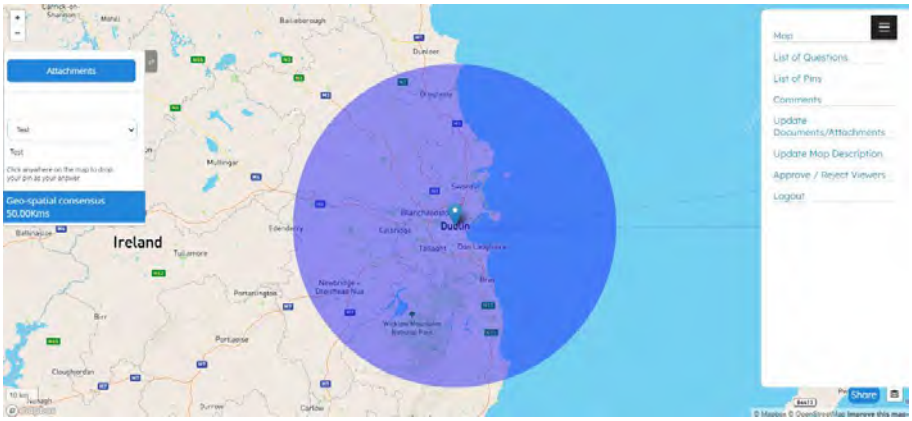



Figure 1: RT-GSCS interface

In our study, we have two types of results: geographical, with visual map outputs and non-geographical results, with spatial data and comments. In this paper, to evaluate spatial data we consider the three measures proposed by Di Zio et al., 2017, denoted by $M_1$, $M_2$ and $M_3$. The first measure, $M_1$, corresponds to the final circle area useful for the identification of outcomes. Nevertheless, this measure is absolute and does not consider the study area boundaries and the size of the initial circle. To address this challenge, we also consider $M_2 = 1 - \frac{FC}{S}$ as second indicator, obtained as the ratio between the final circle's area ($FC$) and the surface ($S$) of Dublin ($S = 117.8\ km^2$). In this case, the indicator shows the degree of geo-consensus, and the more the measure is closer to 1, the more the consensus circle is small compared to the surface. The third indicator, measures the spatial convergence dynamic process, $M_3 = \frac{FC}{IC} \cdot 100$, where $IC$ is the initial circle area (set a priori as 50 $km^2$), and the more the value is higher (close to 100%), the more we have a poor convergence of opinions, the more is close to zero, the higher the convergence. Before finishing the study, we must take into consideration that convergence is not the only stopping criterion to be used to end the exercise (von der Gracht, 2012). For this reason, in our case, we also look for stability with time series analysis.

In the end, we obtain three main scenarios (one for each spatial-question) derived from value judgments and ready for decisions. Nevertheless, spatial data could be analysed in multiple ways and since we have a matrix composed of spatial data and the relative plausibility on a Likert type scale from 1 to 5 expressed by experts, we use ArcGIS PRO to represent data with a Heat Map. A Heat Map is a graphical representation of data in two dimensions that uses colour coding to represent different values, extremely useful in our case to highlight the plausibility of the judgments and possible hotspots. The final result is a map that represents the plausibility in different points of the space based on the available information and can then be used to make better decisions considering the overall territory.

## 3. Results

The results fully answered the research objectives, we obtain 3 main spatial scenarios. The study began officially on November 1, 2022, and concluded on December 5, 2022, after a double stability check. In our panel, out of 62 experts participating in the survey, 26 accepted by placing at least one point on the map. For $Q1$, 58 experts judgments were obtained with 13 comments. For $Q2$, 54 judgments were considered with 16 comments. At the end, for $Q3$, 40 points were recorded with 11 comments. The outputs are presented in Fig. 2.

According to the findings depicted in Fig. 2, the area most likely to face the threat of flooding by the year 2050 ($Q1$), is located in the central part of Dublin city, between the two banks of the River Liffey. The experts believe that a potential flood in this area could have devastating effects, potentially resulting in harm to buildings and infrastructure, disruption of essential services, environmental harm, and loss of life. With regards to $Q2$, the area most susceptible to erosion by 2050, is the eastern coastal

DRAFT

regions of Dublin. The experts believe that coastal erosion in this area could have severe impacts, including the loss of valuable real estate and infrastructure, such as roads, buildings, and homes. It could also pose a threat to public safety by destabilizing cliffs and making them prone to collapse. Furthermore, coastal erosion could lead to the loss of recreational areas and habitats for wildlife, as well as negatively impact the local economy by affecting businesses that depend on the coast for tourism or fishing. In response to $Q3$ the panellists have identified the central area of Dublin as the most likely to be impacted. The experts believe that this area could be affected by various extreme events, such as storms, floods, and heat waves, which could result in significant impacts on the city. Such events could damage or destroy buildings, infrastructure, and public facilities, disrupting daily life for residents and visitors. They could also pose a threat to public safety, causing power outages, landslides, and other hazards. Additionally, extreme events can have social implications, such as increased demand for emergency services and strain on the healthcare system. Finally, these events could also have environmental impacts, altering the local ecosystem and contributing to the degradation of natural habitats.

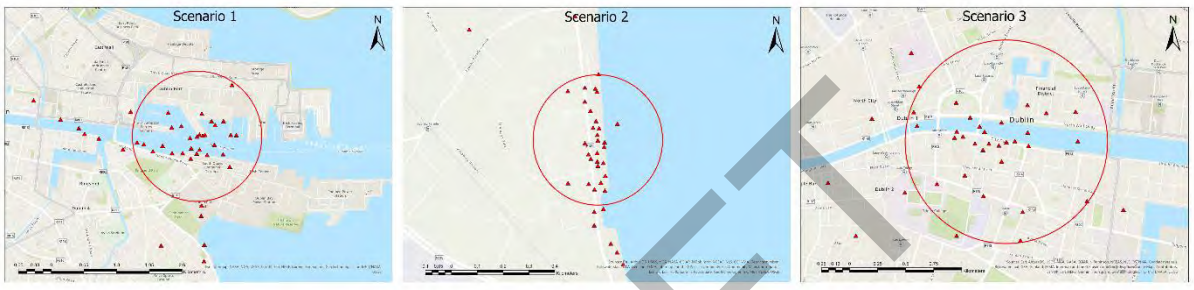



Figure 2: Spatial scenarios

The results of the indicators measuring spatial consensus are shown in Tab. 1, which highlights the convergence of the geographical data obtained from our study. Each question consists of various variables, including the area of Dublin city ($S$), the total number of judgments ($N$) the initial circle ($IC$), the final circle ($M_1$), and the convergence measures $M_2$ and $M_3$.

Table 1: Geographical results

| Scenario | $S\ (km^2)$ | $IC\ (km^2)$ | $FC\ (km^2)\ M_1$ | $M_2$ | $M_3$ | $N$ |
|---|---|---|---|---|---|---|
| Sc.1 | 117.8 | 8.24 | 0.77 | 0.993 | 9.34% | 58 |
| Sc.2 | 117.8 | 3.25 | 0.15 | 0.999 | 4.61% | 54 |
| Sc.3 | 117.8 | 2.92 | 0.54 | 0.995 | 18.49% | 50 |

The experts were able to achieve a significant reduction of the initial circles, greater than 99% in all final outputs (as indicated by $M_2$ in Tab. 1), reaching a high degree of convergence. For Sc.1 and Sc.2, the initial circle was reduced greatly (with $M_2$values of 0.993 and 0.999, respectively). The initial circle for Sc.1 was 8.24 $km^2$ and reduced to 0.77 $km^2$, while the initial circle for Sc.2 was 3.25 $km^2$ and reduced to 0.15 $km^2$. For Sc.3, the initial circle was smaller (2.92 $km^2$) and the final circle was 0.54 $km^2$ with a reduction of 0.995 (as indicated by $M_2$). $M_2$ is a measure of the efficacy of the solution with regards to the specific area of investigation, while $M_3$ measures the consensus dynamic among participants. In a traditional Delphi study, a common measure of consensus is the IQR, with a good consensus considered to be achieved when the IQR is less than 20% of the measurement scale used.

Similarly, in the Spatial Delphi, consensus can be reached when $M_3$ is less than or equal to 20%. In our case, the $M_3$ values of 9.34%, 4.61%, and 18.49% for the three research questions indicate that the experts reached a high level of consensus for Sc.1 and Sc.2, and a slightly lower level of consensus for Sc.3. This could mean that for extreme events it is more difficult to outline future scenarios in respect to erosion and flooding.

DRAFT

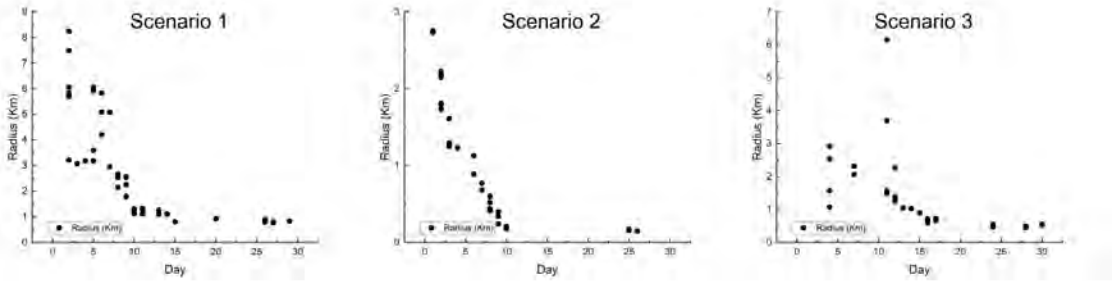



Figure 3: Time series of the sizes of the circles

From the collected data, we have time series of the size of the circles that track the stability of the results over time. Fig. 3 illustrates three-time series, one for each scenario, with the number of days from the start of the process on the horizontal axis and the radius of the circle on the vertical axis. The closer the sequence of points smoothly gets to the horizontal axis, the greater the geo-consensus among the experts. In Sc.1, the radius changed several times before reaching stability from the 18th to the 20th day. However, after we asked for validation, three additional changes took place. This means that the issue of flooding has been quite controversial, in fact, in the Dublin area, many efforts are taking action to mitigate the impacts and experts are dubious about the suitable solutions. Sc.2 had the strongest consensus, which can be seen from the sudden drop and stability from the 10th day, with only two changes made during validation. This means that experts agree on future spatial erosion dynamics finding a common consensus on the Dublin coastal area. In Sc.3, the circle underwent significant changes in the first 15 days – confirming that there has been a lot of debate about places that could be scenarios of extreme events in the future – but reached stability after the 18th day. In all scenarios, we observed stability after the validation phase with no further changes made. The results of the study that evaluated the plasuibility areas of three scenarios were analyzed to determine the scores ($X \sim U(1,5)$) assigned by participants.
The scores can be used to determine the mean ($\mu$), median, and standard deviation ($\sigma$) of each scenario. For Sc.1, the mean score was represented as $\mu_1 = 3.86$, with a median score of $X_1 = 4$ and a standard deviation of $\sigma_1 = 0.68$. For Sc.2, the mean score was represented as $\mu_2 = 4.24$, with a median score $X_2 = 4$ and a standard deviation of $\sigma_2 = 0.57$. For Sc.3, the mean score was represented as $\mu_3 = 4.155$, with a median score $X_3 = 4$ and a standard deviation of $\sigma_3 = 0.74$. These results suggest that scenario $S2$ had the highest mean ($\mu_2$) and median ($X_2$), indicating that participants considered it the most plausible of the three scenarios. At this stage, we present the results of the Heat Map, adopting the method previously illustrated. From what emerged, a noticeable association is observed between the high plausibility of occurrence and the final circle obtained. For Sc.1, the primary cluster of points identified is concentrated within the geo-consensus radius, except for a small cluster of 2-3 points outside of the radius. In Sc.2, a strong plausibility of occurrence is located within the geo-consensus radius, helping to provide a more comprehensive view of the area and highlights the selected areas. Finally, the last scenario depicts greater uncertainty among the experts. Although the strong plausibility of occurrence is located inside the final circle, there is a significant cluster of 11 points in the coastal city of Dún Laoghaire. This is why the $M_3$ measure is close to 20%, as the experts have identified two different plausible areas for extreme events.

Figure 4: Heat Map of the plausibility

DRAFT

## 4. Concluding remarks and future works

This paper adopted "Real-Time Geo-Spatial Consensus System", a novel web-based open platform to perform a Real-Time Spatial Delphi survey in the climate change context. The scenarios developed are a spatial representation of the experts' judgments, useful for decision-makers in empowering immediate decisions. The experts were able to achieve a significant reduction of the initial circles, greater than 99%, in all final outputs, reaching a high degree of convergence. These results suggest that the Real-Time Geo-Spatial Consensus System is a useful tool for decision-makers in the empowering of immediate decisions. The results of the Heat Map showed a noticeable correlation between the high plausibility of occurrence and the final circle obtained. Nevertheless, since the futures are multiple, these scenarios must be combined with others approaches (e.g., mathematical models, GIS analysis) in order to have an even more broad framework.

In future works, different implementations could be done as different spatial analyses (e.g., clustering) or analysis using text mining for expert comments. With regards to the results obtained, could be combined with mathematical models and submitted to a cohort of experts, part of the local authority or governmental bodies in order to develop strategies for immediate action.

**Acknowledgements** The work carried out in this paper was supported by the project SCORE which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003534.

## References

[1] Baker, J., Lovell, K., Harris, N.: How expert are the experts? An exploration of the concept of 'expert' within Delphi panel techniques. Nurse researcher, **14**(1) (2006).
[2] Bishop, P., Hines, A., & Collins, T.: The current state of scenario development: an overview of techniques. foresight, **9**(1), 5-25 (2007).
[3] Dajani, J.S., Sincoff, M.Z., Talley, W.K.: Stability and agreement criteria for the termination of Delphi studies, Technological Forecasting and Social Change, **13**, 83-90 (1979).
[4] Di Zio, S., Pacinelli, A. Opinion convergence in location: A spatial version of the Delphi method. Technological Forecasting and Social Change, **78**(9), 1565-1578 (2011).
[5] Di Zio, S., Rosas, J. D. C., Lamelza, L.: Real Time Spatial Delphi: Fast convergence of experts' opinions on the territory. Technological Forecasting and Social Change, **115**, 143-154 (2017).
[6] Hines, A., Gary, J., Daheim, C., van Der Laan, L.: Building foresight capacity: toward a foresight competency model. World Futures Review, **9**(3), 123-141 (2017).
[7] Linstone, H. A., Turoff, M. (Eds.): The delphi method. Reading, MA: Addison-Wesley, 3-12 (1975).
[8] Rowe, G., Wright, G.: The Delphi technique as a forecasting tool: issues and analysis. International journal of forecasting, **15**(4), 353-375 (1999).
[9] Varho, V., Tapio, P.: Combining the qualitative and quantitative with the Q2 scenario technique—The case of transport and climate. Technological Forecasting and Social Change, **80**(4), 611-630 (2013).
[10] von Der Gracht: Consensus measurement in Delphi studies: review and implications for future quality assurance. Technological forecasting and social change, **79**(8), 1525-1536 (2012).
[11] von Der Gracht, H.A., Darkow, I.-L.: Scenarios for the logistics services industry. A Delphi-based analysis for 2025. Int. J. Prod. Econ. **127** (1), 46–59 (2010).

DRAFT

# Stocks price forecasts using Stochastic Differential Equations: an empirical assessment

Dario Frisardi[a] and Matteo Spuri[a,b]

[a]Department of Statistical Sciences, Sapienza University of Rome, Italy;
dario.frisardi@uniroma1.it, matteo.spuri@uniroma1.it
[b]Bank of Italy, Italy.

**Abstract**

In this analysis we investigated the reliability of Stochastic Differential Equations (SDE) in forecasting stock prices on a short time period in a simplified framework. Four different SDE models were used, each applied on stocks of companies listed on NASDAQ and validated in each month between July 2022 and January 2023. Both the coverage of the confidence intervals estimates and the reliability of the average behaviour estimates were checked. Results show an overall good reliability for stocks with a strong expected percentage increase (above $3\%$).

***Keywords:*** Stochastic Differential Equations, Stochastic Processes, Forecasting, Finance

DRAFT

## 1. Introduction

Stochastic Differential Equations (SDE) models have a variety of applications and can be used in the study of many phenomena concerning different fields of science (e.g. epidemiology, geology, physics, mechanics) [3, 4, 5]. In particular, SDEs, as well as stochastic processes, are widely used for estimating and forecasting stock price trends.

Given the intrinsic uncertainty of such phenomena, progressively more complex models were developed in order to improve the estimation while accounting for several layers of variability (i.e. considering the interaction between different stocks). As a downside, their use has become increasingly complex as well.

Different papers showed how stochastic processes might yield accurate results also in much more simpler environments [8, 11], though this approach is not so well explored as of recent literature. Moreover, existing works focuses much more on intervals coverage without taking into account the trend forecasts. The goal of this study is to assess the predictive reliability in the short run of simpler models, in which each stock is evaluated individually. This was achieved by adapting different stochastic differential equations to the stock price trends. In particular, we focus on stocks with higher expected percentage increase, that was set as an expected increase of at least $3\%$.

The paper is structured as follows: Section 2 introduces the theoretical framework of this work, focusing on a few selected processes for stock estimation and forecasting; Section 3 reports the application and results of this procedure and Section 4 concludes.

## 2. Stochastic Differential Equation models for Finance

We considered four interrelated SDE models with a progressively increasing degree of complexity, thus assessing whether this leads to an improvement in the predictions. Moreover, SDE models were also chosen based on existing literature results. The four SDE models are the following:

### Geometric Brownian Motion

The differential equation of Geometric Brownian Motion (GBM) is defined as follows:

$$\mathrm{d}X_t = \mu X_t \mathrm{d}t + \sigma X_t \mathrm{d}W_t \tag{1}$$

This stochastic process can be used to study the percentage increase in invested capital $\frac{\Delta S_t}{S_t}$, which is equal to the sum of a deterministic trend and stochastic fluctuations. This process is broadly used in studying and forecasting stocks price trends [9].

### Ornestein-Uhlenbeck process

The Ornestein-Uhlenbeck process (OR-UH) is defined as:

$$\begin{cases} \mathrm{d}X_t = \theta(\mu - X_t)\mathrm{d}t + \sigma \mathrm{d}W_t \\ X_0 = x_0 \end{cases} \tag{2}$$

where: $\mu$ is the long-term mean; $\theta$ is the reversion speed, that is to say the tendency to fluctuate around $\mu$ (*mean reversion*); $\sigma$ is the volatility term.

### Cox-Ingersoll-Ross model

The Cox-Ingersoll-Ross (CIR) [1] model is defined as it follows:

$$\mathrm{d}X_t = r(\theta - X_t)\mathrm{d}t + \sigma\sqrt{X_t}\mathrm{d}W_t \tag{3}$$

This process is by construction non-negative ($X_t \geq 0 \ \forall t$) and it can be derived as the square of a OR-UH process. The $r$ parameter corresponds to the speed of convergence of the long-term mean, while $\sigma$ corresponds to the volatility.

### Chan-Karolyi-Longstaff-Sanders process

The Chan-Karolyi-Longstaff-Sanders process (CKLS) [2] is defined as it follows:

$$\mathrm{d}X_t = (\theta_1 + \theta_2 X_t)\mathrm{d}t + \theta_3 X_t^{\theta_4}\mathrm{d}W_t \tag{4}$$

This process can be seen as a generalization of other stochastic processes: it corresponds to a family of stochastic processes that, for given values of the parameter, can lead to the reproduction of different behaviors.

For example, if $\theta_4 = 0$ it corresponds to a OR-UH process, while if $\theta_4 = \frac{1}{2}$ it corresponds to a CIR model and if $\theta_4 = 1$ with $\theta_1 = 0$ it returns to a GBM. [7]

The first two parameters of the CKLS process ($\alpha = (\theta_1, \theta_2)$) represent the long-term mean and the speed of convergence, while the second pair ($\beta = (\theta_3, \theta_4)$) indicate volatility and the dependence of the process on $X_t$, respectively.

# 3. Research Methods and Results

We evaluated the prediction abilities of the different SDE models described in Section 2 on the daily stocks pricing of companies listed on NASDAQ. In particular, we selected the first 750 companies by market cap. Data were retrieved through `quantmod` R package [12]. In order to have multiple periods of validation of this procedure, we used observation up to the first Friday of each of the months between July 2022 and January 2023. For each of the selected stocks, the four SDE models were adapted, deriving the parameter estimates through the maximization of the quasi-likelihood using the `yuima` R package[6].

*Quasi-maximum likelihood* (QML) method consists in the maximization of an approximation of the density function given that the true one is unknown (or infeasible)[13]. This allows to take into account

DRAFT

different properties of the parameter distribution, that might not be considered using methods such as the least squares estimation. QML estimator (QMLE), under mild assumptions, is proved to be consistent and asymptotically normal, thus granting the required properties of an estimator. In formula, it is possible to define the QMLE as:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{\ell}_n \left( \mathbf{X}_n; \theta \right) \tag{5}$$

where $\hat{\ell}_n$ is the *quasi-log-likelihood function*, a function that approximate the true log-likelihood function. The quasi-log-likelihood function is, in general, a nonlinear function of $\theta$, so its maximization has to be computed numerically using nonlinear optimization algorithms.

Forecasts were simulated on the 20 days following the last date of observation. We estimated both the empirical $90\%$ confidence intervals and the average daily values through a bootstrap approach, simulating $10^4$ forecasts.

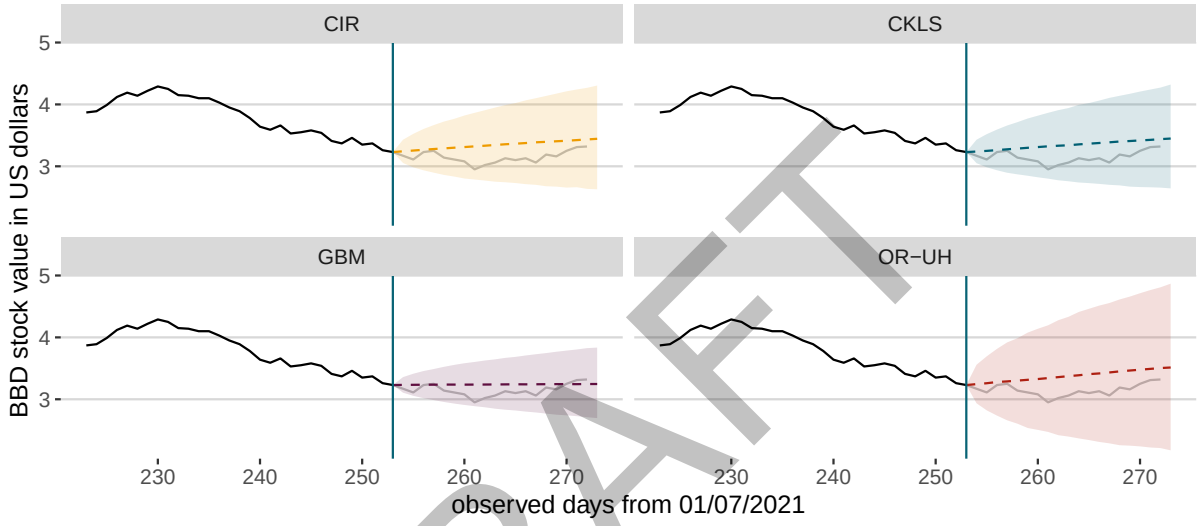



Figure 1: Forecast of average value and confidence intervals for BBD stock in July 2022.

Figure 1 shows forecast of the average daily values and confidence intervals of all four SDE models for BBD (Banco Bradesco SA) stock in July 2022. Results on BBD are reported as an explanatory example given that it is the stock with the highest forecasted increase in July 2022 concerning the CKLS model. Observation after the blue vertical line where not used in the estimation process and were used only in the following step to compute the confidence interval coverage and the average trend reliability.

For each stock we computed the expected increase as the percentage difference between the $20^{th}$ day forecast average value and the last observed value. We then selected stocks with an expected percentage increase of at least $3\%$, in order to evaluate the forecasts for stocks with a marked expected increase. Figure 2 shows the expected percentage increase of each stock in July 2022 with the CKLS process. The one selected by such method are highlighted in the green area.

On these selected stocks, we then evaluated the coverage of the empirical confidence interval, both on the first 10 days and on all the 20 predicted days. Coverage percentages are computed as the percentage of observation lying in the confidence interval region. Each model was evaluated independently.

Results, reported in Table 1, show a high coverage percentage for all four SDE models both on the 10 days forecast and the 20 days forecast (results are based on the aggregate of each month of analysis). The lowest coverage percentages, nonetheless above $87\%$, are those with CIR's forecasts, while GBM and OR-UH models have achieved the highest ones.

DRAFT

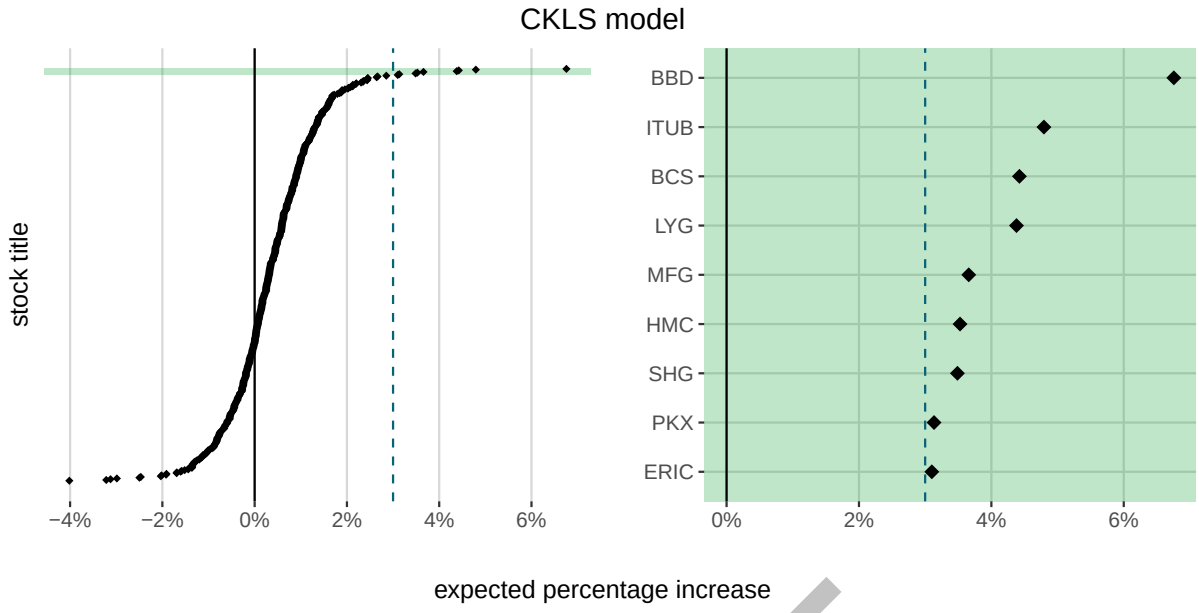



Figure 2: Expected percentage increase and selected titles with CKLS model in July 2022.

| **Models** | 20 days coverage (%) | 10 days coverage (%) |
|---|---|---|
| CIR | 87.74 | 89.68 |
| CKLS | 88.68 | 90.36 |
| GBM | 92.94 | 92.16 |
| OR-UH | 95.71 | 95.48 |

Table 1: Coverage percentage of the 90% empirical confidence intervals for each process of the 20 days and 10 days forecast.

DRAFT

Then we checked the reliability of the percentage increase estimates. For each stock and each month, we estimated an average of the percentage increase fitting a linear regression on the data in the 20 days following the last observed value used by the process, with the following formula:

$$y_t - y_0 = \beta \cdot t \tag{6}$$

where $y_0$ is the last observed value and $y_t$ is the stock observed value $t$ days after $y_0$. It is important to stress that this is not regarded as the true underlying model, but is just a synthetic estimate of the average increase of the process in the 20 following days. Then we derived, for each stock, the expected value 20 days after the last observation and computed the percentage increase.

Plot in Figure 3 reports the expected increase of each stock (on the x-axis) with respect to the actual observed percentage increase (y-axis). We then fitted also a linear regression to highlight whether the model were actually able to predict correctly the increasing trend for each of the four models implemented. As we can see, there is a noticeable correlation between prediction and observations, thus implying that models are predicting correctly, most of the times, the trend of the stocks.

# 4. Conclusions

In this study we assessed the predictive reliability of Stochastic Differential Equations applied to stocks price. We mainly focused on stocks with strong expected percentage increase and on the short term

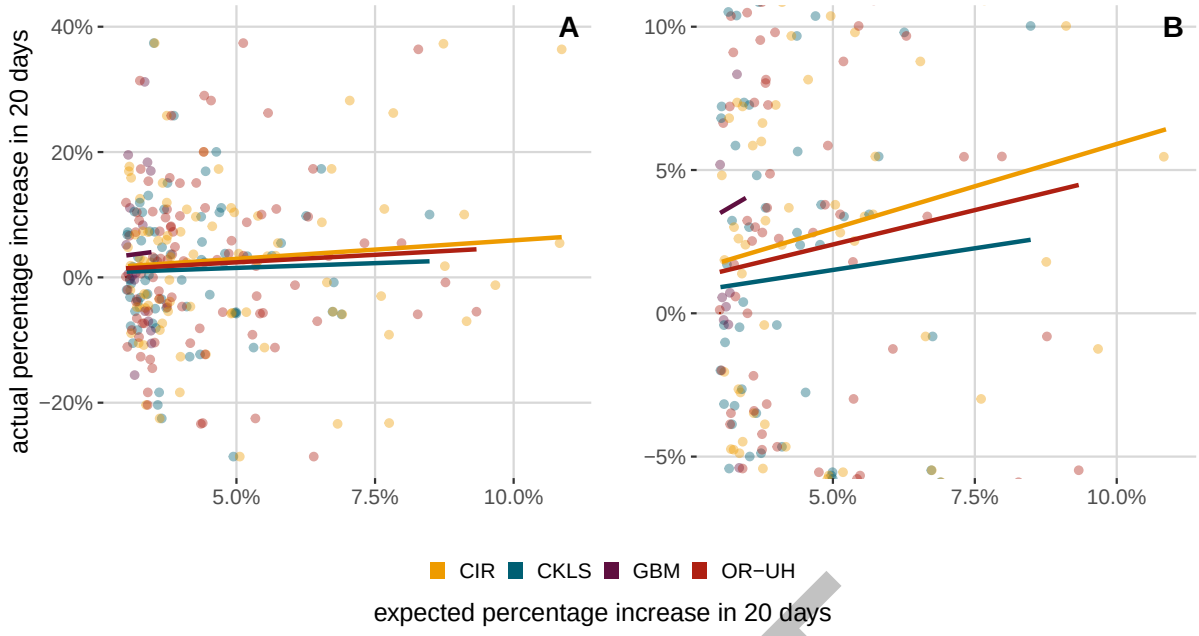



Figure 3: Expected vs. actual percentage increase for each combination of stock, month and model. Complete graph in panel **A** and zoomed graph in panel **B**.

predictions. We developed this analysis in a simplified and feasible framework, selecting four different models (Section 2) applied to each stock independently.

Results show how these SDE models are reliable from a predictive stand point and that are able to capture effectively the increasing trend on a short term analysis. These results are much more impressive if we consider that each stock is analysed individually, without accounting for possible interactions with other ones.

Confidence intervals coverage is very high both for 10 days and 20 days forecasts. It is also noticeable the correlation between the expected and actual percentage increase, as highlighted in Figure 3. Lastly, it was possible to notice how more complex models did not imply a noticeable improvement in the results.

Further developments of this analysis might be the improvement of such predictions by averaging through the different selected models, thus keeping a feasible framework. It could also be useful to take into account other types of stochastic differential equations. Another possible development that might be introduced is to account also for the interaction between different stocks, while applying a penalization to reduce the dimensionality of the adjacency matrix and, thus, of the parameters[10].

DRAFT

# References

[1] John C Cox, Jonathan E Ingersoll Jr, and Stephen A Ross. “An intertemporal general equilibrium model of asset prices”. In: *Econometrica: Journal of the Econometric Society* (1985), pp. 363–384.

[2] Kalok C Chan et al. “An empirical comparison of alternative models of the short-term interest rate”. In: *The journal of finance* 47.3 (1992), pp. 1209–1227.

[3] Cosma Shalizi. “An Introduction to Econophysics: Correlations and Complexity in Finance”. In: *Quantitative Finance* 1.4 (2001), p. 391.

[4] Stefano M Iacus et al. *Simulation and inference for stochastic differential equations: with R examples*. Vol. 486. Springer, 2008.

[5] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[6] Alexandre Brouste et al. “The YUIMA Project: A Computational Framework for Simulation and Inference of Stochastic Differential Equations”. In: *Journal of Statistical Software* 57.4 (2014), pp. 1–51.
[7] Guangqiang Lan, Yunjiao Hu, and Chong Zhang. “The explicit solution and precise distribution of CKLS model under Girsanov transform”. In: *arXiv preprint arXiv:1410.2364* (2014).
[8] Krishna Reddy and Vaughan Clinton. “Simulating stock prices using geometric Brownian motion: Evidence from Australian companies”. In: *Australasian Accounting, Business and Finance Journal* 10.3 (2016), pp. 23–47.
[9] Mohammad Rafiqul Islam and Nguyet Nguyen. “Comparison of financial models for stock price prediction”. In: *Journal of Risk and Financial Management* 13.8 (2020), p. 181.
[10] Alessandro De Gregorio and Francesco Iafrate. “Regularized bridge-type estimation with multiple penalties”. In: *Annals of the Institute of Statistical Mathematics* 73.5 (2021), pp. 921–951.
[11] Shalin Shah. “Comparison of Stochastic Forecasting Models”. In: (2021).
[12] Jeffrey A. Ryan and Joshua M. Ulrich. *quantmod: Quantitative Financial Modelling Framework*. R package version 0.4.20. 2022. URL: https://CRAN.R-project.org/package=quantmod.
[13] Nakahiro Yoshida. “Quasi-likelihood analysis and its applications”. In: *Statistical Inference for Stochastic Processes* 25.1 (2022), pp. 43–60.

DRAFT

# The Added-Worker Effect within Italian Households

Donata Favaro[a], Anna Giraldo[b]

[a] Department of Economics and Management, via del Santo 33, Padova; donata.favaro@unipd.it
[b] Department of Statistical Sciences, via C. Battisti 241, Padova; anna.giraldo@unipd.it

**Abstract**

In this article, we study the relationship between female and male labour supply within Italian households. In particular, by focusing on the relationship between male transitions from employment to unemployment and—at the same time—female partner transitions from labour market inactivity to activity, we evaluate the Added-Worker Effect in Italy at the household level. The analysis is carried out over a long period of time—between 2004 and 2019—on data from the Italian Labour Force Survey (ILFS). To identify the Added-Worker Effect, we adopt a differences-in-differences methodology. By exploiting the richness of information contained in the ILFS data on unemployment status and unemployment risk, we were able to evaluate different "dimensions" of the Added-Worker Effect.

**Keywords:** household labour supply, differences-in-differences, female labour market participation

DRAFT

## 1. Introduction

The aim of the present article is to study the relationship between female and male labour supply within Italian households. In particular, we evaluate whether an Added-Worker Effect (AWE) exists in Italy, focusing on the transition of women from labour market inactivity to unemployment when their male partners move from employment to unemployment.

The literature on the topic goes back to the first contributions of Humphrey (1940) and Woytinsky (1940) and the empirical studies on the Added-Worker Effect (AWE) by Mincer (1962), Heckman and Macurdy (1980), Lundberg (1985). The AWE has been revived with the recent economic crisis. The latest contributions seem to agree much more on the existence of an AWE. Gong (2011), focusing on Australia, found a significant AWE in terms of increased full-time employment and working hours. Bredtmann et al. (2014) investigated the AWE across the European countries (28 countries) in the period 2004-2011 using the European Union Statistics on Income and Living Conditions survey (EU-SILC) and showed that an AWE exists, both at the extensive and at the intensive margin of labour supply. Hardoy and Schøne (2014) found no support for the AWE in Norway, although the AWE was detected in some subsamples. Starr (2014) found, for the USA, that employment rates of women whose husbands were non-employed rose significantly during the recession. Ayhan (2018) showed that the probability of participation in the labour force of Turkish women increases by 15-28% in response to their husband's unemployment. As to the Italian case, facts show that during the recent recession female employment in Italy has increased and has partially counterbalanced the increase in male unemployment (Istat, 2013). According to data from the World Bank the ratio of female to male participation rate increased in Italy over the period 2010-2020. The existing empirical analyses have already provided some evidence of this counter-cyclical trend in female employment compared to the reduction in male employment. Ghignoni and Verashchagina (2016) found that an AWE exists, even if only in cases of serious hardship. More recently, Baldini et al. (2018), studying Italy in the years 2004-2014 using EU-SILC, found a strong and robust evidence that households hit by an employment shock do respond by increasing labor supply. Then, they documented an AWE effect in Italy that affects not only wives but also

teenage children.

Our study evaluates the AWE in Italy during a long period of time—2004 to 2019—by employing the Italian Labour Force Survey (ILFS), a rotating panel provided by the Italian Statistical Institute (Istat). To identify AWE, we employ a differences-in-differences methodology (DD).

## 2. Data and methodology

Our study evaluates the AWE in Italy over a long period of time—between 2004 and 2019—by employing the ILFS, a rotating panel provided by the Istat. The longitudinal data of the ILFS observe individuals across couples of years ($t_0$, $t_1$): in the quarter of entrance in the panel, in the subsequent one (first two quarters of observation), and in the 5th and 6th quarter. In each quarter, new individuals enter the survey, for a share of one fourth of the total sample. The available data are from 2004-05 to 2018-19. Unfortunately, until 2012 Istat only makes available the panel data related to the first quarter of each year. Thus, for reasons of balance and comparability between samples, we carried out our analysis on the first quarter only. This means that our database is made by 15 panels of individuals observed in the first quarter of year $t_0$ and in the first quarter of year $t_1$.

Our analysis focuses on couples—married or cohabiting—with or without children, with partners not retired and not unable to work, in the age range 25-54. To our purpose, we focus on households with unchanged composition in the two occasions ($t_0$ and $t_1$). It is worth mentioning that the original longitudinal data are individual panels without information on the marital status and the relationships within the household. Thanks to the availability of a common household identifier and on the basis of individual characteristics, family relationships were reconstructed.

To identify AWE, we employ a differences-in-differences methodology (DD). Our first definition of treated women includes those women whose partners became unemployed between $t_0$ and $t_1$. Then, by exploiting the richness of the ILFS, we defined a broader group of treated, by including also those men who moved from employment to forms of job protection and those who experienced a reduction in activity or lost jobs other than the main one between $t_0$ and $t_1$. This broader definition of treatment, which is new in the literature, allowed us to consider situations that might reveal an increased risk of losing one's job or a reduction in the available income, which may affect the decision of female partners to enter the labour market.

AWE occurs when the probability of changing employment status from inactive to unemployed or employed is significantly different between treated and untreated women. Then, the equation we estimated to detect AWE is as follows[1]:

$$ES_{it} = \beta_0 + \beta_1 D_t + \beta_2 T_i + \beta_3 T_i D_t + \beta_4 X_{it_0} + \varepsilon_{it} \quad (1)$$

where $ES_{it}$ is the employment status of female $i$ at time $t$. $D_t$ is a dummy with value equal to 0 in $t_0$ and 1 in $t_1$. $T_i$ is a dummy that captures whether the woman is treated or not (1 if treated, 0 if not). $D_t T_i$ is our variable of interest: the parameter $\beta_3$ captures the effect of being treated, compared to not being treated, on the change of employment status of females. $X_{it_0}$ includes several covariates, all evaluated at $t_0$. Among the covariates, we include: cultural proxies (female age cohort and male nationality), female educational level, male educational level, male type of job, male sector of activity, number of children, number of children who work, number of children NEET, number of children under 15 years old and dummies that capture male unemployment risk.

Equation 1 is estimated under different specifications of the treatment ($T_i$) and the outcome ($ES_{it}$).[2] The first treatment we consider is the partner's transition from employed to unemployed. In this case, $T_i$ assumes value zero if the man is employed and 1 if he is unemployed. We then consider other categories of the man's employment status to define the treatment. The aim here is to analyse how the emergence of the partner's risk of losing his job or a change in his economic situation can influence the woman's employment choices. To capture the emergence of a risk in the partner's employment stability, we have considered men's transitions (between $t_0$ and $t_1$) from employment to CIG. CIG stands for Cassa Integrazione Guadagni (Wages Guarantee Fund) and is an institution under Italian law consisting of an economic benefit, provided by the Italian Security System, for workers suspended from the obligation to perform work or working reduced hours. CIG

[1] We follow Angrist and Pischke (2009) and employ a linear model.

[2] In the next section, we will provide a precise description of all the definitions we have adopted for *ES* and *T*.

DRAFT

is a cash grant the Italian Security System pays to workers when companies are in temporary difficulty. In order to capture the effects of a worsening in the economic condition of the family due to a reduction in men's work, we consider men's transitions from working full hours in $t_0$ to reduced hours in $t_1$ and men losing jobs other than the main one between the two years. To complete the analysis, we of course also consider male transitions between the two years from employment to non-employment.

With regard to women's employment transitions, the first objective is to evaluate the AWE in its traditional version, assessing the transition of women from inactive to unemployed. In this case, the outcome variable $ES_{it}$ takes a value of 1 if the woman is unemployed, and 0 if she is out of the labour force. After that, we use further specifications of the outcome variable to capture both different 'degrees' of transition to activity and the effect on labour supply. We evaluate the transition from inactivity not searching to inactivity searching or unemployment by assigning the outcome variable the value zero if the woman is inactive and not searching for work and the value 1 if the woman is inactive but looking for work, or unemployed. Then, we assess the transition from inactivity to unemployment or employment by assigning the outcome variable value 1 if the woman is either unemployed or employed. This specification allows us to detect the extensive margin effect of treatment. Differently, to evaluate the intensive margin effect of treatment, that is the effect on labour supply, we consider only women who already work and observe the transition from part-time to full-time. In this case, the outcome variable $ES_{it}$ takes value zero if the woman works part-time and value 1 if she has a full-time job.

## 3. Results

In this Section, we discuss the results of the estimates of Equation 1 for the different work transitions of women and their partners (see Table 1). Specifically, *T1* is the treatment variable that captures men's transitions from employment towards unemployment and takes on a value of 1 if the man becomes unemployed and a value of 0 if he remains employed. *T2* captures men's transitions from employment to extended CIG or transitions to reduced activity, including job losses other than their primary job. *T3* relates to men's transitions from employment to non-employment—either unemployment or inactivity.

Table 1: The AWE with different treatments and outcomes

| Classes | T1 | T2 | T3 |
|---|---|---|---|
| ES1 | 0.149*** | 0.023* | 0.060*** |
| ES2 | 0.123*** | 0.020 | 0.051*** |
| ES3 | 0.138*** | 0.020 | 0.093*** |
| ES4 | 0.038*** | 0.027*** | 0.028*** |
| ES5 | 0.014 | 0.051*** | 0.023 |

[a] T1: men's transitions from employment to unemployment; T2: men's transitions from employment to CIG /reduced activity/lost jobs other than the main; T3: men's transitions from employment to non-employment. ES1: females' transitions from inactivity to unemployment; ES2: females' transitions from inactivity to unemployment or employment; ES3: females' transitions from inactivity "not searching" to inactivity "searching" or unemployment; ES4: females' transitions from employment "not wishing more hours" to employment "wishing more hours of work"; ES5: females' transitions from part time employment to full time employment.

For women, we adopted different definitions of the dependent variable *ES*, so as to capture different 'degrees' of exit from inactivity and changes in preferences for work involvement. *ES1* is the dependent variable definition that captures women's transitions from inactivity to unemployment; *ES1* takes on a value of 0 if women are inactive in either period and a value of 1 if they are unemployed in $t_1$. *ES2* evaluates transitions from inactivity to the labour force - either unemployed of employed; in this case, the dependent variable assumes a value of 1 in period $t_1$ if the woman enters the labour force between the two periods. *ES3* captures the intention to work and the change of the status from 'not searching' to 'searching' for work. Thus, the sample of women is restricted to the only ones who are inactive and 'not searching' for work in period $t_0$; then, the dependent variable *ES* assumes a value of 1 if the woman is 'searching' for work in period $t_1$ and a value of 0 otherwise, either in period $t_0$ or $t_1$. *ES4* assesses preferences for greater work involvement, i.e. more working hours; in this case, the sample consists of women who are already employed in $t_0$ and who do not wish more hours of work. The dependent variable takes the value 1 in period $t_1$ if women state that they wish more hours of work. We also considered a further specification to assess preferences for greater work

DRAFT

involvement. In this case, we selected only women with a part-time job in period $t_0$ and looked at the transition from part-time to full-time work between the two years; in this specification the variable *ES5* takes the value of 1 in period $t_1$ if women are employed full-time in that period and the value of 0 otherwise.

The results show that an AWE exists within Italian families. The classical definition of AWE, measured by women's transitions from inactivity to unemployment as their partners move from employment to unemployment, compared to women whose partners do not move to unemployment (ES1-T1), is significant and positive. The effect is also positive when the treated group includes only families whose men are "at risk of unemployment" (ES1-T2), meaning that females react to changes in the family economic situation. Interestingly, working women are willing to work more hours (ES4) when the partner loses his job or he is at risk of losing it; this effect is significant for all treatment types. However, the transition from part-time to full-time is weakly significant, only for T2 transition. Possible explanations for this result could be the existence of labour market rigidities in the transformation of part-time work into full-time work or constraints on the supply side of the labour market (due to care activities that limit women's working hours).

**Acknowledgments**: the authors acknowledge the financial support provided by the PRIN project «The Great Demographic Recession – GDR» financed by the Italian MIUR under the PRIN 2017 research, grant agreement n. 2017W5B55Y-003 (PI: Daniele Vignoli).

## References

[1] Angrist J., Pischke S.: Mostly harmless econometrics, an empiricist's companion. Princeton University Press, Princeton (2009)
[2] Ayhan, S., H.: Married women's added worker effect during the 2008 economic crisis. The case of Turkey Rev Econ Household (2018) doi: 10.1007/s11150-016-9358-5
[3] Baldini, M., Torricelli, C., Urzì Brancati, M.C.: Family ties: Labor supply responses to cope with a household employment shock, Rev Econ Household (2018) doi: 10.1007/s11150-017-9375-z
[4] Bredtmann, J., Otten, S., Rulff C.: Husband's unemployment and wife's labor supply – The added worker effect across Europe, Ruhr Economic Papers N. 484 (2014)
[5] Ghignoni, E., Verashchagina A.: Added worker effect during the great depression: evidence from Italy, Int J Manpow (2016)
[6] Gong, X.: The added worker effect for married women in Australia, Economic Record, 87: 414-426, (2011) doi: 10.1111/j.1475-4932.2011.00719.x
[7] Hardoy, I., Schøne, P.: Displacement and household adaptation: insured by the spouse or the state? Journal of Population Economics (2014) doi: 10.1007/s00148-013-0469-5
[8] Heckman, J.J., Macurdy, T.: A life cycle model of female labour supply, Rev Econ Stud, 47, 47-74 (1980)
[9] Humphrey, D.D.: Alleged "additional workers" in the measurement of unemployment, Journal of Political Economy (1940)
[10] Istat: Rapporto Annuale: La Situazione del Paese. Rome, Istat (2013). Available on: www.istat.it
[11] Lundberg, S.: The added worker effects, Labour Econ, 3, 11-37 (1985)
[12] Mincer, J.: Labor force participation of married women: A study of labor supply. In: Universities-National Bureau Committee for Economic Research Aspects of Labor Economics. Princeton University Press (1962)
[13] Woytinsky, W.S.: Additional workers on the labor market in depressions: A reply to Mr. Humphrey, J Political Econ, 48, 735-739 (1940)

DRAFT

# A model for the natural history of breast cancer: application to a Norwegian screening dataset

Laura Bondi[a], Marco Bonetti[b], and Solveig Hofvind[c]

[a]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, and Dondena Research Center, Bocconi University, Milan, Italy; laura.bondi@mrc-bsu.cam.ac.uk
[b]Department of Social and Political Sciences, Dondena Research Center, and Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy; marco.bonetti@unibocconi.it
[c]Section of Breast Cancer Screening, Cancer Registry of Norway, Oslo and Department of Health and Care Sciences, UiT The Arctic University of Norway, Tromsø, Norway; sshh@kreftregisteret.no

DRAFT

**Abstract**

In this work we present some preliminary results on the analysis of data collected by BreastScreen Norway to estimate the natural history of the disease. Learning about the disease occurrence and evolution is crucial to identify the optimal screening schedule, with respect to the age range of the invited women and the lag between successive examinations. The model is a multi-state semi-Markov model with a cure rate structure, where the main quantities of interest are the probability of developing the disease, the age at start of asymptomatic detectability of the disease and the sojourn time, i.e. the time interval during which the disease is screen-detectable but not yet symptomatic.

***Keywords:*** approximate Bayesian computation (ABC), breast cancer, cure rate model, disease history, multi-state model

## 1. Introduction

Cancer screening is defined as the examination of asymptomatic subjects in order to detect tumours before they become evident because of symptoms (1). In the past decades screening programs have been implemented in many countries, therefore making randomized trials difficult to perform and those performed difficult to evaluate due to changes in screening techniques and diagnostic tools. As a consequence, researchers can only rely on observational data collected administratively to learn about the natural history of the disease and to identify the optimal screening policy (in terms of the age range of the women invited, and lag between consecutive examinations), and risk-based tailoring of their schedules.

In this work, we apply a multi-state semi-Markov model proposed in (2) to the breast cancer screening data from Norway (BreastScreen Norway). This model aims at reconstructing the latent process of occurrence and development of breast cancer.

All times are measured from birth of the woman. For those women who do experience the disease, we assume that after the onset of the disease there is a time interval in which not even a screening examination is able to detect the presence of the disease (see Figure 1). The two main quantities of interest are the time to the start of asymptomatic detectability (through screening) of the disease (which we denote by $T_A$) and the time to the symptomatic detection of the disease (denoted by $T_S$). Between

time $T_A$ and $T_S$ the tumour can only be detected through screening (the "sojourn time," denoted by $\Delta$), while at time $T_S$ the disease becomes evident because of symptoms. In other words we have $T_S = T_A + \Delta$. Further, we assume that symptomatic detection occurs exactly when the first symptoms appear.

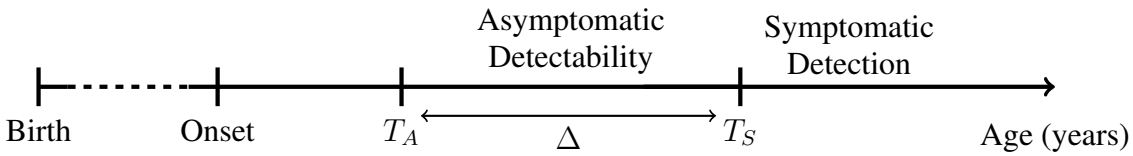



Figure 1: A graphical representation of the natural history from onset until detectability of the disease.

While studying the latent evolution of the disease, we are also interested in studying the probability of insurgence of the disease in a woman's lifetime. For this purpose, the model presents a cure rate structure, i.e. only a proportion of women, which we call the "susceptible proportion", denoted by p with $p \in (0, 1)$ will experience the event of developing breast cancer. Note that these should be considered to be latent cases and not necessarily observed cases.

Given by the limited follow-up, the lag between screening examinations and the impossibility to observe $T_S$ when the tumour is screen-detected, this problem presents a complex missing data structure, which makes the observed data likelihood (3) of the model complicated if not intractable.

DRAFT

## 2. The Norwegian breast cancer screening data

Breast cancer screening in Norway was first introduced in 1995 (and became nationwide in 2005) and targets women aged 50-69 for biennial mammographic screening. We focus on a cohort of women born between 1948 and 1952 without any breast cancer diagnoses before entering the screening program at age 50 (with a small variability due to the invitations rounds). For these women we have information on their screening invitations, attendance and result for each examination, possible date of breast cancer diagnosis (DCIS or invasive), type of detection (in-screening or symptomatic), some additional tumour characteristics, as well as on a number of covariates collected through a questionnaire.

The screening invitations stop at age 69 but the breast cancer diagnoses are updated from the Cancer Registry until May 2022. Therefore the lentgh of follow-up is between 20 and 24 years. We focus only on invasive cancers and on three binary covariates, which divide the women in eight groups as shown in Table 1: having had at least one birth $X_1$ (0=no, 1=yes); level of education $X_2$ (0=low, 1=high); and family history of breast cancer $X_3$ (0=no, 1=yes). These are indeed the three non-race main risk factors for breast cancer (among women with no previous history of the disease) (4).

Table 1 also shows the sample size, the proportion of observed detections, the proportion of symptomatic detections and the mean age at asymptomatic/symptomatic detections for each covariate group.

| Group | $(x_1, x_2, x_3)$ | Size | % Dx | % Symp Dx among all Dx | Mean age Asymp Dx | Mean age Symp Dx |
|---|---|---|---|---|---|---|
| 1 | (0,0,0) | 1240 | 5.2% | 27% | 63.3 | 65.6 |
| 2 | (0,0,1) | 351 | 7.4% | 35% | 63.1 | 63.2 |
| 3 | (0,1,0) | 3731 | 4.9% | 39% | 62.9 | 65.3 |
| 4 | (0,1,1) | 1329 | 9.0% | 38% | 63.1 | 64.0 |
| 5 | (1,0,0) | 16241 | 3.8% | 28% | 63.8 | 66.0 |
| 6 | (1,0,1) | 4627 | 5.7% | 36% | 64.1 | 66.2 |
| 7 | (1,1,0) | 39330 | 4.2% | 33% | 63.7 | 65.1 |
| 8 | (1,1,1) | 13252 | 5.7% | 35% | 64.0 | 65.0 |
| Total | | 80101 | | | | |

Table 1: Observed outcomes in each covariate group. Ages are measured in years. $X_1$= at least one child birth (0:No, 1:Yes); $X_2$=Education level (0:Low, 1:Medium/High); $X_3$=Family history of breast cancer (0:No, 1:Yes).

## 3. Model and preliminary results

Recall the three binary covariates described in Section 2: $X_1$ ="at least one birth," $X_2$ ="high level of education," and $X_3$ ="family history of breast cancer," all coded as 0 = No and 1 = Yes. The susceptible proportion is modelled as function of the observed covariates $\boldsymbol{x} = (x_1, x_2, x_3)$ through the logit link:

$$p(\boldsymbol{x}) = \frac{e^{p_0+p_1x_1+p_2x_2+p_3x_3}}{1+e^{p_0+p_1x_1+p_2x_2+p_3x_3}}.$$

For the subjects who will eventually develop the disease, the disease evolution is described through the joint distribution of the couple $(T_A, \Delta)$. The model assumes that the mean of $T_A$ depends on the covariates linearly, and that the variance of $T_A$ is constant across covariate groups. The distribution of $\Delta$ is defined conditionally on the observed value of $T_A$, and it may depend on the covariates but only indirectly (see below). The definition of the model for the disease process is as follows:

$$T_A \mid \beta_0, ..., \beta_3, \sigma \sim 100 \cdot \text{Beta}(\alpha, \beta);$$
$$\Delta \mid \{T_A = t_A\}, \lambda_1, \lambda_2, \lambda_3 \sim \text{Exp}(\lambda_1 \cdot 1(t_A \leq 55) + \lambda_2 \cdot 1(55 < t_A \leq 65) + \lambda_3 \cdot 1(t_A > 65)),$$

where $E(T_A) = \mu(\boldsymbol{x}) = 100 \cdot \frac{\alpha}{\alpha+\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, $\sigma^2 = 100^2 \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. This model, called "Rescaled beta + piecewise exponential" model, has been introduced in (2), where it was selected among ten competing models to describe breast cancer natural history. We refer to this reference for additional details on the model, such as the prior distribution for the 12 parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma, \lambda_1, \lambda_2, \lambda_3, p_o, p_1, p_2, p_3)$.

Approximate posterior distributions for the model parameters are obtained via likelihood-free inference. More specifically, relying on approximate Bayesian computation (ABC) allows us to avoid deriving the complex observed data likelihood of the model. Details on the metric function employed to measure the distance between the real and generated data are not shown here, but a complete description, together with a discussion of the advantages and drawbacks of ABC in this setting, can be found in (2). The results presented here are based on 200,000 generated datasets.

Figure 2 shows the boxplots of the posterior distributions for the mean of $T_A$, $\mu(\boldsymbol{x})$, and for the susceptible proportion $p(\boldsymbol{x})$ across the eight covariate groups. The left panel of Figure 2 highlights that women with at least one child tend to experience breast cancer later than those without children. From the right panel, instead, clearly emerges that having a family history of breast cancer is associated with an increased risk of belonging to the group of the susceptible subjects. Note that the posterior distributions of the p's are shifted toward slightly higher values than expected from previous studies (see e.g. (5)). This is likely due to the difficulties in estimating a weakly identified cure rate model with limited follow-up, i.e. little data at higher ages.

Given the posterior distributions, one can then compute the predictive distributions for the quantities of interest, $T_A$ and $\Delta$ (see Figure 3). The median predicted values for $T_A$ vary between 69.2 (in group 4) and 72.5 (in group 6). The standard deviation of $T_A$ is estimated to be around 7.5 years. The predictive distributions for $\Delta$ in the three groups defined by the value of $T_A$ have medians 2.4, 4.5 and 0.9 years, respectively. While the first two appear to be in line with what is known by previous studies, the predictive distribution for $\Delta$ in the third group does not seem very reliable. Similarly to what was highlighted when commenting the posterior distributions for the p's, this can be probably attributed to the lack of information about $T_A$ for tumours occurring after the age of 65, given that screening examinations stop at age 69 .

## 4. Conclusions

We have estimated a parametric model to describe the insurgence and the evolution of breast cancer, where the main quantities of interest are the probability of developing the disease ($p$), the start of asymptomatic detectability of the disease and the time of symptomatic detection ($T_A$ and $T_S$). Given the estimated latent disease process, it is possible to simulate the effect of different screening schedules, possibly tailored to the individual risk factors, on the number and kind of diagnoses in the populations.

DRAFT

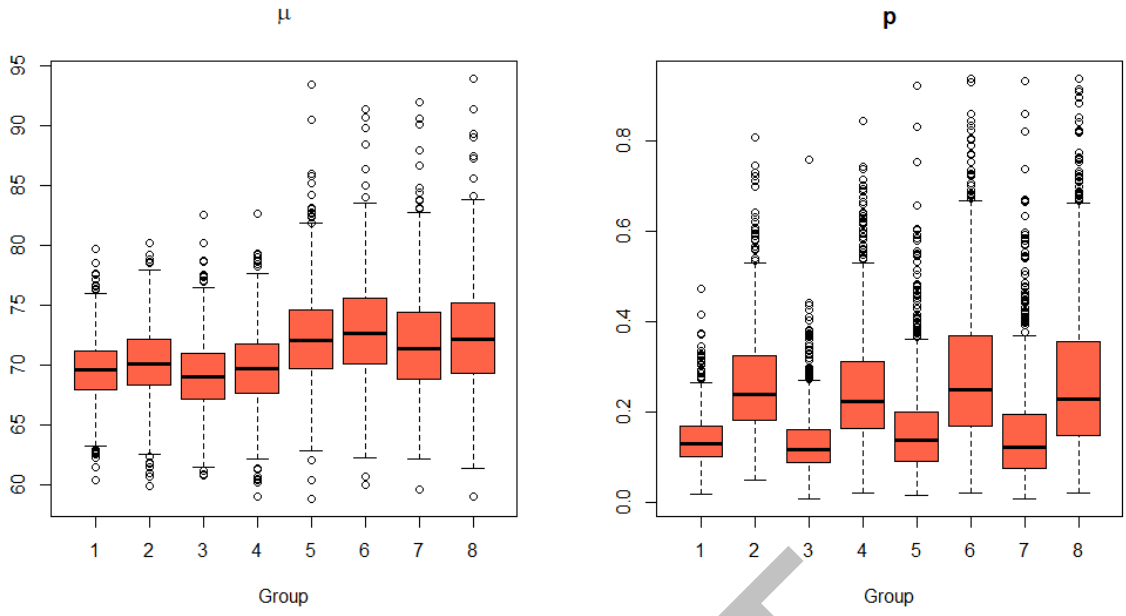



Figure 2: Approximate posterior distributions of the mean age at asymptomatic detectability $\mu(\boldsymbol{x})$ and of the susceptible proportion $\mathrm{p}(\boldsymbol{x})$ across covariate groups.

This work might be extended by including in the model a parameter for the sensitivity of the screening examinations. Indeed, the assumption of a 100% sensitivity, as we made here, corresponds to setting the probability of false negative results equal to zero, which is not realistic (6). Moreover, additional covariates, such as breast density, could be included in the model.

Other interesting analyses could be based on different cohorts of women included in the BreastScreen Norway database, which could confirm the estimated disease process or highlight differences in the occurrence between different generations.

Also, assuming a stable disease population (in which the rate of births and the distribution of ages at tumour onset are constant across calendar time (7)), one could analyse a larger cohort of women at once to exploit the information provided by women at different ages and therefore to obtain a more precise inference. This would however come with a higher computational cost and a longer time needed to perform the ABC simulations.

Future developments of this work will also involve the numerical computation of the observed data likelihood function to perform exact Bayesian inference through MCMC.

DRAFT

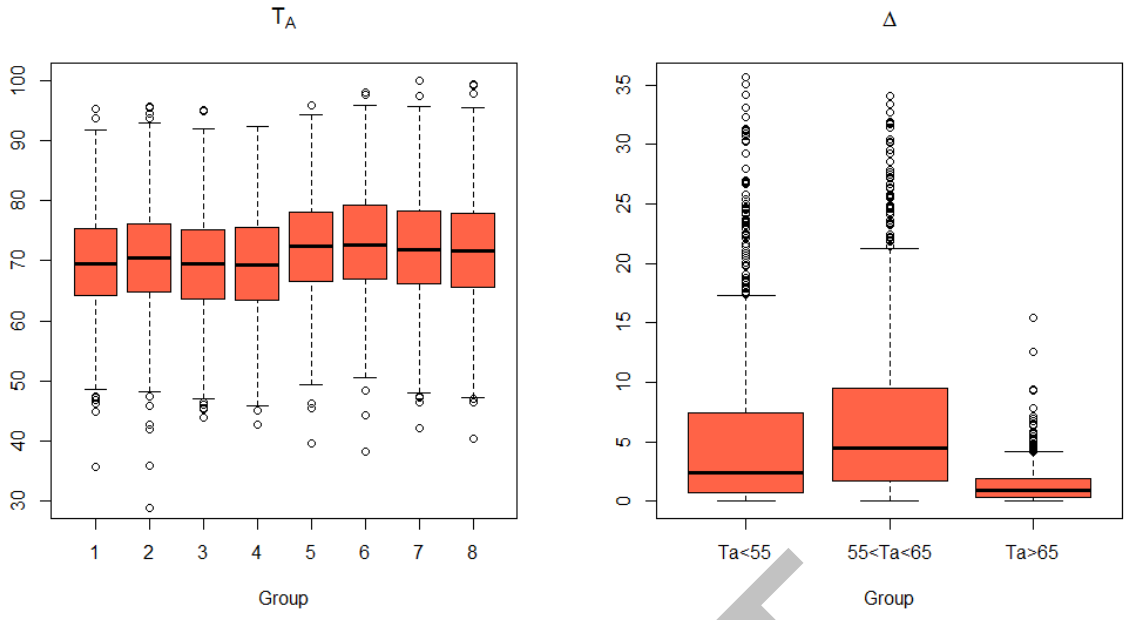



Figure 3: Approximate predictive distributions of the age at asymptomatic detectability $T_A$ and of the sojourn time $\Delta$ across covariate groups.

DRAFT

## References


[1] Van Oortmarssen G, Boer R and Habbema J. Modelling issues in cancer screening. *Statistical Methods in Medical Research.* 1995; 4(1): 33–54.

[2] Bondi L, Bonetti M, Grigorova D and Russo A. Approximate Bayesian Computation (ABC) for the natural history of breast cancer, with application to data from a Milan cohort study. *Statistics in Medicine.* 2023; In press.

[3] Little R and Rubin D. *Statistical analysis with missing data.* Second ed. New York: Wiley, 2002.

[4] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute.* 1989; 81(24):1879-86.

[5] Howlader N, Noone AM, Krapcho M, et al. SEER Explorer. Breast Cancer-Stage Distribution of SEER Incidence Cases, 2007-2016 by Sex. *National Cancer Institute, Bethesda*. 2019.

[6] Abrahamsson L and Humphreys K. A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. *Statistical Methods in Medical Research.* 2013; 25(4): 1620–1637.

[7] Isheden G and Humphreys K. Modelling breast cancer tumour growth for a stable disease population. *Statistical Methods in Medical Research.* 2019; 28(3): 681–702.

# Generalized Bayesian Ensemble Survival Trees: an extension to categorical variables to apply it to real data

Elena Ballante[a,b]

[a]Department of Political and Social Sciences, University of Pavia, Italy;
elena.ballante@unipv.it
[b]BioData Science Unit, IRCCS Mondino Foundation, Pavia, Italy

**Abstract**

Survival analysis is a common framework between different fields of application. The availability of statistical techiniques that allow to model the data at hands is a foundamental area of research. The problem is especially urgent if we consider the need of analyse small dataset, case that is extremely common in biomedical research.
In this paper, a proper Bayesian bootstrap method for ensemble survival tree models is extended to handle categorical variable, in order to apply to real data. Empirical results are shown in a simulated study that shows the potential of the method in terms of improvement of the performances and stability of results. The application on Amyotrophic Lateral Sclerosis patients shows the advantages of the application of the proposed method in real world problems.

***Keywords:*** Survival analysis, Bootstrap, Bayesian nonparametric learning, ensemble models

DRAFT

## 1. Introduction

In a lot of different fields we can encounter problems that require the analysis of time-to-event data. This field of statistical analysis is been developed since 17th century, but still lack of the variety of models that characterize regression and classification framework. This problem especially emerge if we consider the availability of computational tools that make it easily applicable.
Moreover, a very common problem especially in biomedical application is the availability of small dataset. The use of models that can handle this kind of data obtaining stable and reliable is extremely important in that field.
The use of resempling methods to improve models is not new also in the framework of the survival analysis, as in bagging methods (1) and random forest model (2).
The most common resempling method (and the one used in applications) is Efron's bootstrap (3), but some methodological work start to apply different bootstrapping approach in bagging models and random forests. Rubin's bootstrap, also called Bayesian bootstrap, was defined in (4) and consider for this kind of model in (5) and (6). A further bootstraping method based on nonparametric Bayesian statistics, called proper Bayesian bootstrap, was defined in (7) and embedded in predictive models in (8) and (9). The models defined were applied in regression and classification problems, whilst none of these was extended to survival framework.
The aim of this paper is to fully develop the model proposed in (10), extending the bootstrap techniques

to handle categorical variable. With this addition, the model is now available for the application to real world problem.
The paper is organized as follow: Section 2. describes the methodological approach and the novelty proposed, Section 3.describes the simulation setting and the results, Section 4.shows an application to a real dataset and Section 5.contains the conclusions.

## 2. Methodology

The aim of this paper is to extend the proper Bayesian bootstrap bagging tree survival model (10) to manage categorical covariates in order to be able to apply to real data.
As described in more details in (9) and (10), we apply the Proper Bayesian Bootstrap in order to sample from a posterior distribution with Dirichlet distributed weights. The bootstrap resample is generated by a mixture distribution of a prior guess $F_0$ and the empirical distribution $F_n$. The weights of the prior will be referred to as $w$. In this work, each covariate is sampled independently from the others. The new value of the response variable associated to the observations sampled from the prior is generated by an appropriate predictive model. Then, a survival tree model is trained on each of the B bootstrap replicates. The training part of the model described so far is summarized in Algorithm 1.
The prediction step of the algorithm is defined on the basis of (1) and it is explained in Algorithm 2. Observing a new observation $\mathbf{x}_{new}$, the estimated conditional survival function $\hat{S}$ is computed aggregating all the training observation that fall in the same leaves as $\mathbf{x}_{new}$: $\hat{S}_A^B(\cdot|\mathbf{x}_{new}) = \hat{S}_{L_A^B(\mathbf{x}_{new})}(\cdot)$.

**GBEST: Generalized Bayesian Ensemble Survival Trees**

**Input:** Training set $T$
**for** *b in 1:B* **do**
  Sample $(\mathbf{x}_1^*, y_1^*), ..., (\mathbf{x}_m^*, y_m^*)$ from $(k+n)^{-1}(kF_0 + nF_n)$;
  Draw $\mathbf{w}^b$ from $D(\frac{n+k}{m}, ..., \frac{n+k}{m})$;
  Get $\phi^b = \phi(\mathbf{w}^b)$ running weighted survival tree on the new sample $\mathbf{x}_1^*, ..., \mathbf{x}_m^*$
**end**

**Algorithm 1:** Training phase of the GBEST model

**Input:** trained survival trees, test set T
**for** *i in 1:nrow(B)* **do**
  Aggregate the observations $(\mathbf{x}_k^*, y_k^*)$ in the leaves where $\mathbf{x}_{test}$ is fallen
  Predict the conditional survival function as $\hat{S}_{L_A^B(\mathbf{x}_{new})}(\cdot)$
**end**

**Algorithm 2:** Aggregation and prediction phase of the GBEST model

### 2.1 Categorical variables

We consider two different approaches for the bootstrap resampling of the categorical variables.
The first and most common approach is described in (11) and in this paper we will referred to it as *empirical* sampling. The prior distribution is estimated as a categorical distribution where each class has an estimated probability of $p_i = \frac{N_i}{N}$, where $N_i$ is the number of observations of that class and $N$ is the total sample size.
The second approach, inspired by oversampling techniques, assigns to each class the same probability $p_i = \frac{1}{n_{class}}$, without making distinction between more and less represented class in the data. This approach will be referred to as *uniform* sampling.

DRAFT

## 3. Results of the simulated study

In order to show the performances of the proposed method in a controlled setting, a simulation study is performed.
The simulated datasets contain 5 covariates with different balance between categorical and numerical variables. The numerical covariates are simulated as $U(0,1)$, while categorical covariates are sampled with different weights assigned to each categories. Survival times are simulated deploying a Weibull distribution and covariates are associated to the survival time with betas equal to 1. The censorship is fixed to $10\%$ and the sample size is $N = 50$. The number of trees in the GBEST model is set as $B = 100$. The performances are evaluated in terms of integrated Brier score (IBS) in a train-test framework ($75\%$ and $25\%$) and a total number of 100 simulation is performed in order to assess the stability of the results. Mean values and non parametric confidence intervals of the resulting IBSs are presented.

### 3.1 Different kind of sampling

The influence of two different sampling strategies, as described in Section 2.1, is investigated. The simulated datasets are composed of five variable and only one is dichotomous and unbalanced (the probabilities of each category are 0.25 and 0.75). To stress the results, the $betas$ parameters to generate the survival times are set as 1.5 to the categorical variable and 0.5 for the others. The rest of the setting is analogue of the one described in Section 3. The results are depict in Figure 1.

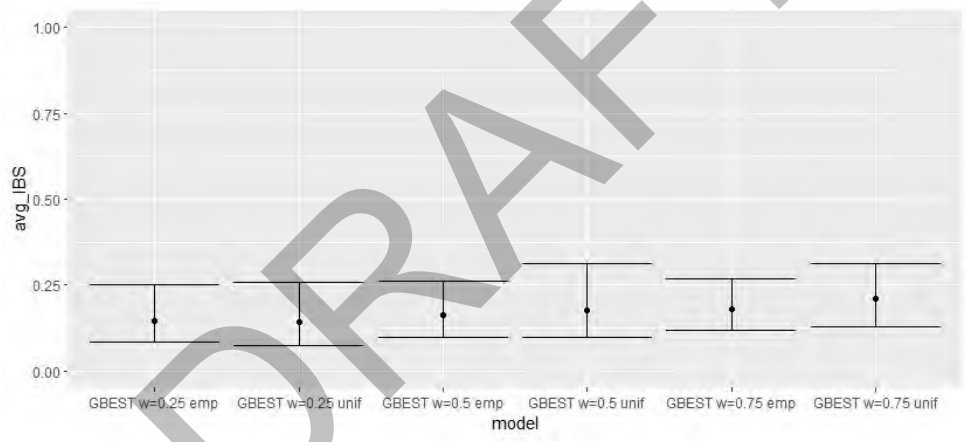



Figure 1: Comparison of mean and non parametric confidence interval of IBS between the two sampling strategies with three different choices of weight

No significant differences between mean IBS values are observed, but a general is showed. The empirical sampling remains stable in terms of length of the confidence interval, instead the uniform sampling shows an increasing variability where the proportion of observation sample from the prior is high ($50\%$ and $75\%$). The empirical sampling is adopted in the next analysis.

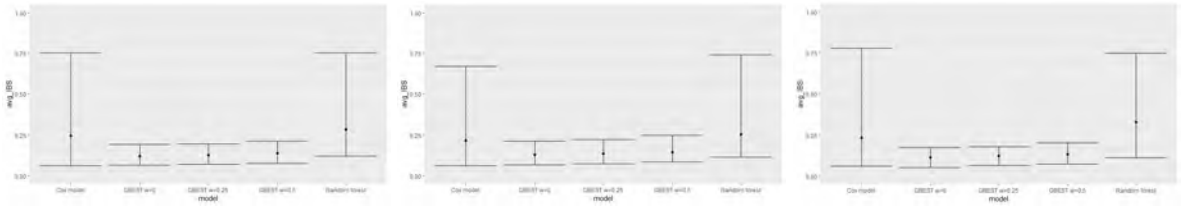
### 3.2 Comparison with literature models

After the sensitivity analysis, the proposed model is compared with the most common models in the survival analysis: the Survival Random Forest (2) and the Cox model (12). In order to obtain more robust results, the analysis are repeated with three different setting of the original dataset.

- First scenario. One categorical variable with 3 balanced levels. The other four variables are generated as $U(0,1)$.
- Second scenario. Three categorical variables. One dichotomous and balanced, one with three balanced levels and one dichotomous and imbalanced ($prob = 0.25, 0.75$). The other variable are generated as $U(0,1)$.

DRAFT

- Third scenario. Five categorical variables. One dichotomous and balanced, one with three balanced levels and three dichotomous and imbalanced variables ($prob = 0.25, 0.75$, $prob = 0.4, 0.6$, $probab = 0.3, 0.7$).



(a) First scenario. (b) Second scenario. (c) Third scenario.

Figure 2: Comparison of mean and non parametric confidence intervals for IBS obtained in cross validation for the 100 simulated datasets in three different scenarios for the proposed model with difference choices of weight and literature models (Random Forest and Cox model)

The GBEST model performs better than the Random Forest and Cox model in all three scenarios, both in terms of averaged error and in terms of stability of the results. The GBEST with $w = 0.5$ seems to introduce a bit of noise with respect to $w = 0.25$ and $w = 0$ but still outperforms the most common survival models.

DRAFT

## 4. Real data application

To investigate the performances of the proposed method, we show an application to a real world problem in the context of biomedical research.
The data analyzed were collected at IRCCS Mondino Foundation. The predictive power of retinal measurement, coupled with clinical information, is investigated in terms of survival time of Amyotrophic Lateral Sclerosis (ALS) patients.
Retinal Nerve Fiber Layer (RNFL) and Ganglion Cell Layer (GCL) are measured by Optical Coherence Tomography (OCT) in 61 ALS patients.
The variables included in the model are: gender, age at onset, site of onset (bulbar/spinal), Percentage of Forced vital capacity (FVC), total score ALSFRS-r, and the OCT parameters, that are RNFL Right, RNFL Left, GCL Right mean and minimum, GCL Left mean and minimum.
Only patients without missing data are included, for a total number of 41 subjects.
The time to event outcome is represented by the survival time from the exam date and the censorship indicator (1 if the patients is dead, 0 if the subject is still alive at the follow up).
Two different settings are considered, a train-test setting where the test is randomly sample as the 25% of the dataset, and a leave one out cross validation exercise (CV). The proposed method (GBEST) with different choices of weights are compared with the most common models in survival analysis, Random Forest (RF) and Cox model (Cox). The results are reported in Table 1.

| Performance | GBEST $w = 0$ | GBEST $w = 0.25$ | GBEST $w = 0.5$ | RF | Cox |
|---|---|---|---|---|---|
| IBS | 0.0859 | 0.0653 | 0.0563 | 0.1120 | 0.1621 |
| CV IBS | 0.1264 | 0.1223 | 0.1273 | 0.1293 | 0.1720 |

Table 1: Results in terms of Integrated Brier Score (IBS) on the OCT data.

The results show that our proposed method performs better in terms of integrated Brier score on the data at hands than classical methods. The major differences can be observed with respect to the Cox model, that performs worse in both settings. Random forest shows slightly worse results than GBEST in the cross validation setting, and the difference is more evident in the train-test setting. Regarding the

different weights assigned to the prior distribution, the best results in both settings are obtained for a weight $w = 0.25$ assigned to the prior, that corresponds to a $25\%$ of the observation sampled from the prior and a $75\%$ from the training data for each bootstrap replicate.

## 5. Conclusion

In this paper, a proper Bayesian bootstrap method for ensemble survival tree models is extended to handle categorical variable, in order to apply to real data. The model proposed has been called Generalized Bayesian Ensemble Survival Trees (GBEST).
Empirical results are shown in a simulated study that illustrates the potential of the method in terms of improvement of the performances and stability of results. The application on Amyotrophic Lateral Sclerosis patients shows the advantages of the application of the proposed method in real world problems.
Further developments include the embedding of covariance structure in the sampling process from prior distribution and the application of non parametric models to label the new generated observations.
The R functions developed for this work will be provided in the Github repository: eballante/pBBBaggTrees.

**Acknowledgments** The author thank Dr Luca Diamanti for the data and Mondino Foundation for financial support.

DRAFT

## References

[1] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger, "Bagging survival trees," *Statistics in Medicine*, vol. 23, pp. 77–91, 2004.
[2] H. Ishwaran, U. Kogalur, E. Blackstone, and M. Laure, "Random survival forest," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008.
[3] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
[4] D. Rubin, "The bayesian bootstrap," *Annals of Statistics*, vol. 9, no. 1, pp. 130–134, 1981.
[5] M. Clyde and H. Lee, "Bagging and the bayesian bootstrap," in *Proceedings of the AISTATS*, 2001.
[6] T. Fushiki, "Bayesian bootstrap prediction," *Journal of Statistical Planning and Inference*, vol. 140, pp. 65–74, 2010.
[7] P. Muliere and P. Secchi, "Bayesian nonparametric predictive inference and bootstrap techniques.," *Annals of the Institute of Statistical Mathematics*, vol. 48, pp. 663–673, 1996.
[8] M. Taddy, C. Chen, and M. Wyle, "Bayesian and empirical bayesian forest," in *Proceedings of the International Conference on Machine Learning*, pp. 967–976, 2015.
[9] M. Galvani, C. Bardelli, S. Figini, and P. Muliere, "A bayesian nonparamtric learning approach to ensamblemodels using the proper bayesian bootstrap.," *Alghorithms*, vol. 14, no. 1, p. 11, 2021.
[10] E. Ballante, "An extension of proper bayesian bootstrap ensemble tree models to survival analysis," in *Book of Short Papers of the 51th Scientific Meeting of the Italian Statistical Society* (A. Balzanella, M. Bini, C. C., and R. Verde, eds.), pp. 1766–1770, Pearson, 2022.
[11] M. Jhun and H.-C. Jeong, "Applications of bootstrap methods for categorical data analysis," *Computational Statistics & Data Analysis*, vol. 35, no. 1, pp. 83–91, 2000.
[12] P. Andersen and R. Gill, "Cox's regression model for counting processes, a large sample study," *Annals of Statistics*, vol. 10, pp. 1100–1120, 1982.

# Joint modelling of hospitalizations and survival in Heart Failure patients: a discrete non parametric frailty approach

Masci Chiara[a], Spreafico Marta[b], and Ieva Francesca [a,c]

[a]Politecnico di Milano, Department of Mathematics, MOX; chiara.masci@polimi.it, francesca.ieva@polimi.it

[b]Universiteit Leiden, Mathematisch Instituut; m.spreafico@math.leidenuniv.nl

[c]CHDS - Center for Health Data Science, Human Technopole, Milan, 20157, Italy

DRAFT

**Abstract**

In this work, we propose a novel joint frailty model assuming bivariate discretely-distributed non-parametric frailties, with an unknown finite number of mass points. This approach allows to detect a latent structure among subjects, clustering them in sub-populations where individuals are characterized by a common frailty value. Our method can be interpreted as an unsupervised classification tool and motivates further investigation into the reasons for similarities within the clustered subjects and dissimilarities across the clusters. This work is motivated by a study of patients with Heart Failure (HF) undergoing ACE inhibitors treatment in the Lombardia region of Italy. Recurrent events of interest are hospitalizations due to HF and terminal event is death for any cause.

***Keywords:*** Joint models; Discrete frailties; Recurrent events; EM algorithm; Heart Failure patients.

## 1. Introduction

*Recurrent* or *repeated events* are common in many clinical and biomedical studies, as patients usually experience the same event multiple times. In the recurrent event framework, classic survival approaches are not suitable as they discard the correlation between subsequent events in the same subject. A wide literature about recurrent events modelling has hence flourished in past years [2,8]. Among others, *frailty models* handle repeated episodes by introducing a random effect which takes a common value for each group of dependent observations and can be used to describe excess risk or frailty of different individuals [2,8]. The time span of an individual's recurrent process could also depend on another terminal event, e.g., the end of the study, loss to follow-up, or death. To jointly analyse both recurrent and terminal processes over time, joint frailty models can be adopted [5,6]. Joint frailty models capture both the correlation among repeated events and the dependence between repeated and terminal processes by incorporating random effects in the two hazard functions. Last advances in this field regard the work proposed in [6], where the authors considered joint models by conditioning on a shared frailty that does not apply equivalently for the two hazard functions and the one proposed in [5], where a joint frailty model with multivariate Gaussian random effects is developed. In contrast to traditional joint shared-frailty models, the authors in [5] yield a more general model where two sets of random effects are used to account for intra-subject correlation of multivariate recurrent event times and individual differences in

mortality hazard rate. Their method allows for a positive or negative association between recurrent and terminal events, distinguishing the origin of their dependence.

We insert within this literature by proposing a joint frailty model with discrete random effects. In the last few decades, a branch of the literature is focusing on the treatment of discretely-distributed random effects [3,4]. Most of the work in this direction has been done in the context of mixed-effects regression models for continuous responses (both univariate and multivariate) and other types of responses in the exponential family. Recently, this approach has been extended in the survival analysis framework, by the introduction of discretely-distributed frailties [1]. The main advantage of modelling discrete random-effects regards the new type of treatment and interpretation of the units at the highest level of the hierarchy, that are clustered into latent subpopulations. When considering events nested within patients or patients nested within health providers, modelling discrete frailties allows to cluster patients or health providers, identifying, for example, long/short-term survivors or more/less successful health providers.

Motivated by the advantages of a discretely-distributed frailty approach in practice, in this article we propose an original extension of the general joint model in [5] to the discrete-frailty framework. In addition to handling heterogeneous susceptibility to the risk and informative censoring, this novel approach can detect a latent structure among subjects, grouping them in sub-populations where individuals are characterized by a common frailty value. Along with the model, we propose an estimation routine via Expectation-Maximization algorithm.

The approach we develop is motivated by a study on patients with Heart Failure (HF) in the Lombardia region of Italy. We face the problem of joint modelling of hospitalizations and survival of patients affected by Heart Failure, with a focus on the effect that a pharmacological treatment based on ACE Inhibitors has on these two processes. The course of HF disease is usually characterized by recurrent hospital admissions [7]. Re-hospitalization events usually herald a substantial worsening of patient's survival prognosis and are terminated by death. The application of our novel approach to this context is hence of interest.

DRAFT

## 2. ACE Inhibitors Dataset

We consider data coming from an administrative database of Regione Lombardia, which records clinical courses and pharmacological prescriptions of subjects affected by Heart Failure. We focus on patients who undergo an ACE inhibitors treatment in the period from January 1st, 2006 to December 31st, 2012. For each patient, the index date coincides with the discharge after the first hospitalization due to Heart Failure. We adopt a *gap times* timescale, i.e. each patients clinical history is declined in repeated observations, characterized by a time-to-event variable **GapEvent** which expresses the days elapsed from the previous patient's hospitalization to the next one. The last gap time of each patient expresses the time elapsed from the last known hospitalization to the terminal event, which may be death or censoring. The nature of each event is kept track of through two dummy variables, respectively **Event** and **Death**.

To assess the effect of the considered ACE inhibitors treatment on survival and hospitalizations, we design a time dependent binary classifier for adherent subjects (variable **Adherent**). At each event in a patient's history, we compute the proportion of days covered by prescriptions of ACE inhibitors since the patient index date; then, if this proportion exceeds a threshold of 80% the patient is considered adherent to treatment, otherwise not. The other three variables we consider in the model are the patient gender (**Sex**) and time-dependent variables, **AgeEvent** and **Comorbidity**, which respectively indicate the age and the number of known comorbidities of a patient at the beginning of the corresponding gap time.

Table **??** reports as an example the data table of a patient in the ACE inhibitors dataset.

## 3. Methods

In the proposed joint model with discrete frailty, we express the hospitalization and death hazards for each patient $i, i = 1, ...N$, as follows

| ID | Sex | Adherent | AgeEvent | Comorbidity | GapEvent | Event | Death |
|---|---|---|---|---|---|---|---|
| 10003004 | F | 0 | 75 | 5 | 229 | 1 | |
| 10003004 | F | 1 | 75 | 6 | 131 | 1 | |
| 10003004 | F | 0 | 76 | 6 | 168 | 1 | |
| 10003004 | F | 0 | 77 | 7 | 353 | 1 | |
| 10003004 | F | 1 | 79 | 7 | 1,153 | 0 | 1 |

Table 1: Data table of patient 10003004.

$$\begin{cases} h_i^R(t|u_i, \boldsymbol{x}_i^R(t)) = h_0^R(t) \exp\{u_i + \boldsymbol{\beta}^T \boldsymbol{x}_i^R(t)\} \\ h_i^D(t|v_i, \boldsymbol{x}_i^D(t)) = h_0^D(t) \exp\{v_i + \boldsymbol{\gamma}^T \boldsymbol{x}_i^D(t)\} \end{cases} \tag{1}$$

where $t$ refers to a gap time with respect to the last known hospitalization event; $h_0^R$ and $h_0^D$ are the hospitalization and death baseline hazard functions, respectively; $\boldsymbol{x}_i^R(t)$ and $\boldsymbol{x}_i^D(t)$ are the observed covariates at time $t$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the estimated coefficients of the two models; $u_i$ and $v_i$ are patient-specific additive frailties. According to our formulation, random effects $u_i$ and $v_i$ are distributed according to a bivariate discrete distribution $P^*$ on $\mathbb{R}^2$

$$[u, v]_i \overset{iid}{\sim} P^* \quad \forall i = 1, \ldots, N. \tag{2}$$

Such discrete and with a finite support measure can be characterized by a vector of points in $\mathbb{R}^2$, $\boldsymbol{P}$, and a vector of weights, $\boldsymbol{w}$. Notice that each weight expresses the probability of a patient to be assigned to a certain point $l$, $l = 1, ..., L$ and thus the sum of the weights is constrained to be unitary. Moreover, the number of points constituting the support of the distribution, $L$, is assumed to be unknown a priori. By considering $L$ as fixed, we can write each patient's contribution to the likelihood of the model as a mixture of $L$ components. Each component coincides with the product of the individual contributions, $\mathcal{L}_i$, to the full loglikelihood $\mathcal{L}$ of two independent Cox models with fixed intercept, modelling respectively the recurrent and terminal event process

$$\mathcal{L}(\boldsymbol{\Omega}; data|z_{il}) = \prod_{l=1}^{L} \prod_{i=1}^{N} \left[\mathcal{L}_i(\boldsymbol{\Omega}; data|[u, v]_i = P_l)\right]^{z_{il}}. \tag{3}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{w}, \boldsymbol{P}, h_0^R(t), h_0^D(t)]$. The abscissa and ordinata of each point $\boldsymbol{P}_l$ specify, respectively, the fixed intercept of the recurrent and terminal event Cox models, while $z_{il}$ are a set of binary auxiliary random variables indicating if a patient $i$ is assigned to the point $l$. In order to obtain estimates for $\boldsymbol{\Omega}$, we design a specific EM algorithm, in which at each iteration the model likelihood is firstly averaged with respect to the $z_{il}$ variables and then maximized.
The EM algorithm is then generalized to cope with an a priori unknown number of support points $L$ through its integration into a wrapper support reduction procedure. The first step consists in the definition of a grid of points in $\mathbb{R}^2$, which ideally covers the region in which the support of the discrete distribution is believed to lie. We evaluate two initialization procedures: the former involves sampling a high number of points from a bivariate Normal distribution, whose parameters are set according to previous knowledge (e.g., looking at the estimates of the model proposed in [5]), and initializing their weights according to the corresponding Normal density; the latter consists in defining a uniform distribution over a rectangle in $\mathbb{R}^2$, whose boundaries are set still according to available knowledge. At each iteration, we merge the couple of points with minimum distance in the actual grid whose Euclidean distance is less than a discretional threshold (*MinDist*), until convergence.

## 4. Results

We fit the nonparametric joint model with discrete frailty by consdiering both the gaussian and uniform initializations and we compare the results with the ones of a disjoint model and the joint models pro-

DRAFT

| Variables | Estimate | StdDev | HR | 95% CI | pvalue |
|---|---|---|---|---|---|
| **Recurrent Events** | | | | | |
| Sex [M] | 0.039 | 0.019 | 1.039 | [1.003,1.079] | 0.034 |
| Adherent [1] | -0.259 | 0.019 | 0.771 | [0.743,0.800] | <2e-16 |
| AgeEvent | -0.015 | 0.001 | 0.985 | [0.984,0.987] | <2e-16 |
| Comorbidity | 0.123 | 0.005 | 1.131 | [1.119,1.143] | <2e-16 |
| **Recurrent Events** | | | | | |
| Sex [M] | 0.169 | 0.073 | 1.184 | [1.025,1.366] | 0.021 |
| Adherent [1] | -0.407 | 0.078 | 0.665 | [0.571,0.755] | 1.7e-07 |
| AgeEvent | 0.039 | 0.004 | 1.041 | [1.032,1.049] | <2e-16 |
| Comorbidity | 0.429 | 0.020 | 1.535 | [1.476,1.597] | <2e-16 |

| **Frailty** | **P1** | **P2** | **P3** | **P4** | **P5** | **P6** |
|---|---|---|---|---|---|---|
| $u$ | -0.466 | -0.194 | 0.079 | 0.231 | 0.468 | 0.679 |
| $v$ | -1.872 | -0.859 | -0.090 | 1.166 | 2.277 | 3.088 |
| $w$ | 0.208 | 0.234 | 0.206 | 0.217 | 0.071 | 0.063 |

Table 2: Summary of the Nonparametric Discrete Frailty model with uniform initialization. For categorical variables, the considered stratum is indicated between brackets. Points of the identified frailty discrete distribution are characterized through their abscissa ($u$), ordinata ($v$) and weight ($w$).

DRAFT

posed in [5,6]. For all models, we consider the following set of covariates: **Sex**, **Adherence**, **AgeEvent** and **Comorbidity**. The nonparametric discrete frailty models are fitted using a *MinDist* value of 0.25, which is initially believed to be suitable to spot significantly different fragility classes of patients. The trained models are compared from two perspectives: coefficients' estimation and random effects' characterization. Table **??** reports the estimates for the nonparametric discrete frailty model with uniform initialization[1] while Figure **??** displays the Hazard Ratio estimates and their respective 95% confidence of all compared models.

By looking at the hazard ratios of the covariates, we observe that male subjects are more prone to risk of hospitalization (HR=1.035) and death (HR=1.197). The covariate **Adherent** results statistically significant at any level for the two processes in all trained models. Being adherent yields a 22.9% decrease in the risk of a new hospitalization (HR=0.771) and a 33.5% decrease in the risk of death. From a clinical point of view, such results finally endorse the efficacy of the ACE inhibitors treatment for heart failure. The covariate **AgeEvent** results, for both processes, statistically significant at all levels. Its effect on the hospitalization hazard is a 1.5% reduction of the risk of hospitalization per year (HR=0.985), while on the death hazard it yields an increase of the risk of 4.1% (HR=1.040). Clinically speaking this can be explained by the fact that part of the risk of experiencing a new hospitalization is replaced by the risk to die when patients get older. The covariate **Comorbidity** results to be in all models statistically significant for both the processes. It yields an increase of 13.1% in the risk of hospitalization and a very high increase of 53.5% in the risk of death per comorbidity registered.

From the frailties characterization point of view, the nonparametric frailty model identifies the discrete distribution reported in Table **??** and visualized in the left-side panel of Figure **??**. The estimated discrete distribution consists in six points disposed in a diagonal pattern and show left skewness: point **P1** and **P2**, associated with a probability of 21% and 23%, identify respectively a *Highly Protected* and a *Protected* subpopulation; Point **P3** is related to a subpopulation *Neutral* to random effects, with almost a 20% probability for a patient to belong to it; Point **P4** identifies a relevant group of patients (*At Risk*) slightly more prone to the risk of a new hospitalization and death, with a probability of 20.5%. Point **P5** and **P6** identify two outlier subpopulations of *Fragile* and *Very Fragile* patients, associated with small probabilities (respectively, 7% and 6%). The effect of these estimated frailties can be appreciated by looking at the induced stratified baseline survival curves. As an example, right-side panel of Figure **??** reports the terminal event process induced stratified baselines.

[1]For the sake of brevity, we report results for the solely uniform initialization case.

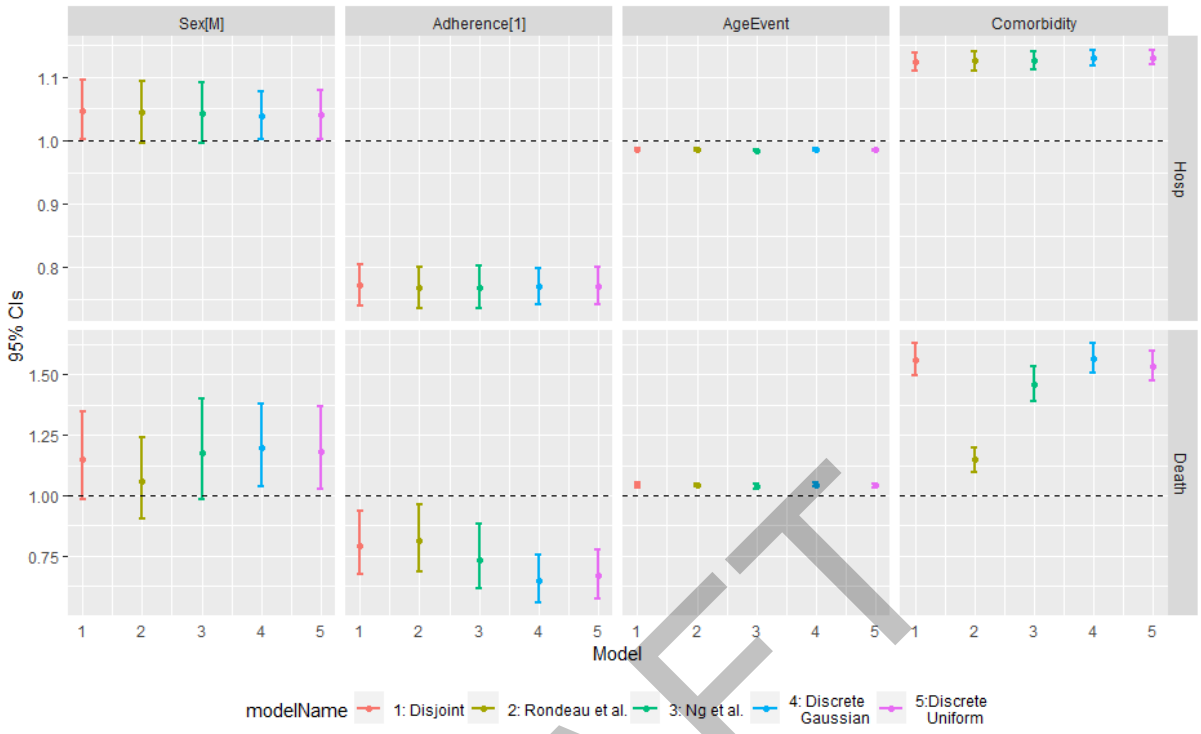



Figure 1: Comparison of estimated hazard ratios and their 95% CI in the trained models. Considered models are: disjoint (pink), Rondeau et al. (pistachio green), Ng et al. (teal), Discrete Nonparametric Frailty with Gaussian Initialization (light blue) and Discrete Nonparametric Frailty with uniform Initialization (purple).

DRAFT

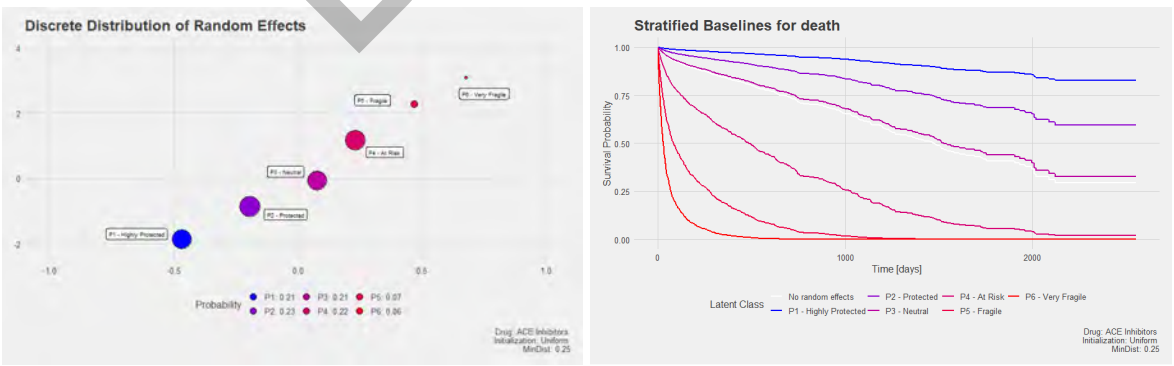



Figure 2: Left-side panel reports the estimated bivariate discrete distribution of random effects, obtained by adopting the uniform initialization case. The color of points ranges from blue (strong subjects) to red (weak subjects), while their size is proportional to the weight. Right-side panel reports the stratified survival probability baseline curves of the terminal event process.

A sensitivity analysis on the *MinDist* threshold is also conducted in order to drive the choice of this important tuning parameter.

## 5. Conclusions

In this work, we propose a methodology to jointly model two dependent longitudinal processes, the first being a recurrent events process, informative for the second, which instead involves events classified as terminal. In our case, the two processes are represented by hospitalizations and departure of patients affected by heart failure who undergo ACE inhibitors therapy. The work actually stems from the idea of expanding previous analyses developed in the field of pharmacoepidemiology, centered on the effect that adherence to drug prescriptions has on the survival outcome of patients, to include in the modelling the recurrent hospitalizations of a patient, in view of the renewed necessity of a proper managing of hospital beds. In this perspective, the proposed joint model with discrete random effects results to be an effective inferential tool. With respect to its counterpart methods in the literature, it yields consistent fixed-effects coefficients estimates, while providing an easy to understand but richer frailty characterization. The identification of classes of patients standing on their fragility level allows for a deeper characterization of patients' profiles. The identified profiles can be further investigated in terms of collateral patients' information.

From a methodological point of view, future work will be dedicated to the definition of a tool for driving the choice of the important tuning parameter *MinDist*.

## References


[1] Gasperoni, F., Ieva, F., Paganoni, A.M., Jackson, C.H. and Sharples, L.: Non- parametric frailty cox models for hierarchical time-to-event data. Biostatistics 21(3), 531-544 (2020).

[2] Kleinbaum, D. G. and Klein, M.: Survival Analysis: A Self-Learning Text. Springer (1996).

[3] Masci, C., Ieva, F. and Paganoni, A. M.: Semiparametric multinomial mixed-effects models: A university students profiling tool. The Annals of Applied Statistics 16(3), 1608-1632 (2022).

[4] Masci, C., Paganoni, A. M. and Ieva, F.: Semiparametric mixed effects models for unsupervised classification of italian schools. Journal of the Royal Statistical Society Series A 182, 1313-1342 (2019).

[5] Ng, S. K., Tawiah, R., McLachlan, G. J. and Gopalan, V.: Joint frailty modeling of time-to-event data to elicit the evolution pathway of events: a generalized linear mixed model approach. Biostatistics (2021).

[6] Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V. and Soubeyran, P.: Joint-frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. International Journal of Epidemiology 8, 708-721 (2007).

[7] Spreafico, M. and Ieva, F.: Functional modeling of recurrent events on time-to-event processes. Biometrical Journal 63(5), 948-967 (2021).

[8] Therneau, T. M. and Grambsch, P.M.: Modeling Survival Data: Extending the Cox Model. New York: Springer (2000).


DRAFT

# Mobility trends in Italy during the first wave of Covid-19 pandemic: analysis on Google data

Ilaria Bombelli[a,b] and Daniele De Rocchi[b]

[a]Italian National Institute of Statistics (ISTAT)
[b]Department of Statistical Sciences, Sapienza University of Rome;
ilaria.bombelli@uniroma1.it, daniele.derocchi@uniroma1.it

**Abstract**

During Covid-19 pandemic, Governments implemented policies to reduce the spread of the virus. In Italy, policies have been implemented starting from 9th March 2020, when in the whole country lock-down policies were adopted. In this study, we analyze mobility data to understand which were the main drivers of mobility during the first pandemic wave. In particular, we analyze Google mobility reports, to study the relative changes in mobility w.r.t. a specific baseline and to analyze several different mobility drivers. In addition, we implement Multilinear Principal Component Analysis to extract relevant features from a multidimensional object. Results show good performances in terms of explained Frobenious norm and two PCs are able to synthesize the trends; finally, the reconstructed trends are also similar to the true original ones.

***Keywords:*** Covid-19, human mobility data, mpca, three-way data

DRAFT

## 1. Introduction

At the beginning of 2020, the SARS-CoV-2 virus and its related disease, known as Covid-19, started to spread all around the world. The first cases were detected in China [1] at the end of 2019; then, on December 31st, Chinese government informed the World Health Organization (WHO) about the detection of some similar cases of respiratory disease due to an undefined agent. On 7th January the virus was identified. Then, on January 31st two Chinese citizens tested positive in Rome and on 17th February the first person affected by the disease in Italy - who has not visited China the months before - was detected. The virus started spreading and the first death occurred on February 20th in Vo' Euganeo (Veneto). In such situation, in the absence of therapies and due to the cases' fast spread, Non-pharmaceutical interventions (NPI) have been implemented by many governments. In particular, in Italy, lock-down policies and mobility restrictions were mainly adopted to reduce the spread of Covid-19. Clearly, the policies had a deep and strong impact on human behavior, especially on human mobility.

The analysis of human mobility data was deeply investigated. For example, [2] compared Covid-19 data and demographic variables with the GPS data in order to study how the restriction orders affect human mobility. In addition, the relationship between Covid-19 transmission and mobility was also analyzed by [3], who focused on 52 countries around the world. Moreover, [4] provided a comprehensive overview of human mobility data and it also compared different data sources to make the researchers and policymakers aware of the nature of the available data sets. A different study has been carried out by [5], who investigated the impact of COVID-19 on the number of people involved in crashes. Finally, [6] studied the effect of the restriction policies on mobility using the geolocalized data from 13 M Facebook users in France, Italy, and the UK.

The research focused on studying mobility data is considered as a data-driven approach which is an extremely helpful tool to support the decision-making process, as the new challenges of digitalization, innovation, and sustainability require.

In this paper, we aim to analyze mobility data provided by Google (more details in Section 2) and to find indicators able to synthesize the mobility data trends related to different categories, as also [7] did for England and Wales.

## 2. Data

Google Covid-19 Community Mobility Reports (GCMR) provide data on human mobility. In particular, mobility reports (GCMR link) share daily mobility data on relative changes compared to a baseline, the day that represents a 'normal' value for that day of the week. In detail, the baseline day is the median value from the 5-week period Jan 3 – Feb 6, 2020. Due to privacy reasons, the absolute values of mobility are not provided. Therefore, for any reported date, the daily relative change is estimated as the percentage change w.r.t. the corresponding baseline weekday. In addition, mobility data refer to 6 different categories which indicate the reasons why the mobility occurred: *grocery, parks, residential, retail, transit, and workplaces*.

In our analysis, the statistical units are the 20 Italian regions; for each of the 6 Google categories, mobility relative changes are provided; finally, the time frame cover from $2020-02-15$ until $2020-08-29$, for a total of 27 weeks.

Therefore, the data structure can be considered as a Three-way data array, having three ways, i.e. rows, columns, and layers, and referring to three modes, i.e. Regions, time, and Google categories, respectively. For each layer (Google category), a rows-by-columns data set provides information on the mobility relative changes during the period taken into account (week 8-week 35) for 20 Italian regions.

To explore the mobility trends, we plot them separately per Google category. In addition, we aggregate the dates referring to 'days' into 'weeks'. In Figure 1, the trends are colored differently, according to the region they refer to. As Figure 1 shows, mobility due to grocery, retail, parks, transit, and workplaces was characterized by a negative change w.r.t. the baseline period, given that lockdown restrictions mainly forbid people to move for other reason than basic needs. In particular, we notice that for transit, retail, and grocery, the relative change w.r.t. the baseline is null starting from week 28; for parks, instead, the mobility trends were set back equal to the baseline starting from week 20 and at week 25 all the regions have the same mobility as the baseline; in addition, from week 25 onward, the mobility towards parks experiences a huge increase in the relative change w.r.t. the baseline, reaching even a positive 500 % of relative change at week 33 in some regions. For what concerns workplaces, the trend, after having experienced a consistent rapid reduction from the starting week until week 15, starts increasing; however, unlike other trends, it remains below 0, meaning that during the analyzed period mobility due to reasons linked to work never came back to the level of the baseline. Totally different is the trend related to residential mobility: in the first weeks, it started a rapid increase in positive relative change w.r.t. the baseline. After that, between weeks 13 and 20, the relative change remain more or less constant at a level of 30 %, then it started slowly decreasing until reaching again level 0 between weeks 25-35, meaning that the mobility during those weeks came back to the baseline level.

## 3. Statistical Analysis

It can be of interest to synthesize the information on mobility data provided by GCMR, by retaining as much information as possible. However, classical techniques, such as the Principal Component Analysis (PCA), cannot be applied, since the three modes (regions, categories, and time) should be taken into account simultaneously.

As [7] noticed, for such data a multilinear PCA (MPCA) is suggested. MPCA has its origin in the well-known Tucker decomposition of a K-Tensor [8], which is used to reduce the dimension of a tensor object, or three-way data structure; the main contributor to the method is [9]. While PCA is used

DRAFT

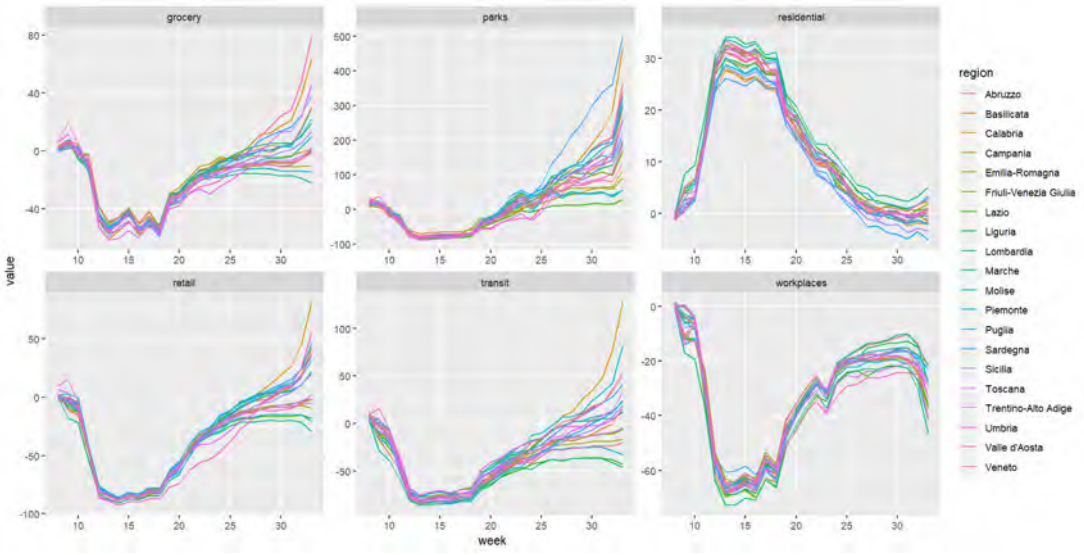



**Figure 1:** Mobility relative changes w.r.t. baseline separated by Italian region and reason (Google Covid-19 Community Mobility Reports (GCMR))

to reduce the dimension of the variables in a units-by-variables data set, MPCA is used to extract the relevant features of a multiway object.

It is worth mentioning that we exclude from the analysis data from parks. Indeed, on the one hand, the analysis performance, measured by the explained Frobenius norm, is significantly reduced if we include this category (less than $50\%$); on the other hand, the trend of mobility towards parks is very different in terms of behavior and in terms of magnitude, as aforementioned.

By applying the `mpca` function (`Rtensor` package) on the tensor (or three-way data) object (built by using `as.tensor` function in the `Rtensor` package), we obtain a good synthesis by using 2 PCs, as the percent of Frobenius norm explained by the approximation is nearly equal to $79\%$. In addition, the estimated data, namely the estimates of all the tensor values after compression, are quite close to the true ones. Indeed, Figure 2 shows that points (true observations) are almost gathered by the central solid lines (estimated trend), even if during the last weeks of the considered period, points spread a bit far from the estimations, meaning that the model is only nearly able to capture the variability of the trends in those weeks.

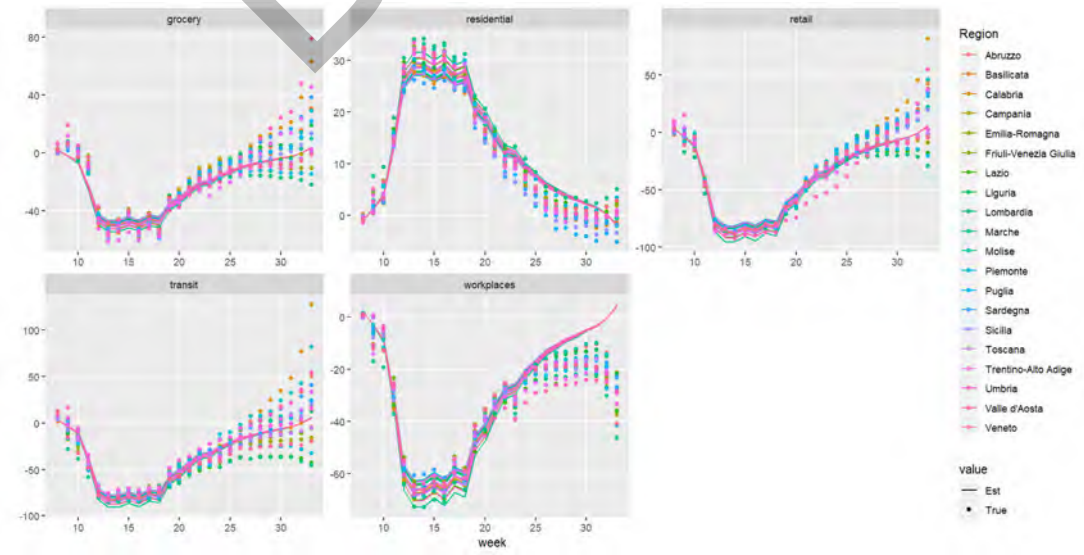



**Figure 2:** Estimated by mpca (2 components) and true observed mobility trends separated by Italian region and reason

DRAFT

By looking at the first two PCs in Figure 3, we realize that the former completely describes mobility for a residential reason, while the latter is a good synthesis of the mobility patterns for the other reasons.

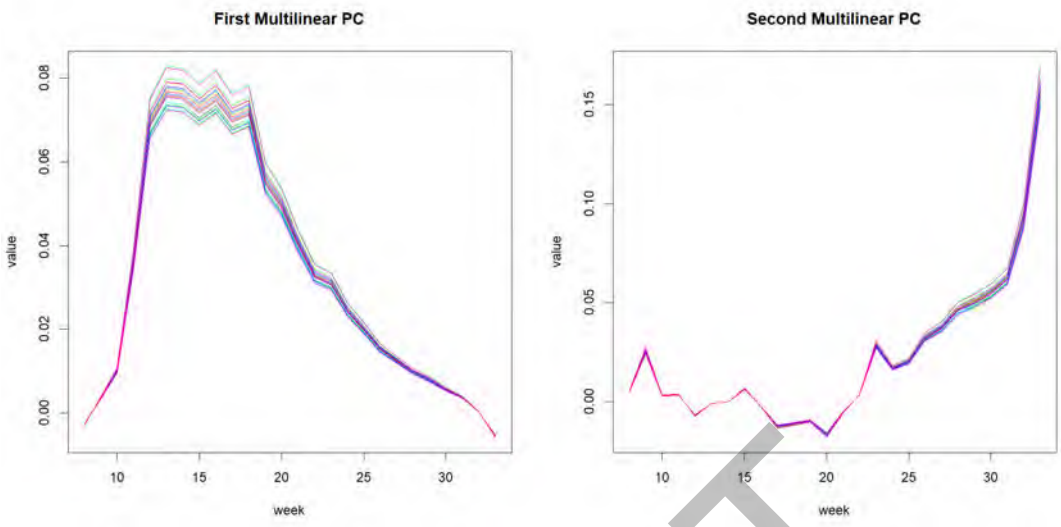



**Figure 3:** Mobility patterns of two PC resulting from mpca

To further describe the mobility trends, we apply the multilinear PCA with three components. By analyzing the results, we obviously obtain an increase in explained variance percentage, moving from 79% to approximately 80%. However, we realize that this gain in terms of explained variance is too small to balance the increase in model complexity. Anyway, for the sake of completeness, some details on the obtained 3 PCs are provided below. From Figure 4, we notice that the third PC allowed us to better capture the variability of the trends in the last weeks of the considered time frame, especially the trends of mobility toward workplaces.

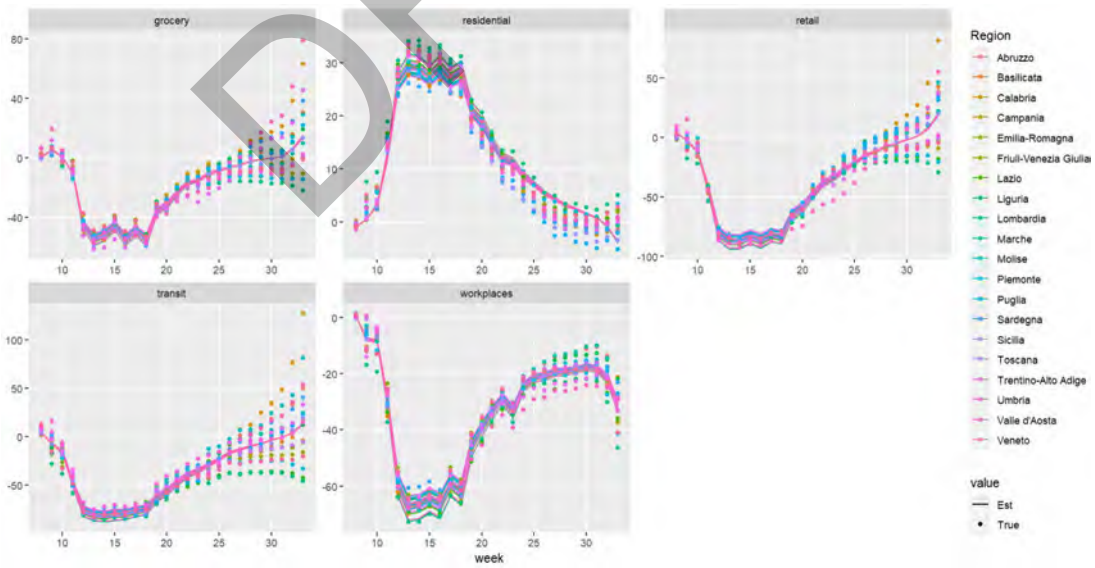



**Figure 4:** Estimated by mpca (3 components) and true observed mobility trends separated by Italian region and reason

In addition, by looking at the first three PC in Figure 5, we realize that the former completely describes the mobility for a residential reason, the third almost describes the mobility towards workspaces, while the second one is a good synthesis of the mobility patterns for the other reasons.

DRAFT

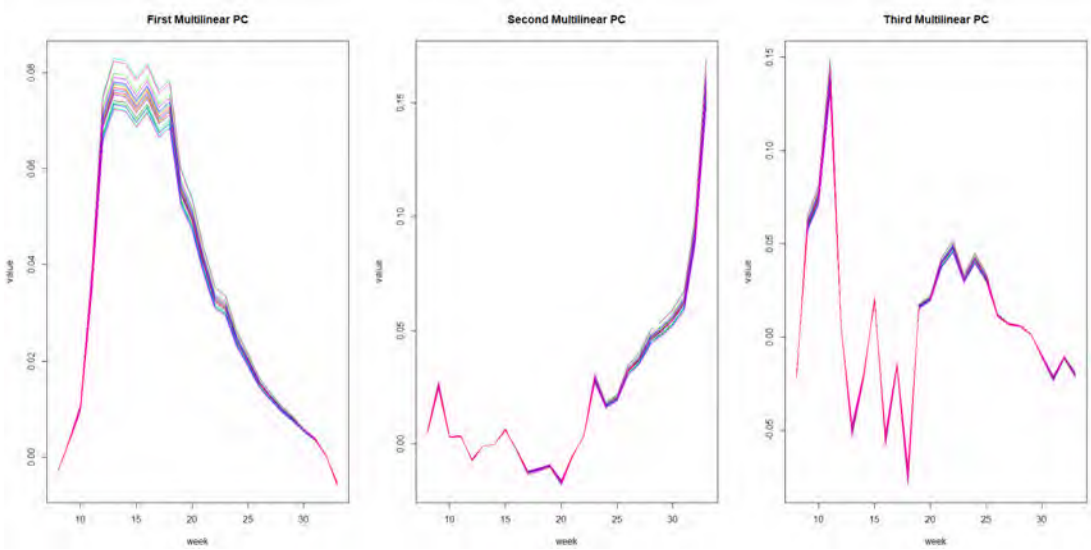



**Figure 5:** Mobility patterns of three PC resulting from mpca

## 4. Conclusion

The analysis allowed us to study how mobility during the first way of the pandemic in Italy changed with respect to the baseline period. Trends show that there was a negative relative change for mobility due to reasons other than residential, while the change was positive and rapid for residential mobility. By reducing the dimension of the three-way data object by using the Multilinear PCA, it was possible to obtain two PCs which are a good synthesis of the 5 trends (one for each mobility driver), as the percentage of Frobenious norm explained shows. In addition, the MPCA performed well also in terms of data reconstruction. Indeed, by estimating the trends and comparing them with the true ones, we experience a good model performance.

Further developments consider an analysis that involves mobility data and mortality data to study the relationship between the two.

DRAFT

## References

[1] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England journal of medicine*, 2020.

[2] Samuel Engle, John Stromme, and Anson Zhou. Staying at home: mobility effects of covid-19. *Available at SSRN 3565703*, 2020.

[3] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, Kylie EC Ainslie, Marc Baguelin, Samir Bhatt, Adhiratha Boonyasiri, Nicholas F Brazeau, Lorenzo Cattarino, Laura V Cooper, et al. Reduction in mobility and covid-19 transmission. *Nature communications*, 12(1):1090, 2021.

[4] Tao Hu, Siqin Wang, Bing She, Mengxi Zhang, Xiao Huang, Yunhe Cui, Jacob Khuri, Yaxin Hu, Xiaokang Fu, Xiaoyue Wang, et al. Human mobility data in the covid-19 pandemic: characteristics, applications, and challenges. *International Journal of Digital Earth*, 14(9):1126–1147, 2021.

[5] Jie Zhang, Baoheng Feng, Yina Wu, Pengpeng Xu, Ruimin Ke, and Ni Dong. The effect of human mobility and control measures on traffic safety during covid-19 pandemic. *PLoS one*, 16(3):e0243263, 2021.

[6] Alessandro Galeazzi, Matteo Cinelli, Giovanni Bonaccorsi, Francesco Pierri, Ana Lucia Schmidt, Antonio Scala, Fabio Pammolli, and Walter Quattrociocchi. Human mobility in response to covid-19 in france, italy and uk. *Scientific reports*, 11(1):13141, 2021.

[7] Ugofilippo Basellini, Diego Alburez-Gutierrez, Emanuele Del Fava, Daniela Perrotta, Marco

Bonetti, Carlo G Camarda, and Emilio Zagheni. Linking excess mortality to mobility data during the first wave of covid-19 in england and wales. *SSM-Population Health*, 14:100799, 2021.

[8] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[9] Haiping Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Mpca: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks*, 19(1):18–39, 2008.

DRAFT

# Tracking attitudes towards COVID vaccines: A text mining analysis

Leonardo Scarso[a], Marco Novelli[b], and Francesco Saverio Violante[a]

[a]Department of Medical and Surgical Sciences, University of Bologna, Via Pelagio Palagi 9, Bologna;
leonardo.scarso2@unibo.it, francesco.violante@unibo.it
[b]Department of Statistical Sciences, University of Bologna, Via delle Belle Arti 41, Bologna;
m.novelli@unibo.it

**Abstract**

It is known how the social media changed our lives and the impact they have on our society today: from the general health state to thoughts on a certain topic, online users share every opinion on online platforms, especially Twitter, which provides real-time information. This article focuses the attention on the analysis of tweets in Italian language relating to an advanced stage of COVID vaccinations, from April to December 2022, with the purpose of extracting the main topics that are depicted in the considered timestamp, through the application of a new indicator, together with sentiment analysis. The results highlight the ability of the proposed analysis strategy in capturing peaks corresponding to particular events and how they were perceived on social media debates.

***Keywords:*** SARS-CoV-2, sentiment analysis, Twitter, vaccination

DRAFT

## 1. Introduction

The outbreak of social media is certainly one of the main events that characterized the first decades of the $21^{st}$ century, with the aim of bringing people together, even from different parts of the globe. It is well known that online platforms have pros and cons: from the diffusion of a large amount of misinformation, to the spread of hate and other negative feelings, worldwide population is free to share whatever they want; only an internet connection is required.

From the general health state to thoughts about a particular topic, everything is shared on online discussions, especially on Twitter, that represents the perfect example of social media: it is featured by its immediacy of the data, allowing for precise information on public opinion.

A strongly debated topic during the last three years is obviously related to the outbreak of the *SARS-CoV-2* virus: the global spread of COVID-19 significantly changed customs and behaviors of the inhabitants, due to the uniqueness of the event. Indeed, this singular occurrence led political and medical institutions to find quick and effective solutions in order to prevent the diseases caused by the infection: principal resolutions consisted on lockdown measures and vaccination campaigns.

Textual analysis could be an effective monitoring tool for the epidemiological situation: it should be applied via sentiment analysis on tweets content (4) or for computing a relative increase indicator based on the amount of data at disposal (1). A further trail of predicting the evolution of the pandemic by a tool capable of providing real-time socio-economic and health crisis is proposed in (5).

In this article we focus on the analysis of Twitter posts about anti-COVID vaccines written in Italian language, from April $15^{th}$, 2022 to December $31^{st}$, 2022, with the purpose of understanding popular

attitudes towards the injection of Pfizer/BioNTech, Moderna, AstraZeneca/Vaxzevria and Johnson & Johnson vaccines in an advanced stage of the campaigns, when the number of infections is notably reduced. Specifically, the attention will be centered on an investigation consisting on the construction of indicators related to textual data combined with the implementation of sentiment analysis.

## 2. Data analysis

In this section we will discuss on the analysis that has been performed, starting from a brief illustration of the structure of the data at our disposal.

### 2.1 Data set description

The built data set contains 593351 tweets written in Italian language, collected day by day from mid April to the end of December 2022, each of them containing one of the following terms or bi-grams: *vaccino COVID*, *vaccinazioni COVID*, the proper vaccines names like *Pfizer*, *Moderna*, *AstraZeneca* and *Johnson & Johnson*, as also some specifics (*BioNTech*, *Vaxzevria*). Each observation consists on several variables; for instance, the type of message that users posted: original tweet (8.8% of the total), retweet (76.7%), reply or quote (14.5%). Only 1949 of them contain information about the geo-location from which the message has been posted; the most frequent ones were shared from Northern Italy (44.33%), followed by the central part (26.83%) and the Southern one (15.44%). The remaining tweets are spread across the different continents, from Australia to North America.
Nonetheless, in this work the attention will be focused only on the text. Before working on it, text has been pre-processed: indeed, punctuation marks and special ones have been dropped, as stop words, in order to obtain a more suitable information for our analysis.

A strategy consisting on the extraction of data information about mentioning vaccines is performed, through the application of two indexes, both based on the daily amount of gathered tweets. The interpretation will also be corroborated with the implementation of sentiment analysis, with the scope of describing the most important events occurred in Italy concerning the anti-COVID vaccination campaigns, and how they have been perceived by Italian citizens.

### 2.2 Indexes comparison

In this work we proposed a new indicator, called *Relative Count of Vaccine* $i$ on day $t$, $RCV_{i,t}$, which is fashioned with the aim of indicating and highlighting how the magnitude of some specific terms can evolve over time. The $RCV_{i,t}$ is the ratio between the absolute count of the vaccine $i$ on tweets of the day $t$, $ACV_{i,t}$ (hence, how many times a vaccine has been mentioned on tweets of a given day), and the number of collected posts on day $t$, $N_t$.

$$RCV_{i,t} = \frac{ACV_{i,t}}{N_t}.$$

In Figure 1 the evolution over time of the four indexes, one for each vaccine, is provided. Except for some days, the most mentioned vaccine is the Pfizer one, as the blue line shows in the plot, followed by, in magnitude order, Moderna, AstraZeneca and Johnson & Johnson, which also characterized the social media discussions in Italian language.

In addition, Table 1 reports some examples of peaks, on which date, vaccine name, respective value and event of interest are attached. For instance, the highest value reached by the $RCV$ is 0.82 on April, $28^{th}$, 2022; it should be interpreted as follows: *the 82% of tweets written on April,* $28^{th}$ *contained at least once the terms 'Pfizer' or 'BioNTech'*. The peak was reached after the report of the president of the Italian Drug Agency (AIFA) about the injection of the fourth dose, that appeared unnecessary; this event caused a large discussion about vaccines, due to the fact that Italian population has always been split up into two parts: from people believing on the anti-COVID vaccines effectiveness, against ones

DRAFT

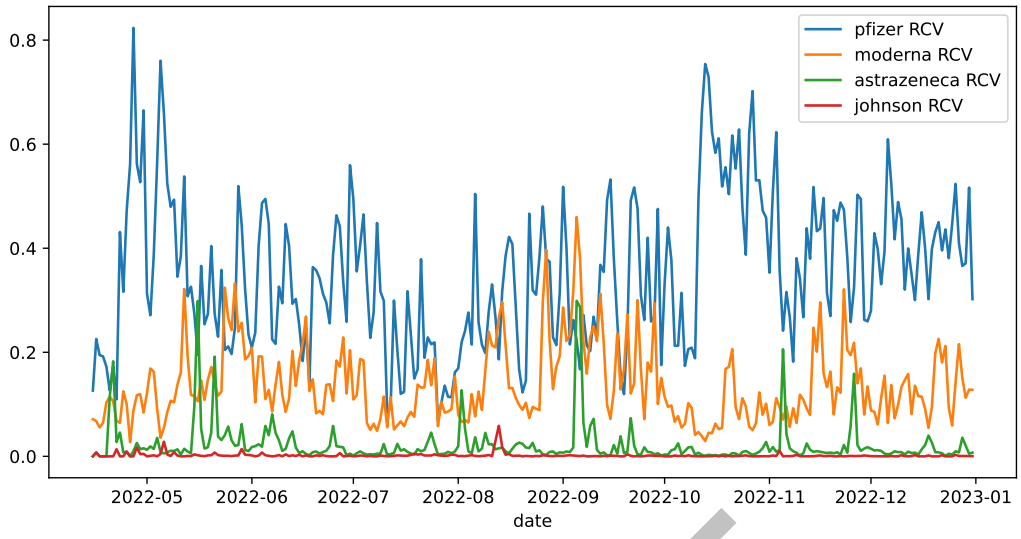



Figure 1: $RCV$ values of Pfizer, Moderna, AstraZeneca and Johnson & Johnson vaccines.

| **Date** | **Vaccine** | ***RCV* value** | **Main debate** |
|---|---|---|---|
| April $28^{th}$ | Pfizer | 0.82 | Injection of fourth dose not necessary |
| May $6^{th}$ | J&J | 0.03 | Injection suspended on United States |
| May $16^{th}$ | AstraZeneca | 0.29 | Death of a woman after the second dose |
| May $25^{th}$–$27^{th}$ | Moderna | 0.33 | Red points on arms as adverse reaction |

Table 1: Description of a few main events related to $RCV$s peaks.

DRAFT

not trusting on how quickly the researchers have found a possible cure for the disease.

The literature on this specific aspect is rather scarce. Gori *et al.* (1) proposed a relative increase indicator, called for simplicity $RII_t$, to describe how the number of daily tweets could be affected by the happening of an event. The $RII_t$ is equal to the the difference among the volume of tweets collected on day $t$, $N_t$, and the one of the day before, $N_{t-1}$, divided by the amount of tweets collected on the previous week, $N_t^{t-6}$, namely from day $t$ to day $t-6$.

$$RII_t = \frac{N_t - N_{t-1}}{N_t^{t-6}}.$$

In Figure 2 a comparison among the $RII$ (in blue) and the sum of the four $RCV$s (in orange) indexes is supplied. Since the two quantities are based on different concepts, there is no similarity between them; as a matter of fact, the trends are different along the timestamp. Indeed, differently from the $RCV$, that has a wider variation range, the $RII$ behaves erratically, spreading around -0.11 and 0.13, suggesting that the relative amount of discussion about vaccines seems to be constant in this forward step of vaccination campaigns, so it is not so informative. Moreover, the indicator proposed in (1) exclusively depends on the amount of collected data: if the daily extraction is not constant, the result could be not informative, since the peaks are only given by a varying total of tweets collected in different days.
On the other hand, the $RCV$s sum, which depends on the count of mentioned terms, expresses how posts regarding vaccinations could be characterized by different peaks about more or less debates on the injection of vaccines, that mainly correspond to the event happened at the same day, or some days before, even when the number of daily tweets is constant (about 2500 per day, in the case in analysis).

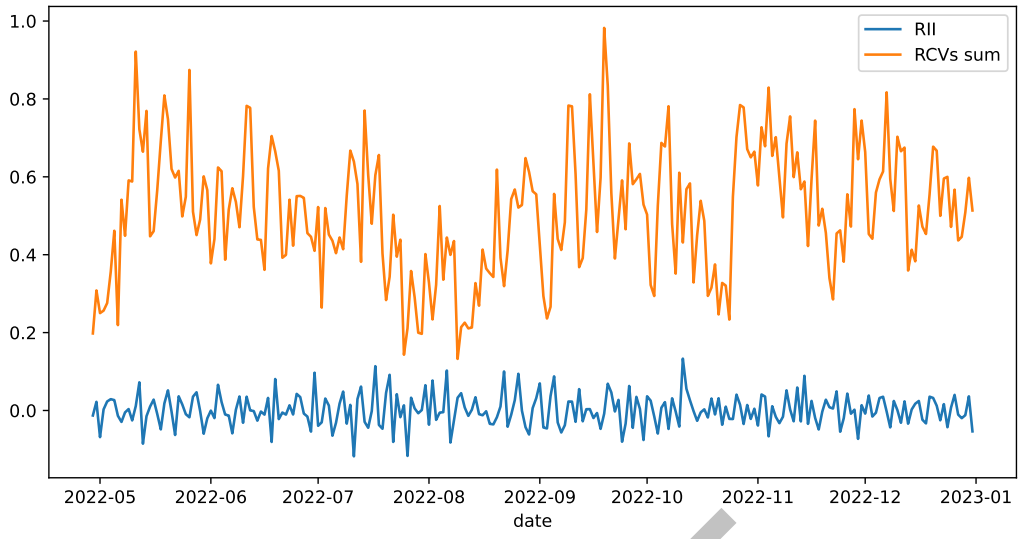



Figure 2: Comparison between the $RCV$s sum and $RII$ values.

DRAFT

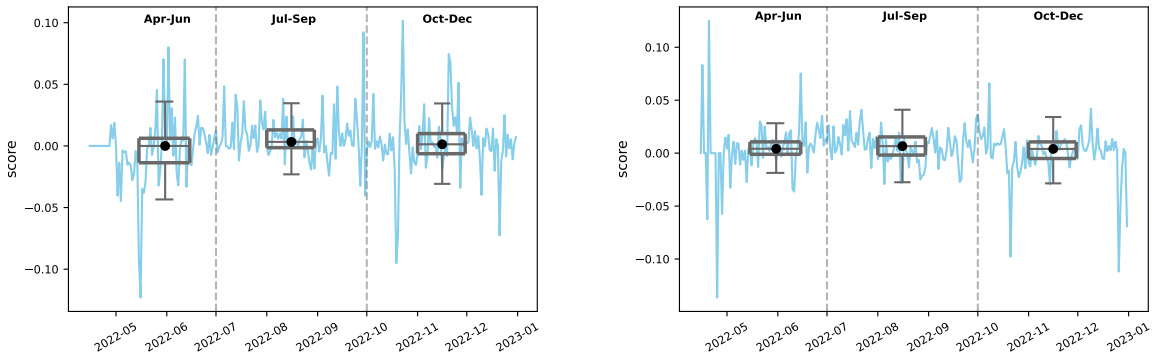
Indeed, the $RCV$ turns out to be an interesting indicator when particular tokens wanted to be found into documents or simple texts. The strength of a simple but effective index like the one suggested in this article is highlighted in the comparison made in Figure 2, which emphasizes the way the indicator seems to adequately describe the temporal evolution of the debates about a given argument, underlining ups and downs caused by particular events that affect our society.

## 2.3 Sentiment analysis

Until now we have seen how an indicator based on the count of terms can be constructed and analyzed to notice whether some debates are present in a given time window and how their magnitude can change. However, social media posts report emotions and feelings, given by the attitude that the users have afterwards an event that could provoke different reactions inside them: from anger and pain, to happiness and joy. Sentiment analysis is a powerful Natural Language Processing technique that can be applied in cases when the subjectivity of a document would be extracted. Based on the implementation of an existing trained pipeline (*spaCy*), the original tweets text has been retrieved, since some emoticons that have been written as ensemble of different punctuation marks (for example the smile '*:)*', which is the most common one) could cause a meaningful variation in the message semantics. The pipeline computes a subjectivity score across -1.0 and 1.0 for each tweet, corresponding to the most positive and the most negative assessments, hence the averaged daily scores are computed, to establish the way the trend evolves.

The following analysis could be thought of as an extension of the one made in (4), where the authors considered the period from December $1^{st}$, 2020 to March $31^{st}$, 2021. Here we split up the timestamp into three sub-periods (April-June, July-September, October-December, 2022). Figure 3 summarizes sentiment analysis results for Pfizer and Moderna vaccines (respectively, panel (a) and (b)), along with a box plot describing the distribution of the scores for each considered sub-period. Note that, due to the limited numbers of daily tweets for AstraZeneca and Johnson & Johnson, those vaccines have been removed from the analysis.

The figure shows that the sentiment for both vaccines varies between -0.15 and 0.15 and the median value for each sub-period lies around 0. However, by looking at the box plot, it can be seen that Pfizer




(a) Sentiment score of Pfizer vaccine.

(b) Sentiment score of Moderna vaccine.

Figure 3: Averaged sentiment score across the considered timestamp for Pfizer and Moderna vaccines.

DRAFT

sentiment scores show a wider range across all the three-months periods, if compared to the Moderna one. This is true also for the interquartile range (IQR): for instance, the first sub-period (April-June) of Moderna has a IQR value of 0.012, whereas, for Pfizer vaccine, the respective value is 0.02.

A non-parametric Kruskal-Wallis test has been implemented in order to assess whether variations in the sentiment scores are observed across the considered time windows. Our results corroborate those obtained in (4), indeed we found a relatively stable sentiment about the Moderna vaccine in the considered period (p-value = 0.176). For Pfizer, instead, there seems to be a difference among the median level of the scores in the three periods (p-value = 0.013): more specifically, a Dunn's test (2) reveals a significant increase between the first and the second period (p-value = 0.010 with the Simes-Hochberg correction).

Despite the fact that the average sentiment seems to be somewhat neutral, the scores exhibit positive and negative peaks, that can be related to those captured by the $RCV$ proposed in the previous paragraph. As regarding Pfizer, the maximum value reached in Figure 3 is of 0.101 on October, $23^{rd}$; on the same day, the value of the related $RCV$ (in Figure 1) is 0.628, corresponding to the period on which the discussion of Pfizer became more intense if compared to the beginning of the month. Doubts about the adverse reactions that the flu vaccine and three or more doses of Pfizer could be encountered are marked in the debates of the third October decade; however, even though the related sentiment is slightly positive, it is preceded by a strong negative peak some days before (-0.073).

Note that few peaks captured by the $RCV$ index are not present when the sentiment analysis is applied. In most cases, the absence of a peak in the sentiment score could be due to the the fact that a large number of posts are retweets of a news posted by institutional profiles (e.g., governmental institutions, newspapers, etc.) (3). Indeed, their message content should not have subjective terms, hence the computed sentiment score should lie around 0, determining the neutral tone the tweets have; here a quantitative approach could help in understanding the magnitude of the discussion. Indeed, an effective 'quali-quantitative' analysis strategy could combine the application of the sentiment analysis along with the $RCV$ index, since the latter has the additional ability of catching particular events.

## 3. Conclusion

In this article we propose an example of how social media posts could be exploited to perform several analysis related to population thoughts and feelings based on a certain event linked to the COVID-19 vaccination campaigns. A new indicator, the Relative Count of Vaccine, has been proposed, which is capable of capturing peaks of debates concerning a particular topic. A sentiment analysis is also

performed, which complements the results obtained by the $RCV$, by evidencing possible positive or negative emotions and hence the direction of the sentiment.
As a possible future extension of the analysis, a larger number of languages could be considered, with the aim of demonstrating if worldwide events related to SARS-CoV-2 vaccination campaigns could be recorded and in which way they were perceived and expressed on social media debates.

## References


[1] Gori, D., Reno, C., Remondini, D., Durazzi, F., Fantini, M.P.: Are we ready for the arrival of the new COVID-19 vaccinations? Great promises and unknown challenges still to come. *Vaccines*, Vol. 9, No. 2 (2021)

[2] Hollander, M., Wolfe, D. A. Wolfe, Chicken, E.: Nonparametric Statistical Methods, Third Edition. *Wiley Series in Probability and Statistics* (2015)

[3] Ji, X., Chun, S. A., Geller, J., Knowledge-Based Tweet Classification for Disease Sentiment Monitoring. *Sentiment Analysis and Ontology Engineering*, pp. 425-454, Springer International Publishing (2016)

[4] Marcec, R., Likic, R.: Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, Vol. 98, No. 1161, pp.544-550 (2021)

[5] Sepúlveda, A., Periñán-Pascual, C., Muñoz, A., Martínez-España, R., Hernández-Orallo, E., Cecilia, J.M.: COVIDSensing: Social Sensing Strategy for the Management of the COVID-19 Crisis. *Electronics*, Vol. 10, No. 3157 (2021)


DRAFT

# Treatment effect assessment in observational studies with multi-level treatment and outcome

Federica Cugnata[a], Paola Vicard[b], Paola M.V. Rancoita[a], Fulvia Mecatti[c], Clelia Di Serio[a], and Pier Luigi Conti[d]

[a]Vita-Salute San Raffaele University; cugnata.federica@unisr.it, rancoita.paolamaria@unisr.it, diserio.clelia@unisr.it
[b]Roma Tre University; paola.vicard@uniroma3.it
[c]University of Milano-Bicocca; fulvia.mecatti@unimib.it
[d]Sapienza University of Rome; pierluigi.conti@uniroma1.it

DRAFT

**Abstract**

In observational studies, one of the main difficulties consists in the comparison of treatment effects. In fact, receiving a treatment is not a "purely random" event, and there could be relevant differences between treatment groups. Propensity score is a popular tool to account for this source of bias. However, its use requires a careful modelization of the dependence relationships of the treatment on the covariates. In this work, we consider a general setting with multiple treatments and discrete multi-valued outcome. We propose to estimate the propensity score by using Bayesian Networks and, based on this, we develop an inferential methodology to evaluate the treatments effect. The performance of the proposed approach have been studied through a simulation study with very promising results.

***Keywords:*** Potential outcomes, propensity score, covariate balance, observational study, Bayesian networks

## 1. Introduction

In medical research, there is a growing interest in evaluating the treatments effect using data from observational studies. These studies have several practical advantages compared to experimental study designs, but the lack of "purely random" treatment allocation may lead to patient characteristics unbalance between the treatment groups and thus to biased assessment of treatment effects. In this setting, several methods based on propensity score are widely adopted to account for this source of bias. In particular, statistical approaches have been developed to match, stratify or weight the samples of the treatment groups. In spite of this, much of the work on propensity score analysis has focused on the case where the treatment is binary, while the use of propensity score when examining multiple treatment conditions has received limited attention [4]. Typically, in the applications with multiple treatments the propensity score is estimated using multinomial or ordinal logistic regression model. The challenge for this approach is choosing the functional form and the correct set of interactions among the covariates to capture their relationship to treatment assignment. A misspecified propensity score may fail to achieve covariate balance between treatment groups. To overcome this issue, in the binary case of two treatments, in [5] we proposed to estimate the propensity score by using Bayesian Networks and based on it, we developed a methodology to evaluate the treatment effect. Estimating propensity score via BNs was

shown to offer substantial theoretical and practical advantages, with respect to more popular approaches such as logistic modelling and machine learning methods. Following this approach, in the present paper we propose to extend the use of the Bayesian Netwok to estimate the propensity score in the multi-level treatment case. Moreover, we consider the case where the outcomes are multi-level. In the literature, when outcomes are continuous or binary, the most commonly way to evaluate the treatment effect is the Average Treatment Effect (ATE). However, this quantity can give limited information expecially for disomogenous responses within groups; furthemore it is not well defined for categorical data with more than two categories. Then, to evaluate the treatment effect in this context we propose an appropriate test for the absence of treatment effect by using a methodology based on inverse probability weighting. A simulation study is set up for evaluating the performance of the proposed approach.

## 2. Method

In formal terms, consider $n$ independent subjects, each receiving a treatment $T$ with $K+1$ levels 0, $1, \ldots, K$. The level 0 denotes the absence of treatment (control group). Let $Y_{(k)}$ the *potential outcome* of a subject when treatment is $k = 0, 1, \ldots, K$. The *observed outcome* for a subject is then

$$Y = \sum_{k=0}^{K} Y_{(k)} I_{(T=k)} \tag{1}$$

where $I_{(T=k)} = 1$ if $T = k$, $I_{(T=k)} = 0$ otherwise. The treatment has no effect if $Y_{(0)} \stackrel{d}{=} Y_{(1)} \stackrel{d}{=} Y_{(K)}$, where $\stackrel{d}{=}$ denotes equality in distribution. Due to the presence of confounding covariates, the assignment-to-treatment mechanism is not purely at random. As an effect, there could be relevant differences among subjects receiving different treatment levels. In the present paper the assignment-to-treatment mechanism is assumed to depend only on *observed* covariates, with no unobserved confouders. Let $\boldsymbol{X} = (X_1 \cdots X_L)$ be the vector of relevant covariates, and denote by

$$p_k(\boldsymbol{x}) = P(T = k | \boldsymbol{X} = \boldsymbol{x}); \;\; k = 0, 1, \ldots, K \tag{2}$$

the propensity score, namely the probability of receiving treatment at level $k$ conditionally on covariates. Due to the nature of the considered application, attention is focused on the case of covariates and potential outcomes that are discrete, finite random variables (r.v.). The assumptions of the analysis are listed below.

H1. *Discreteness*. Each covariate $X_l$ is discrete, finite, and may take nominal values $x_l^1, \ldots, x_l^{s_l}$ with positive probability. The potential outcomes $Y_{(k)}$ are finite, discrete, r.v.s, too, again taking values $0, 1, \ldots, K$ with marginal probability:

$$\theta_k(h) = P(Y_{(k)} = h); \;\; h = 0, 1, \ldots, H, \; k = 0, 1, \ldots, K. \tag{3}$$

H2. *Unconfoundedness*. $T \perp\!\!\!\perp (Y_{(0)}, Y_{(1)}, \ldots, Y_{(K)}) | \boldsymbol{X}$.

H3. *Common support*. There exists a positive real $\delta$ for which $\delta \leq p_k(\boldsymbol{x}) \leq 1 - \delta$ for each $\boldsymbol{x}$ and $k = 0, 1, \ldots, K$.

For the sake of simplicity, the following vector notation is used.

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_0 \\ \boldsymbol{\theta}_1 \\ \cdots \\ \boldsymbol{\theta}_K \end{bmatrix}, \;\; \text{where } \boldsymbol{\theta}_k = \begin{bmatrix} \theta_k(0) \\ \theta_k(1) \\ \cdots \\ \theta_k(H) \end{bmatrix}, \; k = 0, 1, \ldots, K.$$

To estimate the probabilities $\theta_k(h)$, a methodology based on inverse probability weighting is used. Propensity scores $p_k(\boldsymbol{x})$ are first estimated through a Bayesian Network (BN; [1]). As shown in [5], BNs

DRAFT

estimators have several fundamental properties, because they are (universally) consistent and asymptotically efficient, even when the structure of the network is learnt from data. On the basis of the BN estimator $\widehat{p}_k(\boldsymbol{x})$, the estimator

$$\widehat{\theta}_k(h) = \sum_{i=1}^{n} I_{(Y_i=h)} I_{(T_i=k)} p_k(\boldsymbol{x}_i)^{-1} \Big/ \sum_{j=1}^{n} p_k(\boldsymbol{x}_j)^{-1};\ h = 0, 1, \ldots, H,\ k = 0, 1, \ldots, K \tag{4}$$

is constructed. To establish asymptotic, large sample properties of (4) we need a vector notation, similar to that introduced above. Define:

$$\widehat{\boldsymbol{\theta}}_k = \begin{bmatrix} \widehat{\theta}_k(0) \\ \widehat{\theta}_k(1) \\ \cdots \\ \widehat{\theta}_k(H) \end{bmatrix},\ k = 0, 1, \ldots, K;\ \widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\boldsymbol{\theta}}_0 \\ \widehat{\boldsymbol{\theta}}_1 \\ \cdots \\ \widehat{\boldsymbol{\theta}}_K \end{bmatrix}$$

and consider the $(K+1)(H+1)$-variate r.v.

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{5}$$

Under assumptions H1-H3,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{W} \ \text{ as } n \to \infty \tag{6}$$

where $\boldsymbol{W}$ possesses a (singular) Normal multivariate distribution, with null expectation and covariance matrix $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{\Sigma}$ have a complex form, depending on unknown quantities. However, it can be consistently estimated on the basis of sample data. In other terms, it is possible to construct an estimator $\widehat{\boldsymbol{\Sigma}}$ that converges in probability to $\boldsymbol{\Sigma}$ as the sample size $n$ tends to infinity. This result will be mainly used in the sequel to construct test-statistics for the evaluation of possible treatment effects.

## 2.1 Testing for the absence of treatment effect

The main hypothesis to be tested is the *absence of treatment effect*, namely

$$\begin{cases} H_0: & \boldsymbol{\theta}_0 = \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K \\ H_1: & \text{The vectors } \boldsymbol{\theta}_k \text{ do not coincide} \end{cases} \tag{7}$$

If we define the $(H+1)$-dimensional vector $\overline{\boldsymbol{\theta}} = \frac{1}{K+1}\sum_{k=0}^{K} \boldsymbol{\theta}_k$, testing for the absence of treatment effect reduces to the following hypothesis problem

$$\begin{cases} H_0: & \boldsymbol{\theta}_k = \overline{\boldsymbol{\theta}}\ \ \forall k = 0, 1, \ldots, K \\ H_1: & \boldsymbol{\theta}_k \neq \overline{\boldsymbol{\theta}} \text{ for some } k = 0, 1, \ldots, K \end{cases} \tag{8}$$

The basic idea consists in using the test-statistic

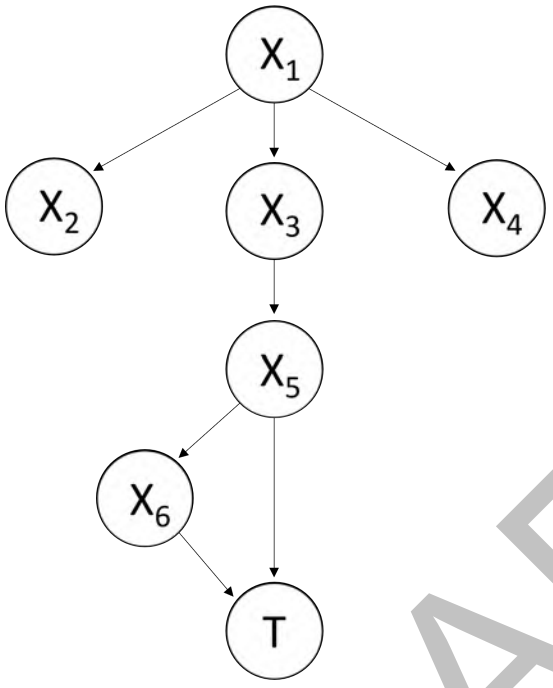
$$D_n^2 = n \sum_{k=0}^{K} \sum_{h=0}^{H} \frac{\left(\widehat{\theta}_k(h) - \widehat{\overline{\theta}}(h)\right)^2}{\widehat{\overline{\theta}}(h)}, \tag{9}$$

which is similar, in principle, to the test-statistic used in a different context by [3].
The test-statistic $D_n^2$ is essentially a sum of measures of divergence of the (probability) vectors $\widehat{\boldsymbol{\theta}}_k$s from $\widehat{\overline{\boldsymbol{\theta}}}$. Thus, in order to compute the rejection region of the test, the distribution of the test-statistic $D_n^2$ is approximated considering a technique which is essentially a combination of the approach pursued in [5], exploiting ideas in [2], with simulations.

DRAFT

## 3. Simulation study

The performance (in terms of significance level and power) of the suggested approach was evaluated through a simulation study considering both a treatment variable and an outcome variable with three levels. Following the procedure showed in [5], the data were randomly generated from the Bayesian network, displayed in Figure 1.




| Event | Probability |
|---|---|
| $X_1 = 1$ | 0.6553 |
| $X_2 = 1 \mid X_1 = 0$ | 0.9961 |
| $X_2 = 1 \mid X_1 = 1$ | 0.4384 |
| $X_3 = 1 \mid X_1 = 0$ | 0.2186 |
| $X_3 = 2 \mid X_1 = 0$ | 0.6238 |
| $X_3 = 1 \mid X_1 = 1$ | 0.2102 |
| $X_3 = 2 \mid X_1 = 1$ | 0.0556 |
| $X_4 = 1 \mid X_1 = 0$ | 0.1600 |
| $X_4 = 2 \mid X_1 = 0$ | 0.6778 |
| $X_4 = 1 \mid X_1 = 1$ | 0.4151 |
| $X_4 = 2 \mid X_1 = 1$ | 0.5150 |
| $X_5 = 1 \mid X_3 = 0$ | 0.4536 |
| $X_5 = 1 \mid X_3 = 1$ | 0.6636 |
| $X_5 = 1 \mid X_3 = 2$ | 0.9726 |
| $X_6 = 1 \mid X_5 = 0$ | 0.3967 |
| $X_6 = 1 \mid X_5 = 1$ | 0.7000 |
| $T = 1 \mid X_5 = 0, X_6 = 0$ | 0.2 |
| $T = 1 \mid X_5 = 0, X_6 = 1$ | 0.3 |
| $T = 1 \mid X_5 = 1, X_6 = 0$ | 0.3 |
| $T = 1 \mid X_5 = 1, X_6 = 1$ | 0.2 |
| $T = 2 \mid X_5 = 0, X_6 = 0$ | 0.2 |
| $T = 2 \mid X_5 = 0, X_6 = 1$ | 0.3 |
| $T = 2 \mid X_5 = 1, X_6 = 0$ | 0.4 |
| $T = 2 \mid X_5 = 1, X_6 = 1$ | 0.2 |

Figure 1: BN used for simulating the data.

Once the treatment variable and the other categorical variables were generated, the potential outcomes $Y_{(0)}$, $Y_{(1)}$ and $Y_{(2)}$ were generated from a Multinomial distribution:

$$Y_{(k)}|X_5,X_6 \sim Multinomial(P(Y_{(k)}=0|X_5,X_6),P(Y_{(k)}=1|X_5,X_6),P(Y_{(k)}=2|X_5,X_6)), \quad k=0,1,2 \tag{10}$$

and linked to the covariates through logistic models:

$$P(Y_{(k)}=0|X_5,X_6) = \frac{1}{1+e^{f_{1k}(X_5,X_6)}+e^{f_{2k}(X_5,X_6)}},$$
$$P(Y_{(k)}=1|X_5,X_6) = \frac{e^{f_{1k}(X_5,X_6)}}{1+e^{f_{1k}(X_5,X_6)}+e^{f_{2k}(X_5,X_6)}},$$
$$P(Y_{(k)}=2|X_5,X_6) = \frac{e^{f_{2k}(X_5,X_6)}}{1+e^{f_{1k}(X_5,X_6)}+e^{f_{2k}(X_5,X_6)}},$$

where

$$e^{f_{10}(X_5,X_6)} = \alpha_0+\beta_1X_5+\delta_1X_6,$$
$$e^{f_{11}(X_5,X_6)} = \alpha_0+\alpha_1+\beta_1X_5+\delta_1X_6,$$
$$e^{f_{12}(X_5,X_6)} = \alpha_0+\alpha_2+\beta_1X_5+\delta_1X_6,$$
$$e^{f_{20}(X_5,X_6)} = \alpha_3+\beta_2X_5+\delta_2X_6,$$
$$e^{f_{21}(X_5,X_6)} = \alpha_3+\alpha_4+\beta_2X_5+\delta_2X_6,$$
$$e^{f_{22}(X_5,X_6)} = \alpha_4+\alpha_5+\beta_2X_5+\delta_2X_6.$$

DRAFT

The performance of the method is evaluated in three different cases: (i) there is no treatment effect; (ii) there is a treatment effect only for a treatment level ($\boldsymbol{\theta}_0 = \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$); (iii) there is treatment effect for all treatment levels ($\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$). For each case we explored scenarios where, except for the treatment, the potential outcomes depended on: (a) no variable among $X_1, \ldots, X_6$ (scenarios S1, S2, S3); (b) two variables $X_5$ and $X_6$ (scenarios S4, S5, S6). The simulation parameters of the different scenarios are reported in Table 1. Finally, the outcome Y was generated as in (1).

Table 1: Parameter values of the models of for simulating the six scenarios.

| Scenario | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $\delta_1$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\beta_2$ | $\delta_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| S1 ($H_0$ true) | -1 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 |
| S2 ($H_0$ false - $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | -1 | 1 | 0 | 0 | 0 | -1 | 0.5 | 0 | 0 | 0 |
| S3 ($H_0$ false - $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | -1 | 2.5 | 1.5 | 0 | 0 | -1 | 1.5 | 1 | 0 | 0 |
| S4 ($H_0$ true) | -1 | 0 | 0 | 1 | -2 | -1 | 0 | 0 | 2 | -3 |
| S5 ($H_0$ false - $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | -1 | 1 | 0 | 1 | -2 | -1 | 0.5 | 0 | 2 | -3 |
| S6 ($H_0$ false - $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | -1 | 2.5 | 1.5 | 1 | -1 | -1 | 1.5 | 1 | 2 | -3 |

Four sample sizes (n =500, 1000, 2500, 5000) have been explored with 1000 Monte Carlo runs for each combination of scenario and sample size. Table 2 summarizes the Monte Carlo rejection probabilities of the null hypothesis for different scenarios and sample sizes. The simulated significance level is always lower than the nominal level 5% (S1 and S4) and when the null hypothesis $H_0$ of absence of treatment effect is false (S2, S3, S5, S6), simulated the power is at least 0.38 when the sample size is 500 and greater than 0.8 as the sample size increases.

Table 2: Rejection probabilities (nominal significance level 0.95)

| Scenario | n=500 | n=1000 | n=2500 | n=5000 |
|---|---|---|---|---|
| S1 ($H_0$ true) | 0 | 0.01 | 0 | 0 |
| S2 ($H_0$ false - $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | 0.6 | 0.93 | 1 | 1 |
| S3 ($H_0$ false - $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | 1 | 1 | 1 | 1 |
| S4 ($H_0$ true) | 0.01 | 0 | 0 | 0 |
| S5 ($H_0$ false - $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | 0.38 | 0.84 | 1 | 1 |
| S6 ($H_0$ false - $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_1$) | 1 | 1 | 1 | 1 |

DRAFT

## 4. Remarks and perspectives

In the present paper, a new method to assess the treatment effect in case of multi-level treatment and multi-level outcome is presented. In particular, (i) we proposed to estimate the potential outcomes distributions applying a methodology based on the inverse probability weighting with the propensity score estimated using a Bayesian Network, and (ii) we developed a new test for the absence of treatment effect. A simulation study showed the good performance of the developed test.
In the future, the proposed approach will be also evaluated through an application to real data of prostate cancer patients. Moreover, this work will be extended developing tests for stochastic ordering among treatments effects.

## References

[1] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J.: Probabilistic Networks and Expert Systems. Springer Verlag, New York (1999).

[2] Hirano, K., Imbens, G. W., & Ridder, G.: Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica, 71, 1161-1189 (2003).
[3] Marella, D. and Vicard, P.: Bayesian network structural learning from complex survey data: a resampling based approach, Statistical Methids & Applications, 31, 981-1013 (2003).
[4] McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F.: A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Statistics in medicine, 32(19), 3388-3414 (2013).
[5] Vicard, P., Rancoita, P. M. V., Cugnata, F., Briganti, A., Mecatti, F., Di Serio, C., & Conti, P. L.: Can Bayesian Network empower propensity score estimation from Real World Data?. arXiv preprint arXiv:2302.07663 (2023).

DRAFT

# Are European consumers willing to pay the true price for sustainable food?

Luca Secondi[a], Mengting Yu[b]

[a] University of Tuscia, Department for Innovation in Biological, Agro-food and Forest Systems
secondi@unitus.it ,
[b] University of Tuscia, Department of Economics, Engineering, Society and Business Organization
mengting.yu@unitus.it

**Abstract**

Our current food systems are not sustainable as the negative impacts on our environment, society, and individuals' health are still intense. The transition to sustainable food systems is set as one of the Sustainable Development Goals. Existing literature suggests that the true price of food regulates a more sustainable food system while building up trust in consumers and further persuading their purchase intention. Indeed, consumers play a key role to accelerate the transition to sustainable food systems. The question is to what extent consumers are willing to pay the true price for sustainable food. What is influencing consumers' Willingness to Pay? Can values which are conveyed through a true price increase purchase intention? We carried out an in-depth analysis of a large-scale survey targeting individuals' knowledge and attitudes towards food sustainability in EU-27. Our findings suggest the importance of the attitudes towards the environment, and organic products as well as of economic affordability that policymakers and stakeholders should consider.

**Keywords:** Willingness To Pay, true price, food sustainability, ordinal logistic regression

DRAFT

## 1. Introduction

Before buying a food product, do you check the price? Do you fully understand the information that the price label conveys? What if you would be told that 10% of the price represents the value of "green production" and the guaranteed welfare of the workers? Will you buy this product even if it is 10% more expensive than another one without any specifications on the same shelf?

In the 2021 Food Systems Summit report, the scientific group pointed out that our current food systems were still not sustainable (Hendriks et al., 2021), with a specific focus on the impacts of the current food systems on our environment, society, and health (FAO et al., 2020). A transition towards more sustainable food systems had been called to reduce the related impacts and costs while offering affordable and healthy food to all. To quantify these potential benefits, we could project a 21-37% cut of the net anthropogenic GHG emissions (IPCC, 2019), a 690 million undernourished population and more than 10 million lives saved (GBD, 2019; FAO et al., 2020).

In recent years, much research has been carried out aimed at improving the sustainability of the food system. Among the various perspectives of analysis and factors studied, "externalities" have been introduced. Indeed, the costs and the benefits of those "food" externalities are not included in the market prices, and as a result, both consumers and the stakeholders of the food systems are not aware of the negative hidden effects (Baker et al. 2020), classified into environmental, social, health, and economic externalities. The inclusion of those issues within the product prices would help realize the true cost to the stakeholders in the value chain and the true price to the consumers. Scholars have carried out studies on the True Cost Accounting (TCA) approach with the externalities to unfold the hidden impact on the environment, economy, and society to the industrial stakeholders, consumers, and policymakers; meanwhile, studies on consumer perception and potential acceptance of the "true price" are increasing, because ultimately the acceptance of the "true price" by the market (consumers) will accelerate the transition to sustainable food systems. Moreover, stakeholders in the food systems will gain the consumers' trust when they successfully demonstrate the values with the true price (Taufik, van Haaster-de Winter, and Reinders 2023).

However, previous studies carrying out comprehensive research on consumers' perception and acceptance of the true price are limited. Most of the existing studies have covered the selected typical

externalities or have targeted a specific area, for instance, the consumer's motivation and perception towards environmental and social externalities via food product labelling such as carbon footprint, fair trade, rainforest protection, and animal welfare (Grunert, Hieke, and Wills 2014a); Shao (2018) studied the consumers' Willingness To Pay (WTP) relating to the environment protection in China – an analysis on both micro (individual) factors and macro (city) influences.

The aim of this paper is to provide an in-depth analysis of the potential linkages between EU-27-country consumers' WTP for the true price of food. We consider the individual-level factors including personal habits and attitudes and conduct the first exploration of the role of heterogeneity at territorial (country) levels.

## 2. Data and descriptive analysis

Our study refers to the latest available Special Eurobarometer (SE) survey (93.2), carried out between 3 August to 15 September 2020 on a representative sample of approximately 28,300 individuals across 27 EU Member States (European Commission, Brussels 2021). The survey was designed to collect public opinions towards the strategy of transforming our food system set by SDGs. With the aim of capturing the individuals' appetite for change, a wide range of questions was designed around food buying and eating habits, knowledge about food sustainability and sustainable diet. Meanwhile, to engage the whole food system, citizens' attitudes and opinions towards the roles they play in sustainable food systems were examined.

The survey included a complete section dedicated to food sustainability to explore consumers' food buying and eating habits, knowledge about sustainable food and diet, and opinions of the key roles to play in food sustainability. The attitudes of the interviewees were captured by an ordinal scale of measurement ranging from *Totally Agree* to *Totally Disagree categories*. To study the sample population's WTP for sustainable food, we specifically referred to the following statement *"You are prepared to pay 10% more for agricultural products that are produced in a way that limits their carbon footprint"* as our target variable.

Our descriptive analysis on the SE 93.2 microdata shows a country-level heterogeneity towards WTP for sustainable food among 27 EU Member States (Figure 1) with overall responses identifying 22.26% of interviewees totally agree with spending more for sustainable food and 42.56% tend to agree on spending more (up to 10%) on sustainable food products, while nearly 11.66% disagreed with a firm attitude.

Among 27 EU Member States, Germany (39.5%), Denmark (39%), the Netherlands (38.9%), Sweden (34.3%), Cyprus (33.4%), Luxembourg (32.7%), and Slovenia (31.7%) showed the highest WTP when considering the individuals' opinions of *Totally Agree*. When we merged the opinions of *Tend to Agree* with *Totally Agree*, we obtained more countries in our results, for example, Ireland and Belgium. On the other hand, Lithuania (20.1%), Bulgaria (19.9%), Hungary (18.7%), Latvia (18.2%), and Slovakia (17.6%) had the highest disagreement voices when we only considered the individuals' opinions of *Totally Disagree*. Again, if we included the answers of *Tend to Disagree*, Portugal had the highest (sample) population (54.7%) who voted *Totally Disagree* and *Tend to Disagree*.

DRAFT

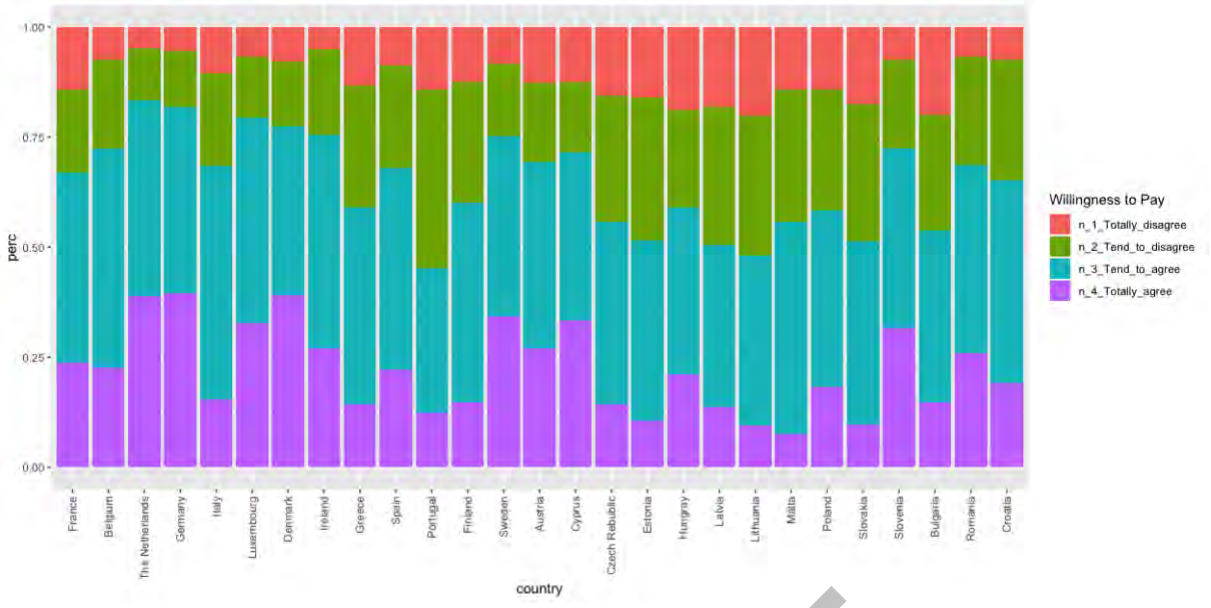



Figure 1: Willingness-to-Pay in percentage by country

Youths (those aged between 15 and 35 years old) showed higher WTP, and so did the education level. An individual's WTP rose with his/her education level. Focusing on the individual's subjective social classification which was divided into 5 categories as underlined in Figure 2, the upper middle class (36%) showed more WTP than the higher class (33%). Gender and residence areas did not display diverging patterns. People working in management positions in the companies had the highest WTP (29%), followed by self-employed (25.9%) consumers and students (25.4%). Indeed, unemployed and retired-or-unable-to-work-due-to-illness groups reported a higher disagreement rate (14.5% and 13.9% respectively), although a proportion of retired-or-unable-to-work-due-to-illness group was willing to pay more for sustainable food (21.5%), which was higher than the figures of the employed (not management roles; 21%), housewife or houseman (19%), and the unemployed (18%).

DRAFT

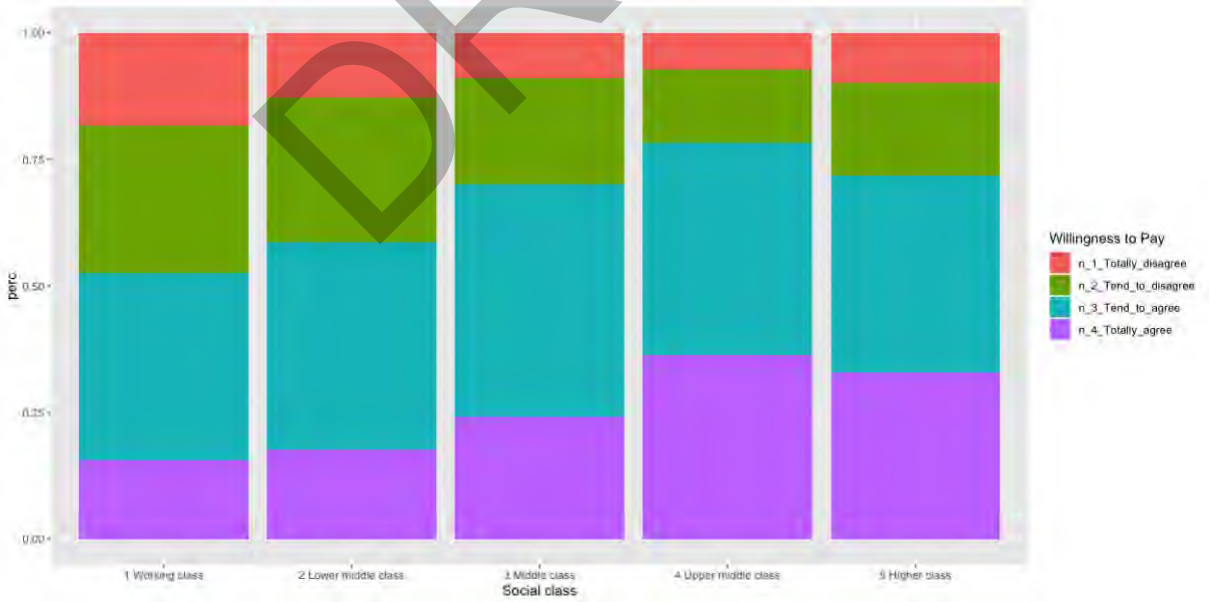



Figure 2. Willingness-to-Pay in percentage by subjective social-class

Safety, cost, nutrition, taste, and origin of the food products were important conditions when most individuals made their purchase decision. Most people connected sustainable food with three sets of descriptions: nutritious and healthy, affordable, and with little or no pesticides. Moreover, most people thought that a sustainable diet represented health and was beneficial to the local economy in the first place. Regarding individuals' opinions towards the key roles to play in a sustainable food system, 82.3%

thought food producers and manufacturers should play a key role, followed by government or NGOs (9.2%) and retailers (4.3%). Only 1.2% considered themselves an important role in food sustainability. Most interviewees believed that making sustainable food more affordable (55.3%), more accessible in the shops (48.7%), and clearer labelling (37.5%) would help them to adopt healthy and sustainable diets. More people agreed that public institutions should offer sustainable food (52.7% *Totally Agree*; 40.8% *Tend to Agree*) and 44% of the sample population strongly agreed that information about sustainability should be compulsory on food labels, and another 43.5% held the opinion of *Tend to Agree*.

8.9% and 57.5% of interviewees reported that they had a frequent sustainable diet or took a sustainable diet most of the time, respectively; only 4.3% never had a sustainable diet. Interestingly, we found out that, although sustainable-food frequenters were most willing to pay for more expensive sustainable food (33.6%), the disagreement in this group was not the least (12.2%), but people who claimed a sustainable diet most of the time had the least disagreement (8.8%).

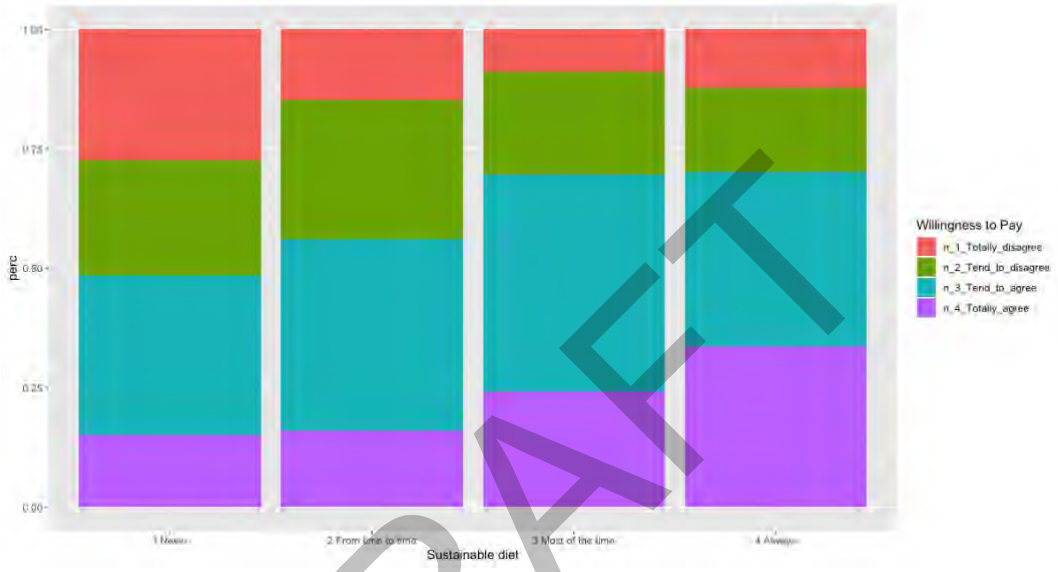



Figure 4. Willingness-to-Pay in percentage by the sustainable diet

## 3. Methodology

Given the above-mentioned target variable, we considered the WTP in our analysis as an ordered variable with a ranking nature. We decided to implement the Ordinal Logistic Regression (OLR) model, which has been widely used to analyze ordinal and categorical outcomes (Shao, Tian, and Fan 2018; Loan et al. 2019; Long and Freese 2006; McCullagh 1980). In our data-cleaning process, we reorganized the values of all ordinal variables, so that they shared the same rising tendency with other explanatory variables.

WTP=1 represents the lowest WTP and acceptance of the price increase (up to 10%) and WTP=4 stands for the highest WTP and the least tolerance to the price change. As the basis of the OLR model, we set up the following OLR model:

$$WTP_i^* = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i \quad (1)$$

Where $\boldsymbol{x}$ is the vector of explanatory variables; $\boldsymbol{\beta}$ is the vector of parameters; $\varepsilon$ is the random error and $i$ represents the individual respondents. The latent WTP for more expensive (up to 10% price increase) sustainable food, ***WTP*,*** is a linear function of the vector of explanatory variables in the OLR model.

Therefore, we used the observed responses "*Totally disagree*" (WTP=1), "*Tend to disagree*" (WTP=2), "*Tend to agree*" (WTP=3), and "*Totally agree*" (WTP=4) to construct the ordinal categories with the cut points $\boldsymbol{C_j}$ as following:

DRAFT

$$WTP = 1 \Rightarrow Totally\ disagree\ (0\%\ price\ increase), if\ -\infty < WTP_i^* \leq C_1$$

$$WTP = 2 \Rightarrow Tend\ to\ disagree\ (0 - 10\%\ price\ increase), if C_1 < WTP_i^* \leq C_2$$

$$WTP = 3 \Rightarrow Tend\ to\ agree\ (0 - 10\%\ price\ increase), if C_2 < WTP_i^* \leq C_3$$

$$WTP = 4 \Rightarrow Totally\ agree\ (10\%\ price\ increase), if C_3 < WTP_i^* \leq +\infty \quad (2)$$

The probability that respondent ***i*** chooses ***k*** is expressed accordingly:

$$\begin{aligned} P_{ik} &= pr(WTP_i = k|x_i) \\ &= pr(WTP_{ik} = 1|x_i) \\ &= pr(C_{k-1} < x_i'\beta + \varepsilon_i < C_k) \\ &= F(C_k - x_i'\beta) - F(C_{k-1} - x_i'\beta) \end{aligned} \quad (3)$$

## 4. Ordinal Logistic Regression (OLR) model: provisional results

The OLR estimation results suggested that single or married (or with a partner) individuals tended to have higher WTP than divorced, separated or widow individuals. People aged between 56 to 75 years showed the highest level of disagreement with sustainable food products with increased prices. WTP decreased among housewives or housemen. Individuals who self-reported to be from a higher social class indicated a higher WTP. Interestingly, individuals with a lower education still showed a higher WTP significantly.

Compared with the demographic factors, individuals' knowledge and attitudes towards sustainable food had more influence. The individuals who had a sustainable food diet tended to show a higher WTP. People who had strong ethics and beliefs and took convenience as an important factor for general food purchase decisions had higher WTP, but those who cared about the environment, food origin, nutrition, and safety when buying food showed lower WTP. In terms of promoting sustainable food to consumers, it might be more difficult to convince those who voted for affordability. However, sustainable food promotion through education campaigns on its health benefits, simply facilitating better menus with sustainable food choices or ensuring a clearer label and more access to sustainable food products in the shops, might potentially increase the individual's WTP.

17 countries (Total: 27 EU Member States) were found to be significantly associated with individuals' WTP. The Netherlands, Germany, Luxembourg, Denmark, Ireland, Sweden, Slovenia, and Croatia indicated a higher WTP; Italy, Greece, Estonia, Latvia, Lithuania, Malta, Poland, Slovakia, and Bulgaria showed the opposite results (lower WTP).

## 5. Conclusion and limitation

We find out that the younger generation tends to accept the price increase regarding sustainability, and this echoes the similar findings that students have a higher WTP as well. Higher education results in higher WTP, but the richer (higher social class) does not necessarily show a higher WTP. However, demo-social-economic factors seem less influence on individuals' WTP in comparison with an individual's knowledge and attitude towards food sustainability, which was agreed by other scholars (Laroche, Bergeron, and Barbaro-Forleo 2001).

The results suggest a positive relationship between the consumer's attention to food products' affordability and the WTP. Therefore, a true price should persuade consumers that transparent pricing from a sustainable food system indeed is affordable and responsible for our planet and society. People who relate environmental issues (pollution, carbon footprint, etc), minimal processes and short supply chains with food sustainability displayed higher WTP. Stakeholders in food systems should consider specifying the true price which underlines these extra values of producing sustainable food to win the trust of the consumers and consumers' purchase intention (Taufik, van Haaster-de Winter, and Reinders 2023). Our results show that food producers or manufacturers fail to persuade consumers to accept sustainable food products with higher prices now, and they are expected to play the most important role to promote food sustainability. In the lower stream of the food value chain, more sustainable food

DRAFT

choices with easier access to food outlets (retail and food service) could be helpful to increase consumers' acceptance of more expensive sustainable food products. Meanwhile, consumers with higher WTP expect better labelling to communicate the sustainable contributions or impacts of a food product, and regulations to be implemented in the value chain for food sustainability development.

Different countries display different influences on their citizens' WTP. Previous studies incline that in different nations with various GDP levels, the public concern ranges from survival and material issues to post-material values, such as welfare and environment, therefore the WTP varies; however, it is common that complex and even contrary relations exist between the motivation or concern about sustainability and its implementation (Shao, Tian, and Fan 2018; Grunert, Hieke, and Wills 2014b). The hidden factors influencing individuals in a country include many, for example, policies, living standards, traditions and so on. However, it reminds us that the influence from a country level could be more efficiently assessed. Therefore, researchers and governments should take a leading joint role in examining the food sustainability transition in depth.

Since the survey was conducted in 2020 and released in 2021, we foresee the limitation of the data, and we aim to update our analysis results when the new wave of micro-data will be released. In addition, our quantitative analysis was not completed after the implementation of the OLR model and further work is still ongoing to refine micro-level variables by considering specific country-level factors in the model selection and specification. Lastly, an extension of this analysis could be to continue our research through a TPA approach based on the latent WTP threshold values.

## References


[1] Baker, Lauren, Guillermo Castilleja, Adrian De Groot Ruiz, and Adele Jones. 2020. "Prospects for the True Cost Accounting of Food Systems." Nature Food 1 (12): 765–67. https://doi.org/10.1038/s43016-020-00193-6.

[2] European Commission, Brussels. 2021. "Eurobarometer 93.2 (2020)Eurobarometer 93.2 (2020): Europeans, Agriculture and the CAP, Making Our Food Fit for the Future – Citizens' Expectations, and Attitudes of Europeans towards Tobacco and Electronic Cigarettes (COVID-19 Pandemic): Europeans, Agriculture and the CAP, Making Our Food Fit for the Future – Citizens' Expectations, and Attitudes of Europeans towards Tobacco and Electronic Cigarettes (COVID-19 Pandemic)." GESIS Data Archive. https://doi.org/10.4232/1.13706.

[3] FAO, IFAD, UNICEF, WFP, & WHO. (2020). The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets. Food and Agriculture Organisation of the United Nations, Rome, FAO. https://doi. org/10.4060/ca9692en.

[4] GBD (2019). Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. https://doi.org/10.1016/S0140-6736(19)30041-8.

[5] Grunert, Klaus G., Sophie Hieke, and Josephine Wills. 2014a. "Sustainability Labels on Food Products: Consumer Motivation, Understanding and Use." Food Policy 44 (February): 177–89. https://doi.org/10.1016/j.foodpol.2013.12.001.

[6] Hendriks, Sheryl, Adrian de Groot Ruiz, Mario Herrero Acosta, Hans Baumers, Pietro Galgani, Daniel Mason-D'Croz, Cecile Godde, Katharina Waha, and Dimitra Kanidou. 2021 "The True Cost and True Price of Food.", available at https://sc-fss2021.org/wp-content/uploads/2021/06/UNFSS_true_cost_of_food.pdf

[7] IPPC (2019). Climate change and land, an IPCC spe- cial report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems, IPPC, www.ipcc.ch/report/srcc

[8] Laroche, Michel, Jasmin Bergeron, and Guido Barbaro-Forleo. 2001. "Targeting Consumers Who Are Willing to Pay More for Environmentally Friendly Products." Journal of Consumer Marketing 18 (6): 503–20. https://doi.org/10.1108/EUM0000000006155.

[9] Loan, Le Thi Thanh, Yoshifumi Takahashi, Hisako Nomura, and Mitsuyasu Yabe. 2019. "Modeling Home Composting Behavior toward Sustainable Municipal Organic Waste Management at the Source in Developing Countries." Resources, Conservation and Recycling 140 (January): 65–71. https://doi.org/10.1016/j.resconrec.2018.08.016.

[10] Long, J. Scott, and Jeremy Freese. 2006. Regression Models for Categorical Dependent Variables Using Stata, Second Edition. Stata Press.

[11] McCullagh, Peter. 1980. "Regression Models for Ordinal Data." Journal of the Royal Statistical Society: Series B (Methodological) 42 (2): 109–27. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.


DRAFT

[12] Shao, Shuai, Zhihua Tian, and Meiting Fan. 2018. “Do the Rich Have Stronger Willingness to Pay for Environmental Protection? New Evidence from a Survey in China.” World Development 105 (May): 83–94. https://doi.org/10.1016/j.worlddev.2017.12.033.
[13] Taufik, Danny, Mariët A. van Haaster-de Winter, and Machiel J. Reinders. 2023. “Creating Trust and Consumer Value for True Price Food Products.” Journal of Cleaner Production 390 (March): 136145. https://doi.org/10.1016/j.jclepro.2023.136145.

DRAFT

# Can the reliability of composite indexes be impacted by uncertainty of individual indicators?

Caterina Giusti[a], Stefano Marchetti[a], and Vincenzo Mauro[b]

[a] Department of Economics and Management, University of Pisa, Pisa, Italy; caterina.giusti@unipi.it, stefano.marchetti@unipi.it

[b] Department of Political Science, Communication and International Relations, University of Macerata, Macerata, Italy; vincenzo.mauro@unimc.it

**Abstract**

In recent years, there has been a significant transformation in how important social and natural phenomena are assessed and measured, with traditional single-variable measurements being replaced by multidimensional approaches. Key to these approaches are composite indexes (CIs), which are real-value functions that synthesise multiple achievements of a group of units.

Each of the dimensions under study are typically synthesised through one or more variables, usually referred to as indicators. The reliability of CIs is affected by estimation errors, such as sampling error, that might be non-negligible when indicators are obtained through estimation processes.

In this paper, we propose a methodology based on a parametric bootstrap technique to evaluate the effect of uncertainty in indicators on the reliability of composite indexes, with a focus on indicators whose sampling error may be correlated, such as indicators obtained from a common sample survey. The aim is to apply this method to a CI proposed to evaluate Italian regions' environmental performances. The proposed methodology can be generalized to address other types of errors, such as measurement, non-response, or non-sampling errors.

**Keywords:** bootstrap, correlated sampling errors, ecological indicators

DRAFT

## 1. Introduction

In recent years, the way in which significant social and natural phenomena are measured and evaluated has undergone a significant transformation. This is largely due to the widespread agreement among the scientific community regarding the complex and multi-faceted nature of subjects such as well-being, poverty, and the environment, and the need to have a more comprehensive approach to defining and measuring them. Traditional methods of measurement, which were based on a single variable, are now frequently being replaced by multidimensional techniques (Sen, 1999; Chakravarty and D'Ambrosio, 2006; Alkire and Foster, 2011; Bossert et al., 2012; Nicholas et al., 2017).

These methods tackle the subject of investigation from a wider viewpoint by incorporating a large number of dimensions. Depending on the theoretical model selected, these dimensions (often referred to as "achievements" or "indicators") are either considered as outcomes of the underlying concepts under study, or as their root cause.

Many studies that are based on a multidimensional approach use composite indices (CIs) as a tool to reduce the complexity of data. In recent years, the use of CIs has significantly increased. Bandura (2009) noted the use of hundreds of CIs covering a wide range of topics, including well-being, the environment, and child poverty. The growth in both structured and unstructured data, often referred to as "big data", is a major factor that is driving the development of multidimensional frameworks. The abundance of data has increased interest in multidimensional approaches, creating at the same time new theoretical and methodological challenges.

The indicators used to define the CIs are often obtained from different datasets, or are measured on subgroups of the target population. This sometimes results in a mixed framework, where some indicators are measured from the overall population and others from samples of the population. This can cause difficulties in measuring the impact of sampling variability on the aggregated CIs.

This is particularly true for complex areas, where the variables included in the analysis can be extremely diverse and belong to different fields of study. For instance, the sustainable development index provides a one-dimensional metric to evaluate country-specific information on different dimensions like economic, environmental, and social.

Moreover, measures based on subjective processes, such as opinions or perceptions, are often derived from sample surveys, adding an additional source of heterogeneity between indicators.

Sampling variability is only one of the potential sources of error that must be taken into account. Other sources of error, such as non-response and measurement errors, can also affect the data used to calculate the individual indicators that make up the composite index (CI). In this paper, we only focus on sampling error as a source of uncertainty affecting the indicators, but our methodology can be expanded to consider other sources of error.

Traditionally, most of the literature on the development of CIs focused on the effects of normalization procedures, the combination of different indicators (linear or non-linear), and the sensitivity of CIs to methods of aggregation and weighting. Some authors examine the variability among the indicators included in the CIs, while others consider the uncertainty in the construction process of a CI. However, as Ceccarelli et al. (2020) point out, the sensitivity of CIs due to uncertainty in the indicators has often been overlooked in the literature, and finding measures of accuracy for a composite index remains an unsettled problem.

## 2. Estimating uncertainty in composite indicators

Mauro at al. (2021) attempted to address the issue introduced above by studying the impact of uncertainty due to the sampling error in the true values of indicators on the final aggregate score. They presented a method for obtaining estimates of confidence intervals and their variability.

The method was applied to real data indicators that were combined to create a CI that represented the environmental performance of different regions in Italy. The proposed method effectively captures how the sampling error in individual indicators affects the corresponding error in the CI. The results of the study also provided insight into how errors CIs should be interpreted, such as ranking the environmental performance of Italian regions.

Mauro at al. (2021) proposed a solution by accounting for the sampling error in the aggregation phase of CIs through the use of a parametric bootstrap technique. The sampling distribution of each indicator was replicated to obtain a distribution, which was used to determine the standard errors and confidence intervals for the CI. In order to test the results, the proposed technique was therefore applied to four aggregation methods: arithmetic mean, geometric mean, Mazziotta-Pareto index and multidimensional synthesis of indicators.

In their work, Mauro et al. (2021) rely on a three basic hypothesis, namely:

1. The sampling errors of the indicators are independent of each other within units, so that the sampling error of indicator *j* for unit *i* does not depend on the sampling errors of the other *k-1* indicators of unit *i*, where *k* is the number of indicators available for each unit.
2. The estimates are unbiased with regards to the survey design.
3. The variance, $\sigma_{ij}^2$, of the estimates is known for all *i* and *j*. Usually, $\sigma_{ij}^2$ is unknown and estimated from the sample data, but the estimator is often smoothed and treated as the true sampling variance.

The first assumption rules out potential correlations among the sampling errors of the indicators. Specifically, the approach assumes that the covariance matrix is diagonal, which may be overly restrictive: for example, if the indicators are obtained from the same surveys, they are likely to present correlated sampling errors. This case is referred to as "horizontal correlation".

Let $\widehat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{e}_i$, where $\widehat{\boldsymbol{\theta}}_i = \left[\hat{\theta}_{i1}, \dots, \hat{\theta}_{ik}\right]^T$ is the vector of the estimates for the *k* indicators for unit *i*, $\boldsymbol{\theta}_i = [\theta_{i1}, \dots, \theta_{ik}]^T$ the vector of their true (unknown) values and $\boldsymbol{e}_i = [e_{i1}, \dots, e_{ik}]^T$ is the vector of

DRAFT

their sampling errors. Therefore, we can assume $\boldsymbol{e}_i \sim D_k(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i$ is its covariance matrix and $D_k$ is a multivariate distribution, usually a multivariate normal of dimension $k$, because of the central limit theorem.

Using a parametric bootstrap, we estimate the variance of CIs, taking into account the sampling variability of single indicators and their horizontal correlation. The proposed technique is as follow:

1. Generate $B$ bootstrap vectors $\widehat{\boldsymbol{\theta}}_i^b = \left[\hat{\theta}_{i1}^b, \dots, \hat{\theta}_{ik}^b\right]^T$ sampling from $D_k\left(\widehat{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma}_i\right)$, so that $\widehat{\boldsymbol{\theta}}_i^b$ mimic a realization of the unknown sampling distribution of $\widehat{\boldsymbol{\theta}}_i$, which is $\widehat{\boldsymbol{\theta}}_i \sim D_k(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$; then stack together the $n$ vectors $\widehat{\boldsymbol{\theta}}_i^b$ to obtain the matrix $\boldsymbol{X}^b$;
2. Starting from $\boldsymbol{X}^b$, compute $B$ standardized matrix $\boldsymbol{\Xi}^b$, where the entries of this matrix are obtained as $\xi_{ij}^b = \frac{x_{ij}^b - \min\left(x_{1j}^b, \dots x_{nj}^b\right)}{\max\left(x_{1j}^b, \dots x_{nj}^b\right) - \min\left(x_{1j}^b, \dots x_{nj}^b\right)}$ or $\xi_{ij}^b = \frac{\max\left(x_{1j}^b, \dots x_{nj}^b\right) - x_{ij}^b}{\max\left(x_{1j}^b, \dots x_{nj}^b\right) - \min\left(x_{1j}^b, \dots x_{nj}^b\right)}$ according to positive or negative polarity of the indicator respectively;
3. Using $\Xi^b$, obtain estimates of CIs, denoted as $\tau_i^b$, $b = 1, \dots, B$ and $i = 1, \dots, n$;
4. Obtain estimated variance from the $B$-vector of CIs: $\hat{V}(\tau_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\tau_i^b - \bar{\tau}_i\right)^2$, where $\bar{\tau} = \frac{1}{B}\sum_{b=1}^{B}\tau_i^b$.

Usually, when indicators are means, totals or proportions, the distribution $D_k$ is a multivariate normal distribution. The covariance matrix of the sampling errors is considered known here, and in real applications it has to be estimated from survey data.

Our goal is to measure if and how the relaxation of the first assumption by Mauro et al. (2021) affects the overall uncertainty of the CIs obtained with different aggregation methods. In their case study, Mauro et al. (2021) analyse data from multiple sources, including administrative archives and two national sample surveys (EU-SILC and HBS), as well as the Istat sample survey on Aspects of Daily Life, to address this limitation. All the estimates were therefore calculated assuming independence between sampling errors, although this was not the case from some of the indicators coming from the same survey. In general, the hypothesis of independence may not hold for analyses where indicators are obtained from a limited number of datasets or a unique dataset. In these cases, the correlation between the variances of the indicators can significantly affect the overall uncertainty of the CIs if not properly addressed. Therefore, the uncertainty in the CIs estimates can be caused by both the variance of individual indicators and the covariance matrix, whose structure is not diagonal. The purpose of this paper is to provide a preliminary assessment of the impact of indicators' sampling errors on CIs in a more general context.

A design-based simulation has been carried out using data described in section 3. A correlation between 0.3 and 0.8 has been set among indicators coming from the same survey, and main results indicates that the proposed bootstrap perform very well, particularly for the AMPI. Detailed results of the simulation are not reported here and are available upon request to the authors.

DRAFT

## 3. Application

As already underlined, CIs are nowadays widely used to measure complex phenomena in many fields of studies. Especially in the last years, with the increase of data available from surveys, administrative archives and alternative data sources, the awareness that many phenomena can be better measured using a set of indicators, instead that as single one, is widely recognized. A noteworthy example is the Italian National Statistical Institute's (Istat) Equitable and Sustainable Well-being (BES), where 12 domains of analysis are each measured by several indicators to obtain a detailed framework for monitoring well-being in Italy. Aggregating the information of the well-being indicators into a reduced number of composite indicators, for example one indicator within each well-being domain, can be crucial for policy-making and benchmarking when the units of analysis are the Italian regions or provinces (Ciommi et al., 2017).

Based on the BES, Mauro et al. (2021) proposed a CI indicator that can be used to monitor the environmental performance of Italian regions. The CI is composed by 14 indicators (Table 1): 7 coming

from administrative archives and therefore not affected by sampling error, and seven coming from sample surveys. Specifically, four indicators come from the Istat survey on 'Aspects of daily life', two from the EU-SILC survey ('European Union - Statistics on Income and Living Conditions') and one from the HBS ('Household Budget Survey'). These indicators are affected by sampling error, and as shown by Mauro et al. (2021), this kind error also affects the corresponding CI.

Table 1: Indicators used to define the proposed CI to evaluate Italian regions' environmental performance: name, source, polarity. Indicators non affected by sampling error are in italics.

| Indicator | Source | Polarity |
|---|---|---|
| Dissatisfaction for the landscape deterioration (%) | Istat survey on Aspects of daily life | - |
| Concern for landscape deterioration (%) | Istat survey on Aspects of daily life | - |
| Satisfaction with environmental quality (%) | Istat survey on Aspects of daily life | + |
| Concern for biodiversity loss (%) | Istat survey on Aspects of daily life | - |
| Noise from neighbours or from the street (%) | Istat survey EU-SILC | - |
| Pollution, grime or other environment problems (%) | Istat survey EU-SILC | - |
| Share of car transportation expenses on overall transportation expenses (%) | Istat survey HBS | - |
| *Losses in drinking water supply network (%)* | *Istat census on water availability for civil use* | - |
| *Exceeding of the NO2 annual limit for the protection of human health (%)* | *Istat environmental data on regional capital cities* | - |
| *Urban green areas per inhabitant* | *Istat environmental data on regional capital cities* | + |
| *Wastewater treatment failures (%)* | *Istat census on water availability for civil use* | - |
| *Protected natural areas (%)* | *Istat data based on Ministry for the Environment information* | + |
| *Energy from renewable resources (electricity, %)* | *Istat data based on Ministry for the Environment information* | + |
| *Waste sorting (%)* | *Istat data based on Ministry for the Environment information* | + |

DRAFT

However, Mauro et al. (2021) used the hypothesis of independence of the indicators' sampling errors, although this assumption is not correct in the case of the four indicators coming from the survey 'Aspects of daily life'. The aim of the current work is to remove the hypothesis of independence and to derive the corresponding effect in terms of variability of the CI.

The preliminary results obtained by relaxing the assumption of independence between errors in the estimates of the indicators still depend on some minor assumptions that could be revised, extended, or relaxed to encompass more general situations.

To our knowledge, there is no literature on estimating the impact of indicators' sampling error on CIs under this framework. The methodology proposed in this paper must therefore be considered as a first tentative to address this issue.

## References

Alkire, S., Foster, J., (2011) Counting and multidimensional poverty measurement. Journal of Public Economics. 95, 476–487.

Bandura, R., (2011) Composite indicators and rankings: Inventory 2011. Technical report. New York: Office of Development Studies, United Nations Development Programme (UNDP) (2011th ed.).

Bossert, W., Chakravarty, S., D'Ambrosio, C., (2012) Poverty and time. The Journal of Economic Inequality 10, 145–162.

Ceccarelli, C., Guandalini, A., Martini, A., Pontecorvo, M.E., 2020. Accuracy evaluation of lfs-bes indicators: A regional assessment. Social Indicator Research, 1-14
Chakravarty, S.R., DAmbrosio, C., (2006) The measurement of social exclusion. Review of Income and Wealth, 52, 377–398.
Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., Chelli, F., (2017) A new class of composite indicators for measuring well-being at the local level: An application to the equitable and sustainable well-being (bes) of the italian provinces. Ecological Indicators 76, 281–296.
Mauro V., Giusti C., Marchetti S., Pratesi M., (2021) Does uncertainty in single indicators affect the reliability of composite indexes? An application to the measurement of environmental performances of Italian regions, Ecological Indicators, 127
Nicholas, A., Ray, R., Sinha, K., (2017) Differentiating between dimensionality and duration in multidimensional measures of poverty: Methodology with an application to china. Review of Income and Wealth 60, 48–74.
Sen, A.K., (1999) Development as freedom. Oxford University Press, Oxford.

DRAFT

# Initial Coin Offerings and ESG: allies or enemies?

Alessandro Bitetto[a] and Paola Cerchiello[a]

[a]University of Pavia; alessandro.bitetto@unipv.it, paola.cerchiello@unipv.it

**Abstract**

Initial Coin Offerings (aka ICOs) have gained a prominent interest in the FinTech world as an alternative way to fund raising for innovative and cutting edge business ideas. So far, academics have studied drivers of success without posing specific attention to the products or activities proposed by the ICOs. In this paper, we investigate the possible nexus between ICOs and Enviromental, Social and Governance (ESG) indicators, by studying a set of 621 ICOs. Specifically, we extract keywords related to ESG from whitepapers associated to each ICO and build a variable which acts as a signal of attention to sustainability topics. Our research hypothesis concerns the evaluation whether ICOs oriented towards ESG are more likely to successfully raise expected funds. Preliminary results confirm such hypothesis.

***Keywords:*** ICO, Green, Sustainability

DRAFT

## 1. Introduction

Nowadays themes like Environment, Social Change, and Governance are becoming more and more important. Over 90% of CEOs believe that ESG indicators are critical to their company's earnings and progress. Indeed, the inclusion of environmental, social, and governance aspects are playing an important role in investment and emission processes, promoting innovation and the expansion of sustainable finance [7]. We could state that, for a company, Environmental, Social, Governance investments and reporting represent one of the ways to keep up with the market. As a matter of fact, companies with stronger ESG propositions tend to have higher growth, higher worker efficiency, lower volatility, cost decrease, and fewer institutional interventions. Furthermore, in recent years, start-ups and the most innovative businesses turn to alternative sources of capital instead of classic channels, such as Initial Coin Offerings (ICOs). An ICO is a new way to fund businesses and initiatives, it is one of the blockchain-based processes that allow the emission of an utility token rather than a security or equity token. The growing popularity of the ICOs is clearly due to several related benefits, such as the high level of offered return on investment, high liquidity, fast financing, cost minimization and high availability, which are increasingly encouraging innovative investors and businesses to abandon traditional financing methods. However, it is also a young and ever-changing market full of significant risks.

Previous literature on this topic is still scarce. Most relevant papers that consider the effect and success of environmental initial coin offerings were published in recent years, and the studies examine the success of environmental ICO measured by the total funding in the actual ICOs and the long-term survival of the projects [4]. Moreover, there are some articles that study how environmental issues have led to many new trends in technology and financial management. They analyze the relationship between Fintech and sustainability. These studies explain how, in recent years, investors' attention to environmental issues is increased and how investors, that are concerned about such issues, reduce the probability of long-term failure. Such considerations are consistent with the fact that investors concern regarding climate change influences investment decisions and resource allocation. Moreover, the trends in the fintech sector regarding the environmental, sustainable, and governance factors boost the business

performance of financial institutions [11]. As stated in the McKinsey article [6], it is proven that a strong ESG proposition can guarantee long-term success for the company. Therefore, those ESG plans are not only a feel-good exercise but they are important for the growth of the company [3]. Some studies, for example, found that a company's ESG performance is positively related to stock market returns during the financial crisis and, furthermore, they suggest that ESG may play a significant role in company success during the Covid-19 pandemic [12]. It is worth mentioning that in recent years, the obligatory tools have been enhanced based on the issuer's sustainability performance, with characteristics that may change depending on the achievement of specific goals. Green Bonds, for example, are relatively new financial instruments that have experienced extraordinary growth since 2007. They are obligations and their emission is linked to projects that have a positive impact on the environment, such as energy efficiency, renewable energy production, and sustainable land use. Moreover, since 2020, other types of instruments were going to be added. Among those there are the sustainability-linked bonds, which have amassed a total value of 120 million euros as of December 2021, equal to 12% of the annual volume of ESG emissions [1]. Sustainability-Linked Bonds (SLB) are obligations with financial and structural characteristics that vary depending on the achievement of predefined goals related to the issuer's sustainability performance [2]. However, when it comes to making a financial decision, Europe is not in the first place in terms of environmental sustainability. In Indonesia, India, and China, on the other hand, when a financial institution is selected for supporting a new product or service, it is evaluated also in terms of environmental sustainability. Although the limited literature on the nexus between ESG and ICOS, a recent paper by Guzman et al. [5] investigate whether the attention to global warming increases the total funding raised in an environmental ICO by leveraging a set of 324 ICOs and Google trends information. In this paper, we aim at improving such analysis by analysing a wider data set composed of 621 ICOs and extracting specific references to ESG pillars from whitepapers through appropriate deep learning methods.

Our paper therefore puts a special emphasis on whether ESG dimensions influence ICOs performances. Thus, we propose to investigate the role played by a ESG flag covariate, appropriately built as described in the following section, in predicting the probability of success when collecting the expected amount of funds during the funding round. To this end, we use textual analysis techniques for creating a proper sustainability flag variable; afterwords, we fit logistic models with several specifications along the ESG dimensions and controls.

DRAFT

## 2. Data

As database, we consider the Token Offering Research Database by Paul P. Momtaz [10]. Such database contains more than 6,400 ICOs. It comprises several important pieces of information. Besides the available variables, we focus on the whitepapers' links and we download and use them to gather the information we require. By examining each whitepaper, we are able to learn about the company's industry, the number of words and pages, the level of technical and financial expertise and, most importantly, whether it is a company compliant with the ESG principles or carrying a related business idea. Indeed, we analyze each document searching for the ones which are related to sustainability and environment topics, looking both at those having as final purpose the sustainability and/or the environment. After screening for available whitepapers and relative ICOs with no missing data for the other variables of interest, we end up with a database containing 621 ICOs, spanning from 2014 to 2019. Our target variable, similarly to previous literature (for example [9]), is the binary flag of ICOs success/failure, evaluated as the ratio of raised funds and the hard cap, i.e. the maximum amount of funding expected to be raised. If the ratio is above 0.5, we assign success, failure otherwise.

By using the methods previously described, we are able to collect 16 independent variables. The country variable indicates the continent of each ICO divided into America, Africa, Asia, Europe, and Oceania. The accepting variable *CRYPTOACCEPT* indicates what type of cryptocurrency each ICO accepts, i.e. Bitcoin (BTC), Ethereum (ETH) and others. Every ICO is divided into those who admit Ethereum and those who accept Bitcoin and those who accept both. The sector variable shows the type of industry that corresponds to each ICO. We have several sectors: internet publishing and entertainment

providers; IT services and software development; environmental services; healthcare industry and research services; real estate industry; investment and financial services; Civic and Social Organizations; Manufacturing and logistics; consumer services; mining; facilities services. The *ERC20* variable is a dummy one that indicates whether the token offering adheres to the technical ERC20 standard. The technical standard is known as Ethereum Request for Comment 20 (or ERC20), and it specifies a set of rules that a token built on the Ethereum blockchain must follow [8]. So, in other words, ERC-20 establishes a standard for token fungibility. These tokens have a property that makes each token identical (in type and value) to another token. The variable *RATING* indicates the overall project rating on ICObench and it is based on the consensus of industry experts. It ranges from 1 to 5 (”poor quality” to ”good quality”). The *BOUNTY* variable is a dummy one indicating whether the token offering has a bounty program that rewards individuals (usually in the form of free tokens) for marketing activities that promote the offering and the startup. Bounty programs are rewards given to groups of participants for various activities related to an ICO. The *SOCIAL* variable is a numeric one that counts how many social networks (Linkeding, Twitter, Facebook, etc.) the ICO is advertized on. The *PRESALE* variable is a dummy stating whether the ICO had a pre-sale round or not. The *DURATION* variable is a numeric one that reports the number of days the ICO lasted. The *BTCOPENING* variable records the logarithm of the Bitcoin opening price (in USD) at the ICO starting date and the *TOKENPRICE* reports he logarithm of the ICO token price (in USD).

White papers have been fully analyzed through advanced textual analysis techniques based on Bidirectional Encoder Representations from Transformers (BERT) architecture, in order to extract information about the characteristics of the proposed business idea. In particular, we use pretrained model specifically tailored on ESG indicators and financial related vocabularies. The outcome of the model is a probability score for each classification class, e.g. Environmental, Social, Governance, estimating how much pertinent the whitepaper text is. In this way, we do not perform a topic-independent analysis, but we specifically focus and elicit the possible presence of sustainability related keywords. Such step is crucial for building the *ESGFLAG* covariate used in the analysis. Other BERT models are used to extract the continuous Financial Sentiment *FINSENT* (1 for positive, 0 for neutral, -1 for negative) and the level of technical terms *TECHLVL* (from 0 to 1), defined as the frequency of occurrence of tech-related words, e.g. Block-chain, Cloud, etc. Additionally, the length of the whitepaper *PAPERLENGTH* indicates the number of pages associated each paper. The variable *WORDSNUM* indicates the number of words into each paper.

DRAFT

## 3. Methodology and Results

We start the research by analyzing the Token Offering Research Database by Paul P. Momtaz [10]. We apply BERT model to each whitepaper, extracting the ESG topic probability and creating the related flag, as well as all the other text-related variables. At the end of this analysis, 150 ICOs are flagged as ESG-compliant. Due to collinearity, we remove the variable *WORDSNUM* (positively correlated with *PAPERLENGTH*).

We then fit a logit model with OLS estimation, taking into account for year-quarter, country and sector fixed effects, as well as clustering the error by country. We divide our analysis into two scenarios: in the first we only consider the impact of ICOs technical and financial specifications and in the second we include the information extracted from the whitepapers. Both scenarios include the key variable *ESGFLAG*, in order to validate our research hypothesis. Given the imbalance in the target variable, we opt for a weighted logit model, so as to mitigate the impact of the ”failure” class. Table 1 reports the results of both scenarios. Results are stable over the two scenarios. In particular, the type of cryptocurrency and the ERC20 protocol have a positive impact on the success, as BTC and ETH are the most traded cryptocurrencies and the ERC20 guidelines ensure a robust level of security. A positive effect of the presence of a bounty schema and the amount of social platforms is confirmed by the involvement of fundraisers in advertising the ICO. The presence of a pre-sale round and a short duration of the ICO increase the probability of success because they reflect the trust of the investors and the appeal of the product. The price of BTC at the beginning of the ICO and the small price of each sold token still contribute positively

in increasing the likelihood of success, as they reflect the global crypto market trend. In the second scenario, the length of the paper, the positive financial sentiment and the technical level of the content of the whitepaper additionally increase the probability of success. In both scenarios, we observe that the the success of an ICO is promoted when the project shows an interest in ESG topic.

Thus, preliminary results appear to confirm the nexus between ICOs success and ESG. The attention towards sustainability related topics in general seem to favour fund raising activities. This in line with a public audience tendency in evaluating better every activity connected to ethic and responsible behaviour. Such analysis will be further improved and robustified by enlarging the dataset and evaluating more control variables and scenarios.

DRAFT

Table 1: Predicting ICOs succcess with logistic model.

| Variable | Baseline | Whitepaper Content |
|---|---|---|
| CRYPTOACCEPT_BTC | 0.0961** | 0.00705* |
| | (0.0391) | (0.1834) |
| CRYPTOACCEPT_ETH | 0.0379 | 0.0136* |
| | (0.124) | (0.1751) |
| ERC20 | 0.112*** | 0.184*** |
| | (0.0409) | (0.0610) |
| RATING | 0.763*** | 0.544* |
| | (0.169) | (0.279) |
| BOUNTY | 0.0631*** | 0.102*** |
| | (0.0451) | (0.0834) |
| SOCIAL | 0.180** | 0.263* |
| | (0.0910) | (0.147) |
| PRESALE | 0.0759* | 0.102 |
| | (0.0424) | (0.1944) |
| DURATION | -0.244*** | -0.250*** |
| | (0.0279) | (0.0476) |
| BTCOPENING | 0.692 | 0.0721 |
| | (0.190) | (0.1991) |
| TOKENPRICE | -0.153 | -0.234 |
| | (0.29) | (0.198) |
| ESGFLAG | 0.292** | 0.0221*** |
| | (0.049) | (0.0123) |
| PAPERLENGTH | | 0.0143 |
| | | (0.487) |
| FINSENT | | 0.0787*** |
| | | (0.0299) |
| TECHLVL | | 0.131*** |
| | | (0.313) |
| Observations | 621 | 621 |
| Pseudo $R^2$ | 0.53 | 0.51 |
| Year-Quarter effects | Yes | Yes |
| Country effects | Yes | Yes |
| Sector effects | Yes | Yes |
| Clustered Std. Err. | Country | Country |

*Notes:* The table reports coefficients and their standard error (in parentheses). The outcome variable is the binary flag of ICO's success/failure and all variables are defined in Section 2. Data span over the period 2014-2019. Estimation method is OLS with standard errors clustered by ICO's country. The bottom part of the table reports which fixed effects are used in each model specification. First column reports the baseline model, second reports the model that includes variables extracted from whitepapers. The *, ** and *** symbols denote the p-values at $10^{th}$, $5^{th}$ and $1^{st}$ significance level, respectively.

DRAFT

## References


[1] Annunziata, Rita. (2022, 29 April). Sustainability-linked bond: cos'è e perché piace ai mercati. We Wealth.

[2] Antilici Paola, Mosconi, Gianluca, Russo, Luigi. (2022, 27 April). Quando innovazione finanziaria e finanza sostenibile si incontrano: i Sustainability-Linked Bonds. Numero 22. Banca d'Italia. ISSN 2724-6418.

[3] Feldman G. Putting Uncle Milton to Bed: Reexamining Milton Friedman's Essay on the Social Responsibility of Business. Labor Studies Journal. 2007;32(2):125-141.

[4] Guzmán, A.; Pinto-Gutiérrez, C.; Trujillo, M.-A. Signaling Value through Gender Diversity: Evidence from Initial Coin Offerings. Sustainability 2021, 13, 700.

[5] Guzmán, A.; Pinto-Gutiérrez, C.; Trujillo, M.-A. Attention to Global Warming and the Success of Environmental Initial Coin Offerings: Empirical Evidence. Sustainability 2020, 12(23), 9885

[6] Henisz, Witold, Koller, Tim, Nuttal, Robin. (2019, November). Five ways that ESG creates value. Mckinsey.

[7] Hoffman, Andrew John, (2018, 1 January), The Next Phase of Business Sustainability, Stanford Social Innovation Review, 16(2): 34-39., Ross School of Business Paper No. 1381.

[8] Mansouri, Sasan and Momtaz, Paul P., (2022, February 13). Financing Sustainable Entrepreneurship: ESG Measurement, Valuation, and Performance.

[9] Meoli Michele, Vismara Silvio, Machine-learning forecasting of successful ICOs, Journal of Economics and Business, 2022, vol. 121, issue C

[10] Momtaz, P.P. (2021) Token Offerings Research Database (TORD)

[11] Nizam, Esma, Ng, Adam, Dewandaru, Ginanjar, Nagayev, Ruslan, Nkoba, Malik Abdulrahman, 2019. "The impact of social and environmental sustainability on financial performance: A global analysis of the banking sector," Journal of Multinational Financial Management, Elsevier, vol. 49(C), pages 35-53.

[12] Reijonen, J. (2021). The importance of ESG factors for company performance during Covid-19 pandemic.


DRAFT

# On the impact of intraclass correlation in the ANVUR evaluation of academic departments

Giorgio E. Montanari[a] and Marco Doretti[a]

[a]University of Perugia, Department of Political Science, via Pascoli 20, 06123 Perugia (Italy);
giorgio.montanari@unipg.it, marco.doretti@unipg.it

**Abstract**

In Italy, academic departments are ranked according to a function of the scores assigned by ANVUR (an appointed national agency) to some of their scientific products. We show how intra-department correlation among these scores affects the overall validity of currently-in-use procedures. For this correlation, we also outline an identification and estimation strategy based on the available data.

***Keywords:*** department ranking, performance index, research output evaluation

DRAFT

## 1. Introduction

The Italian Agency for the Evaluation of Universities and Research Institutes (Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca - ANVUR) is a public institution monitored by the Ministry of University and Research. In accordance with the Italian law, one of its tasks consists in periodically producing a ranking of academic departments reflecting their overall scientific quality level. Such a ranking serves as a tool to identify a restricted set of top departments (departments of excellence), which can compete to gain access to dedicated funds. The ranking methodology is based on a standardized departmental performance index (Indice Standardizzato di Performance Dipartimentale - ISPD). For each department, ISPD is computed as a function of scores assigned to a selection of scientific products (papers) written by its scholars. Specifically, the scores of another periodic department-level evaluation process (Valutazione della Qualità della Ricerca - VQR, also conducted by ANVUR) are taken [1].

Recently, a noticeable debate around the overall soundness of the ANVUR ranking methodology has risen [2]. The main critique lies in the fact that in the 2017 and 2022 rankings (based on the 2011-2014 and 2015-2019 VQR processes, respectively), the number of departments with extremely high and low ISPD values is sensibly higher than expected. Possible causes, however, have not been deeply investigated so far. In this paper, we address this problem by considering the most likely (in our view) reason of this discrepancy: the presence of some degree of intra-department correlation among the product scores.

## 2. The ANVUR standardized departmental performance index

For the 2015-2019 VQR process, Italian academic departments have been asked to submit a number of scientific products equal to three times the number of staff members, with sporadic exceptions. Each product belongs to a scientific field (Settore Scientifico Disciplinare - SSD) and, after the evaluation process, it receives a VQR score equal to 0, 0.2, 0.5, 0.8 or 1, with higher scores denoting higher quality

levels. To remove heterogeneity in evaluation standards among SSDs, product scores are then standardized with the mean and standard deviation of all scores associated to products of the same SSD. Thus, for each SSD, standardized scores have null average and standard deviation equal to 1.

Let $N_d$ be the number of evaluated scientific products for the $d$-th department ($d = 1, \dots, D$). Further, denote by $Z_{dji}$ the standardized score for the $i$-th product of the $j$-th SSD, with $i = 1, \dots, N_{dj}$ and $j \in \text{SSD}_d$ (the set of SSDs in the $d$-th department), where $N_d = \sum_{j \in \text{SSD}_d} N_{dj}$. The ANVUR ranking procedure computes, for each department, the scaled average

$$Z_d \equiv \sqrt{N_d}\bar{Z}_d = \frac{\sum_{j \in \text{SSD}_d} \sum_{i=1,\dots,N_{dj}} Z_{dji}}{\sqrt{N_d}}$$

($\bar{Z}_d$ being the sample average) and defines the standardized departmental performance index as

$$\text{ISPD}_d = 100 \cdot \Phi(Z_d),$$

where $\Phi(\cdot)$ denotes the Cumulative Density Function (CDF) of the standard normal distribution (note that published ISPD values are rounded to the nearest semi-integer). Clearly, this index ranges from 0 to 100, with higher values denoting better performances. Specifically, $\text{ISPD}_d$ is interpreted as the percentage probability that the product set of a virtual department with the same SSD structure as $d$ - formed by replacing each product of the department $d$ with one selected at random from those of the SSD it belongs to - realizes a scaled average score not greater than that of $d$. Indeed, under such a random process, $Z_{dji}$ is a standardized random variable. Thus, assuming the absence of correlation among the $Z_{dji}$ variables, each department's scaled average $Z_d$ is approximately distributed as a standard normal random variable by virtue of central limit theorem. Since $N_d$ is typically greater than 60 (with a few exceptions), the approximation is reliable.

The ISPD is used to rank departments in the same way as an order statistic since, under the above assumptions, it has a uniform distribution on the 0-100 interval. Specifically, denoting by $F(x)$ the CDF of $\text{ISPD}_d$ evaluated at $x$, we have

$$\begin{aligned} F(x) &= P(\text{ISPD}_d \leq x) = P(100 \cdot \Phi(Z_d) \leq x) = P(Z_d \leq \Phi^{-1}(x/100)) \\ &= \Phi(\Phi^{-1}(x/100)) = x/100. \end{aligned}$$

## 3. The impact of intra-department correlation

The approach described in Section 2. hinges on the assumption that the $Z_{dji}$ variables are uncorrelated. However, such an assumption might be unrealistic due to the fact that scholars submit more than one product to the VQR procedure, as well as to departments' tendency to aggregate scholars with similar scientific quality levels, especially within SSDs. This generates a non-null intra-department correlation among the $Z_{dji}$ variables, that we model as $\rho_d = \text{Cor}(Z_{dji}, Z_{dj'i'})$ for all $i = 1, \dots, N_{dj}$, $i' = 1, \dots, N_{dj'}$ and $j, j' \in \text{SSD}_d$ such that $ji \neq j'i'$. Hence, the variance of the scaled department average modifies to

$$V(Z_d) = 1 + (N_d - 1)\rho_d \equiv \sigma_d^2, \tag{1}$$

becoming department-specific. Since it is reasonable to assume that $\rho_d$ is positive, we have $\sigma_d^2 > 1$, meaning that the distribution of each $Z_d$ has heavier tails than the standard normal.

In this setting, the comparability of the $\text{ISPD}_d$ values - if computed without a proper modification - is in doubt. To acknowledge this fact, suppose that two departments with the same number of products, say 150, but different intra-department correlations, say 0.02 and 0.05, realize the same scaled average, say 2.0. For both departments, ISPD is equal to 97.7, but the true percentage probabilities of performing better than the corresponding random virtual departments are 84.2% and 75.4%. Conversely, if the common scaled average were equal to -2.0 (*ceteris paribus*), the two true percentage probabilities would be 15.8% and 24.7%, whereas the ISPD value would be 2.3 for both departments. Clearly, analogous discrepancies would arise for departments with the same scaled average and correlation, but with different number of products. Hence, the comparison is no more unaffected by the size of the departments, as it would be if all the $\rho_d$ values were equal to zero.

DRAFT

## 4. Identification and estimation

The heterogeneity of the $\sigma_d^2$ variances implies that the observed department scaled averages, and ISPD values in turn, cannot be treated like realizations from the same distribution as in Section 2. In fact, the former can be thought of as realizations of a stratified sample, with one observation per stratum (i.e., department). The underlying observed random variable, denoted by $S$, is characterized by the mixture density

$$f_S(s) = \frac{1}{D}\sum_{d=1}^{D} \phi(s; 0, \sigma_d^2), \tag{2}$$

with $\phi(x; 0, \sigma^2)$ representing the probability density function at $x$ of a $N(0, \sigma^2)$ variate. Straightforward calculations show that $E(S) = 0$ and that

$$V(S) = \frac{1}{D}\sum_{d=1}^{D} \sigma_d^2 = 1 + \frac{1}{D}\sum_{d=1}^{D}(N_d - 1)\rho_d \equiv \sigma^2, \tag{3}$$

which is greater than 1 since all the $\sigma_d^2$ terms also are; see Section 3.

In this framework, ISPD values can be thought of as independent realizations of $100 \cdot \Phi(S)$. Thus, under the assumption that the distribution of $S$ is approximately normal, the fact that $\sigma^2 > 1$ provides an immediate explanation as to why a surplus of departments with low and high values generates. In practice, the empirical histogram of the ISPD values, rounded to the nearest semi-integer, will approximately assume a "bath-tub shape" (or U-shape), with two peaks in correspondence of extreme values 0 and 100 as well as densities symmetrically increasing as the index moves away from the center of the distribution. Such a trend is recognizable in the right tail of both the 2017 and 2022 ISPD distributions (the only portions of data made available by ANVUR), while an U-shaped histogram appeared in an independent analysis performed on the whole 2017 data [2]. For the latter, it has been reported that 119 out of 766 departments (15.54%) had ISPD $< 0.5$, and exactly the same number had ISPD $> 99.5$ [2].

The $\sigma^2$ parameter can be identified from the ISPD distribution in a number of ways. The simplest approach relies on the aforementioned assumption that $S \approx \sigma Z$, with $Z$ being a standard normal variate. Thus, letting for any $\alpha \in (0, 0.5)$

$$P_\alpha = P(\text{ISPD}_d \leq 100 \cdot (\alpha/2) \ \cup \ \text{ISPD}_d \geq 100 \cdot (1 - \alpha/2)),$$

simple algebra shows that $P_\alpha \approx 2\big(1 - \Phi(Z_{\alpha/2}/\sigma)\big)$ and, consequently,

$$\sigma \approx \frac{Z_{\alpha/2}}{\Phi^{-1}(1 - 0.5P_\alpha)}, \tag{4}$$

where $Z_{\alpha/2}$ is the standard normal distribution quantile leaving $\alpha/2$ probability at its right. Clearly, the goodness of the approximations in the above formulas depends on how close to normality $S$ is. Hence, for a given $\alpha$, estimation of $\sigma$ can be performed via (4) after estimating $P_\alpha$. In this respect, for the 2017 ANVUR data the estimated proportion $\hat{P}_{0.01} = 2 \cdot 119/766 = 31.07\%$ has been reported [2], leading to $\hat{\sigma} \approx 2.541$.

Unfortunately, this strategy for the (approximate) identification of $\sigma$ does not enable identification of all department-specific correlations $\rho_d$, unless additional assumptions are made. For example, assuming $\rho_d = \rho$ in (3) returns

$$\hat{\rho} = \frac{\hat{\sigma}^2 - 1}{\bar{N} - 1}, \tag{5}$$

where $\bar{N} = D^{-1}\sum_d N_d$. More generally, assuming that $\rho_d$ and $N_d$ are (statistically) uncorrelated, it is possibile to identify the average correlation $\bar{\rho} = D^{-1}\sum_d \rho_d$, whose estimate is still given by the right-hand side of (5). For the 2017 ANVUR data, we have $\hat{\rho} = \hat{\bar{\rho}} \approx 0.042$, where $\bar{N}$ is computed starting from information in [2] and considering that in the 2011-2014 VQR process each scholar had to submit two products (again, with rare exceptions).

DRAFT

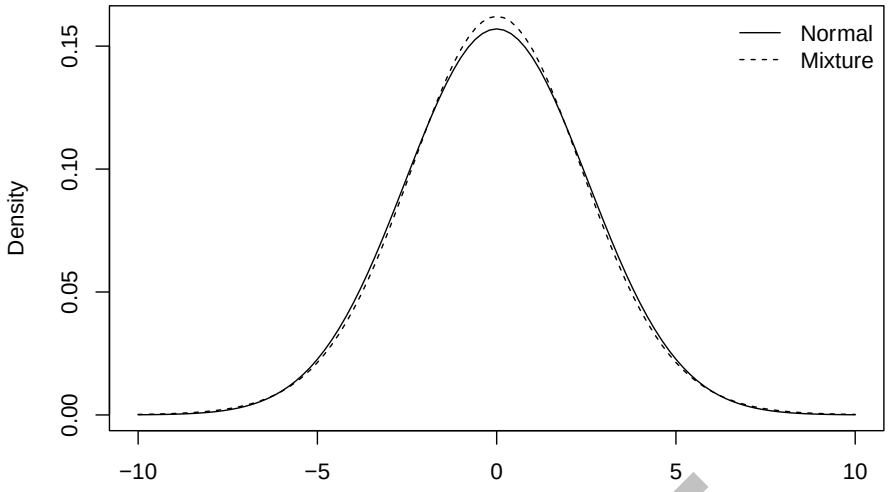



Figure 1: 2017 ANVUR data: density of the $N(0, \hat{\sigma}^2)$ distribution (undashed) and of the mixture of equally-weighted centered normal distributions with variances $\hat{\sigma}_d^2$ (dashed, $d = 1, \ldots, D$).

When the correlation level is assumed constant across departments, it is also possibile to identify and estimate all the $\sigma_d^2$ parameters by replacing $\rho_d$ with $\hat{\rho}$ in (1). Under this approach, for the 2017 ANVUR data the estimated department-specific variances $\hat{\sigma}_d^2$ range from 1.969 to 20.499, with average $\hat{\sigma}^2 = 6.456$ and standard deviation equal to 2.090. These estimates have been plugged-in in (2) to build the estimated mixture density $\hat{f}_S(\cdot)$, which is plotted in Figure 1 together with the $N(0, \hat{\sigma}^2)$ density. As shown by the graph, the two functions are quite close, which makes the $S \approx \sigma Z$ assumption reasonable.

DRAFT

## 5. Further remarks and possible extensions

In this paper, we have motivated why ISPD values are highly concentrated in the distributions tails, with almost one third of Italian academic departments assuming the extreme values (0 or 100) in the 2017 ranking (in 2022, this phenomenon is even more pronounced). This owes to the fact that the variances of scaled averages are greater than 1, due to the correlation among the VQR scores of scientific products in the same department. Importantly, departments with (rounded) ISPD equal to 100 (as well as to 0) might severely differ in their scaled averages and, thus, in their performance. This may be unveiled by computing ISPDs with the CDF of a normal distribution with null expectation and variance equal to that of the observed scaled averages, $\hat{\sigma}^2$.

Generally speaking, the intra-department correlation is department-specific. Even if it were the same in all departments, the variances of scaled averages would not be constant because of the different size of the departments. Thus, scaled averages are not comparable, since they are drawn from distributions with different variability, and for this reason the ISPD-based department ranking might be misleading.

Access to the VQR scores assigned to each scientific product would enable direct identification and estimation of department-specific correlations $\rho_d$ and, in turn, of variances $\sigma_d^2$ via (1). In this framework, the $Z_d/\hat{\sigma}_d$ ratio could be approximately thought of as a draw from a standard normal distribution, although some adjustments are needed because variance estimation has to be properly accounted for. The adoption of an index given by $100 \cdot \Phi(Z_d/\hat{\sigma}_d)$ would be the first step toward a fairer and more appropriate performance measure.

Future research could also deal with refinement of the $\rho_d$ modeling, via the decomposition of this correlation level into a number of sub-components. Such an approach would allow to account more explicitly for a number of additional features as, for example, the fact that multiple scientific products belong to the same person, and that each department's scholars are further grouped into SSDs. However, sample size issues are likely to arise if $N_d$ is not big enough.

## References


[1] ANVUR (2022). Nota metodologica sul calcolo dell'indicatore ISPD. Rapporto tecnico (link).

[2] Redazione ROARS (2022). VQR: la lista segreta dei 120 dipartimenti con zero in pagella (link).


DRAFT

# Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes

Luigi Biggeri[a], Stefano Marchetti[b], Caterina Giusti[b], Monica Pratesi[c], Francesco Schirripa Spagnolo[b], and Gaia Bertarelli[d]

[a]University of Florence-Dagum ASESD Centre; luigi.biggeri@unifi.it
[b]University of Pisa-Dagum ASESD Centre; stefano.marchetti@unipi.it, caterina.giusti@unipi.it, francesco.schirripa@unipi.it
[c]Istat-University of Pisa-Dagum ASESD Centre; monica.pratesi@istat.it
[d]University of Venice-Dagum ASESD Centre; gaia.bertarelli@unive.it

DRAFT

**Abstract**

The aim of this work is to estimate monetary poverty indicators at provincial level in Italy taking into account the different price levels within the country. To account for the local price levels, we compute Spatial Price Indexes (SPIs) using retail scanner data on retail prices. The SPIs are used to adjust the poverty line when computing poverty indicators at provincial level that are estimated using Small Area Estimation (SAE) models.

***Keywords:*** Poverty mapping, Spatial price indexes, Scanner data

## 1. Introduction

Poverty measures at sub-national and local level play an important role in setting policy actions against poverty and social exclusion.

Particularly, the local poverty indicators are relevant both for a detailed planning of the policies actions and for the citizens to evaluate their effect. However, there are still open problems to compute adequate sub-national poverty indicators. One of the main issue in estimating these indicators at sub-national level (such as Italian provinces) is the need to account for the cost-of-living differences in the comparison of poverty between different territorial areas. To assure that the poverty line(s) represent approximately the same standard of living across the different areas, Spatial Price Indexes (SPIs) are used to adjust the national poverty line at the local level. Spatial price indexes are measures of differences in price levels across areas, essential to compare economic well-being indicators. However, at the sub-national and local level, direct estimates of poverty measures are not accurate because sample surveys on income or consumption expenditures are usually designed so that direct estimators lead to reliable estimates only for larger domains (states, regions). Small area estimation (SAE) comprises the methods for obtaining more precise estimators in local areas, making use of the common features of the areas. A wide range of methods have been proposed and used in literature to obtain reliable small-area estimates (mostly model-based estimators (4)).

The costs involved for carrying out ad-hoc surveys for collecting price data and the labour-intensive analyses necessary for processing traditional price data have limited attempts to regularly produce SN-SCPIs (Sub-National Spatial Consumer Prices Indices). Because of this, using big data such as business transaction data can be a suitable solution to the challenges that National Statistical Institutes (NSIs)

encounter in comparing consumer prices at different territorial levels (2). Since 2018, Istat has included scanner data of grocery products (excluding fresh food) provided by the market research company AC-Nielsen in the production process of the consumer price indices. Due to the high coverage of transactions and the high territorial coverage, scanner data represent an useful source in order to compare price levels among geographical areas within a country at a very detailed territorial level (such as provinces).

Therefore, in this work we use scanner data to compute Spatial Price Indexes (SPIs) at provincial level in Italy. Then the poverty incidence at provincial level is estimated by using different poverty lines adjusted using the SPIs. In this way, the poverty estimates consider the different purchase power within the country. The scanner database referred to the years 2017 and 2018 were provided by an agreement between Istat and ASESD Dagum Centre signed to implement the tasks of the MAKSWELL project (www.makswell.eu) (for more details please refer to the work by Pratesi, Giusti, Marchetti, Biggeri, Bertarelli, Schirripa Spagnolo, Laureti, Benedetti, Polidoro, Di Leo, Fedeli of the Makswell "Deliverable 3.2 - Guidelines for best practices implementation for transferring methodology").

## 2. Estimation of Spatial Consumer Price Indexes for the Italian Provinces

Spatial Consumer Price Indexes for the Italian Provinces needs to be estimated to adjust the national poverty line and to take into account the different price levels within the country. To estimate the SN-SCPs for each of the 103 (out of 110) Italian provinces, two-step procedure has been followed, adapting the approach of (7).

In the first step, we computed the average unit price at provincial level, by considering the unit value prices from the consumer side. In applying the principle of comparability, we did not follow a very tight way by considering the comparisons of the 'like to like' items (products). Instead, we applied the principle at a different level, the level products' groups, and exactly at the level of the 102 groups of the European Classification of Individual Consumption according to Purpose (ECOICOP) 8-digit classification.

Define the weighted mean price $\bar{p}_{ij}$ for ECOICOP-8-digit $j$ and province $i$. Let $r_{ijk}$ and $q_{ijk}$ be the annual turnover and the total quantity sold[1] respectively of item $k$ belonging to ECOICOP-8-digit $j$ in province $i$. These quantities are estimated by Istat using the scanner data of the products sold in modern distribution chains referring to the year 2017 and the sampling weights computed according to the survey design (we refer to Deliverable 3.2 of the MAKSWELL project for further details). In order to obtain, for each item the annual price per gr. or ml., the turnover per item $(r_{ij}/q_{ij})$ is divided by the quantity of the item $ijk$ in terms of gr. or ml. $u_{ijk}$:

$$p_{ijk} = \frac{\frac{r_{ijk}}{q_{ijk}}}{u_{ijk}}.$$

Then, for each item we define its relative weights in term of turnover as

$$w_{ijk} = \frac{r_{ijk}}{\sum_{k=1}^{n_{ij}} r_{ijk}},$$

where $n_j$ is the number of items in the $j$th ECOICOP-8-digit aggregation and the $i$th province. Finally, the weighted mean price is:

$$\bar{p}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} p_{ijk} w_{ijk}.$$

Therefore, $\bar{p}_{ij}$ is the weighted mean price per gr. or ml. for products in ECOICOP-8-digit $j$ and province $i$.

The second step is devoted to the aggregation of 102 average level of prices to estimate the provincial SPI. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces.

[1] Which are the expenditure and the quantity purchased by consumers.

DRAFT

To compute the SPIs at provincial level we adapt a Country Product Dummy (CPD) model (3). The products are aggregated by province and ECOICOP-8-digit classification, for a total of 103 provinces and 102 ECOICOP-8-digit. Note that not all the ECOICOP-8-digit aggregates are present in all the provinces. The CPD model we propose is as follows:

$$\log \bar{p}_{ij} = \alpha_0 + \alpha_i D_i + \beta_j I_j + \varepsilon_{ij}, \quad i = 1, \ldots, 103 \quad j = 1, \ldots, 102, \tag{1}$$

where $D_i$ is a vector equal 1 if the mean price is in province $i$ and 0 otherwise, $I_j$ is equal 1 if the mean price belongs to $j$th ECOICOP-8-digits and 0 otherwise. The index $i$ is for the provinces and the index $j$ is for the ECOICOP-8-digit. The error $\varepsilon_{ij} \sim N(0, \sigma^2)$.

To take into account the different level of the turnover between the ECOICOP-8-digit aggregates we estimate the model (1) using weighted least squares, where the weights are computed as

$$wls_{ij} = \frac{\sum_{k=1}^{n_{ij}} r_{ijk}}{\sum_{k=1}^{n_i} r_{ijk}},$$

the ratio between the total turnover of one aggregate in one province and the total turnover in the province ($n_i$ is the number of items in the $i$th province).

Model (1) – as it is specified – is not identified, because the $D_i$s vectors are a linear combination of the constant. Therefore, we impose the constraint $\alpha_1 = 0$ so that the model is identified. Once the model is estimated, from the data we obtain the estimates of the SPIs at provincial level by $\exp(\hat{\alpha}_i)$, where $\hat{\alpha}_i$ is the estimate of $\alpha_i$. The coefficient $\alpha_i$ is the difference of fixed effects connected with the province $i$ compared with the base province $i = 1$. To use as a reference Italy instead of area 1, the coefficients $\hat{\alpha}_i$ has been adjusted following (6).

The aim of this work is to use the estimated SPIs to adjust the national poverty line at the province level. The SPIs estimated according to model (1) are based on mean prices of specific headings (ECOICOP-8-digit), therefore the adjustment of the national poverty line is not poor specific. As an alternative, the method can be easily extended to produce SPIs related to the first quintile of the distribution of the price of each specific product, assuming that poor purchase the cheaper items of the product. For example, figure 1 reports two choropleth maps of estimated SPIs based on model (1, left) and on an adjusted the model that considers the quantile 0.2 of the unit prices to obtain the estimates of spatial price indices related to the cheaper prices for each Italian provinces, which we denote as SPI($Q_{0.2}$)'s (right).

The results we obtained are somehow expected. Indeed, provinces in the south of Italy show SPIs smaller than 1, while provinces in the north show values greater than 1. However, there are exceptions, provinces in the north-east Alps mountains show SPIs below 1, even if they are close, both considering the mean and the quantile 0.2 of unit prices.

DRAFT

## 3. The impact of the local cost-of-living differences on the measure of the poverty incidence

To evaluate the poverty incidence at provincial level in Italy we estimate the Head Count Ratio using Household Expenditure Survey (HES) data in Italy. To take into account the different price levels, the national poverty line is adjusted for each province using the SPI($Q_{0.2}$) values opportunely weighted (adapting the idea in (5)):

$$nPL_i^* = nPL \times (\lambda_i SPI_i + 1 - \lambda_i) \tag{2}$$

where $nPL$ is the national poverty line, $nPL_i^*$ is the adjusted poverty line for province $i$, $\lambda_i$ is the estimated share of food consumption in province $i$ and $SPI_i$ is the SPI($Q_{0.2}$) for province $i$. The quantities $\lambda_i$'s are estimated from the HES 2017 as the provincial mean of the ratios between the food expenditure and the total consumption expenditure:

$$\lambda_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} \frac{p_{ij}}{t_{ij}} w_{ij}, \tag{3}$$

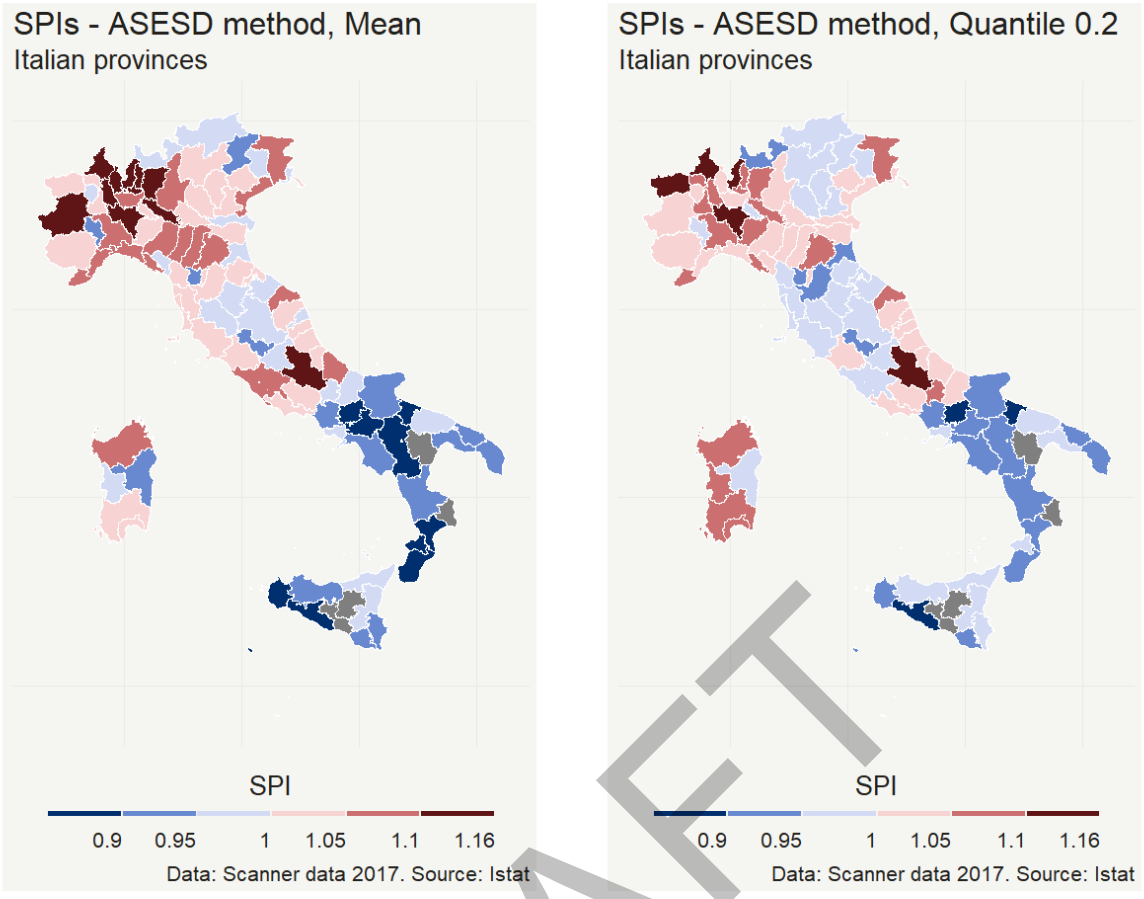



Figure 1: Choropleth map of SPIs obtained using mean unit prices (left) and quantile 0.2 of unit prices (right).

DRAFT

where $n_i$ is the sample size in province $i$, $w_{ij}$ is the survey weight of household $j$ in area $i$, $p_{ij}$ is the food expenditure of household $j$ in area $i$ and $t_{ij}$ is the total consumption expenditure of household $j$ in area $i$. The survey weights have been calibrated to sum to the total households at provincial level. Although the $\lambda_i$'s are estimated at the provincial level – thus possibly unreliable because of small sample size – we judge the direct estimates suitable for our purpose.

Having computed the adjusted nPLs, we then calculated the corresponding direct estimates of the poverty rates. As the variability of the direct estimates was too high (approximately half of the provinces a CV greater than 30%) we estimated a Fay-Herriot (FH) model (1) with the following auxiliary variables: the ratio between number of taxed persons over the population, and the ratios between the number of persons with $i$. income coming from salary, $ii$. income coming from pensions and $iii$. income lower than 10,000 euros per year, over the number of taxed persons. These data come from the Italian tax agency database 2017.

The Empirical Best Linear Unbiased Predictor (EBLUPs) obtained with the FH model showed a gain in efficiency with respect to direct estimates. We obtained a CV smaller than 16% in 37 provinces, while half of the provinces had a CV smaller than 20%.

Figure 2 maps the Head Count Ratio (HCR) computed using the price-adjusted poverty lines referring to the Italian provinces[2]. As we can see, the results confirm the well-known north/south divide, with HCR values that are generally higher in the south of the country, lower in the north.

We also computed the EBLUPs without any adjustment of the national poverty line, using the same small area model as for adjusted EBLUPs. Figure 3 reports the comparison of the two set of EBLUPs estimates: as we can see, using the SPI($Q_{0.2}$) to adjust the poverty lines, the HCRs in northern and central

[2]The HCR value has not been estimated for out of sample provinces in the HES data. Specific predictions could be made for this provinces.

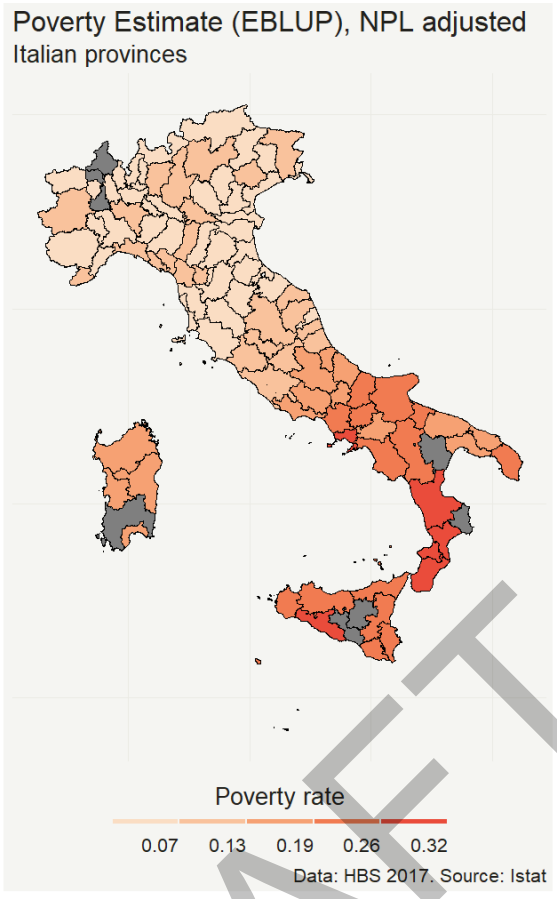



Figure 2: Poverty rate at provincial level in Italy: EBLUPs computed using the adjusted national poverty line.

provinces slightly decrease.

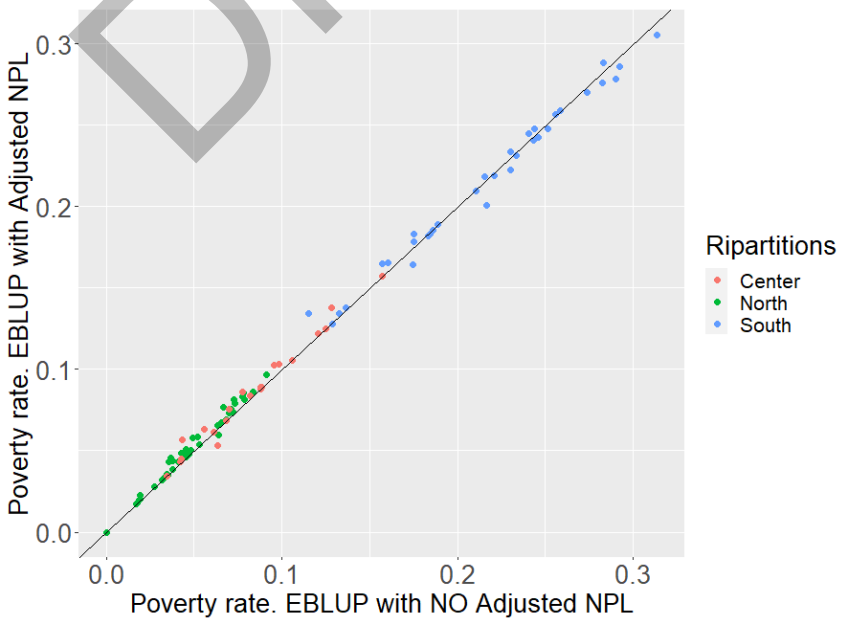



Figure 3: Poverty rate at provincial level in Italy: provincial EBLUPs estimates using the SPI($Q_{0.2}$) adjusted vs not adjusted national poverty line.

DRAFT

The results obtained suggest that the methodology can be extended to include other Spatial Price Indexes, therefore adjusting the national poverty line with other components of households' consumption expenditure. Indeed, our results suggest the products included in the scanner data represent a relevant but still limited share of the total household consumption expenditure, approximately equal to the 20%. Therefore, by including other consumption expenditure components, such as for example the expenditure for the rent, the national poverty line could be adjusted in a more complete manner.

## References


[1] Fay III, R. E., Herriot, R. A.: Estimates of income for small places: an application of James-Stein procedures to census data. J. Am. Stat. Assoc. **74**. 269–277 (1979).

[2] Laureti, T., Polidoro, F.: Using scanner data for computing consumer spatial price indexes at regional level: An empirical application for grocery products in Italy. J. Off. Stat., **38**, 23–56 (2022).

[3] Laureti, T., Prasada Rao, D.S. : Measuring spatial price level differences within a country: Current status and future developments. Estud. Econ. Apl., **36**, 119–148 (2018)

[4] Pratesi, M., ed: Analysis of Poverty Data by Small Area Estimation. John Wiley & Sons (2016).

[5] Renwick, T., Aten, B., Figueroa, E., Martin, T.: Supplemental poverty measure: A comparison of geographic adjustments with regional price parities vs. median rents from the american community survey. Technical report, Bureau of Economic Analysis (2014).

[6] Suits, D.: Dummy variables: Mechanics v. interpretation. Rev. Econ. Stat., **66**, 177–180 (1984).

[7] World Bank Group: Operational Guidelines and Procedures for Measuring the Real Size of the World Economy. 2011 International Comparison Program, Washington, DC (2015).


DRAFT

# Smart Composite Indicators Measuring Corporate Sustainability: A Sensitivity Analysis

Camilla Salvatore[a], Annamaria Bianchi[b], and Silvia Biffignandi[c]

[a] Utrecht University; c.salvatore@uu.nl
[b] University of Bergamo; annamaria.bianchi@unibg.it
[c] CESS; silvia.biffignandi@outlook.com

**Abstract**

Augmenting traditional data with digital trace data is an emerging trend in statistics, as it offers the possibility to explore novel aspects not covered by traditional data sources. This study focuses on corporate sustainability and proposes to augment traditional data (from a commercial database) with social media data under a composite indicators perspective. We present and apply a modular framework for data augmentation, by building on previous research. The innovative aspect of this paper is the evaluation of the quality of the resulting indicator through a sensitivity analysis. Integrating data from different sources requires more attention than classical sensitivity analysis, where different aggregation strategies and weighting procedures are compared. It is also necessary to consider the unique nature of the innovative data source. In our study, we also test the stability of the results to different data selection strategies, data cleaning, and analytical choices for extracting relevant information from the text. We conclude with recommendations for researchers interested in augmenting traditional data with digital trace data.

**Keywords:** Data Augmentation, Corporate Sustainability, Social Media

DRAFT

## 1. Introduction

Measuring sustainability has become increasingly important and, in recent years, also businesses and organizations have recognized the importance of prioritizing environmental and socially responsible behaviours [7]. Measuring and evaluating the commitment of businesses towards sustainability is important to policy makers and researchers, especially with respect to the goals of the Agenda 2030. With the availability of innovative data, such as big or digital trace data, researchers can explore new aspects of sustainability phenomena, which are not covered by traditional data sources. By combining traditional and innovative data, it is possible to gain a more complete understanding of sustainability and identify areas where action is needed.

At the same time, informed decision making relies on good quality data. Traditionally, such data are represented by survey and also administrative/commercial database data. Surveys are considered the gold standard: structured, checked for quality and with well-established statistical error framework [2]. Some examples are the European Company Surveys, the Business and Consumer Surveys (BCS) and other surveys carried out by National Statistical Institutes. Alongside surveys, other popular sources for business statistics are administrative or commercial business data [6]. These are still structured data, quality is checked and improved when necessary. However, these data are not primarily collected for statistical or research purposes. For that reason, they are usually referred to as secondary data. Business registers, documents from local authorities (e.g., tax authority), and law-mandatory reporting are all examples of administrative data. Commercial business data are provided by private companies, for example, Bureau van Dijk, Bloomberg, and Refinitiv.

More recently, the digital transformation has resulted in the emergence of new sources for business and economic statistics [3]. For example, social media posts, annual reports, businesses websites and newspaper articles can be used to study new aspects or getting additional information about companies.

In this respect, the production of statistics using traditional data enhanced with new data available from online sources are referred to as smart statistics. One of the advantages of smart statistics is the ability to augment the information, thereby providing richer insights to the topic of interest.

This paper focuses on the concept of corporate sustainability, and extends on a previous study where a methodology for constructing smart composite indicators by augmenting traditional data with digital trace data has been proposed [15]. In this study quality aspects are addressed. Corporate sustainability is closely related to the notion of Corporate Social Responsibility (CSR), which involves implementing activities aimed at improving firms' reputation and positively impacting society [5]. Social media data present an opportunity for researchers and policy makers to monitor business behaviours concerning sustainable development and the Agenda 2030 by investigating online communication about CSR activities.

In Section 2, we summarize the modular framework proposed in previous papers, as well as the data used. However, the quality of the resulting indicator has not been extensively studied and evaluated before. Thus, we aim to address this topic in the current paper. Section 3 evaluates the stability of results through a sensitivity analysis, considering not only classical aspects like aggregation strategy and weighting, but also novel aspects related to the innovative data source, such as data selection, data cleaning, and analytical choices for extracting relevant information from the text. In Section 4, we conclude with recommendations for researchers interested in building composite indicators using both traditional and digital trace data.

## 2. The Case Study

As part of an ongoing research, this work is based on two papers where a smart composite indicator measuring corporate sustainability (SMART-INDEX) is constructed following the original modular framework proposed by Bianchi et al. [4] and Salvatore et al. [15]. It is based on the modular organization into three layers introduced by Ricciato et al. [10]. We briefly present the results of a previous work and then we move to the original contribution of this paper, the sensitivity analysis. More details about the theory of composite indicators can be found in [9,8].

The SMART-INDEX is derived from the combination of a traditional index (TRAD-INDEX), already available in a commercial database, and an original social media-based indicator (SM-INDEX). In the following paragraph we explain how to derive such indicator using the modular framework. Figure 1 presents the adaptation of the general framework to our case study.

The first layer involves digital trace data collection and their transformation into structured data. To the purpose of our analysis and as part of an ongoing research project, we consider the same data used in Salvatore et al. [14,15]. They refer to the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies as of the composition in August 2020. Following the tasks in the first layer, we identified and retrieved the data form the official Twitter accounts of the companies. Given that companies may have several Twitter accounts, we focused primarily on CSR accounts and, in case these are not available, on the news or multipurpose ones. The objective is to reduce the noise (no-CSR tweets) in the data. The Tweets refers to the 2019 year and the total number of messages retrieved is 25,148. In order to extract the relevant information from the text, the Structural Topic Model ([11]) has been applied and topics have been assigned to CSR dimensions: social, economic, environmental, and mixed. For more information about the data and the topic model methodology, please refer to Salvatore et al. [14,15]. The output of the first layer is a set of elementary indicators which are the base for the construction of the SM-INDEX.

From a theoretical point of view, social media communication differs in content (the topic discussed) and modality (the way it is conducted). Thus, we consider two dimensions to build the SM-based indicator. The first one refers to the communication content in tweets, i.e., to the text which refers to the communication of CSR activities in one of its dimensions, economic, social, environmental, and general (or mixed). The second one refers to communication modality (media richness from tweets metadata; the higher the media richness, the more effective the communication will be [1]).

The second layer involves the practical construction of the SM-INDEX (Figure 2). The composite indicator for the content dimension is constructed by considering the output of the topic model as its elementary indicators. Specifically, the proportion of text devoted to each CSR dimension for each tweet is used. We assume that these proportions are substitutes (compensatory aggregation) with the same importance (no weight). To obtain the composite indicator, we take the sum of these proportions at the

DRAFT

tweet level and then aggregate them at the firm level by taking the arithmetic mean (first innovative indicator).

The composite indicator for the modality dimension is based on tweets' metadata. Similar to the content dimension, we consider elementary indicators to be substitutes (compensatory aggregation) with the same importance (no weight). The elementary indicators used are the presence of hashtags, mentions, media, and links (binary variables). For each tweet, we sum these individual indicators, obtaining a score between 0 and 4. We then aggregate these scores at the firm level by computing the arithmetic mean (second innovative indicator).

Once the modality and the content indexes are constructed, it is necessary to combine them to obtain the SM-INDEX. In this case we propose to apply the Mazziotta-Pareto index (MPI) [8] which is partially compensatory recognizing that the two dimensions are equally important but partially substitute to gain efficiency in CSR communication. Indeed, a deficiency in the content can be partially compensated by effective communication (and vice versa).

For an in-depth description of the aggregation approach, please refer to Salvatore et al. [15].

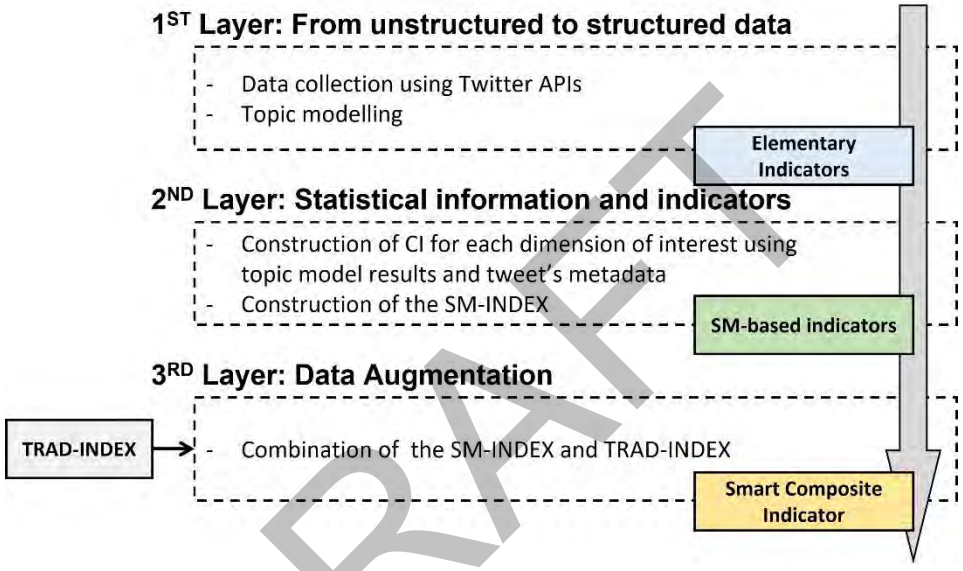



Figure 1: Modular Framework for Smart Composite Indicator adapted to our case study.

DRAFT

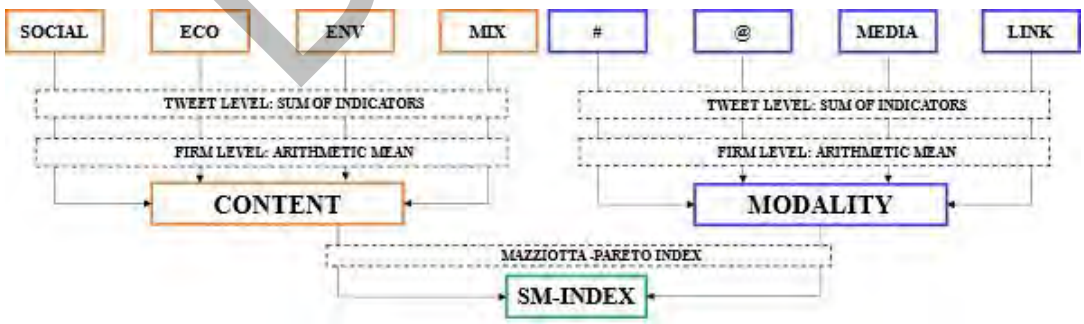



Figure 2: SM-INDEX construction strategy.

Finally, the third layer entails the data augmentation step. As TRAD-INDEX we consider the CSR-Strategy Score provided by Refinitiv, a commercial database. It reflects a company's practices to integrate economic (financial), social and environmental dimensions into its day-to-day decision-making process and it ranges between 0 and 100.

Considering the SM-INDEX and the TRAD-INDEX, it is possible to build a combined innovative smart indicator (SMART-INDEX). For the aggregation of SM-INDEX and TRAD-INDEX, we propose to apply the MPI (Figure 3). Indeed, we assume that the two dimensions are partially compensatory, i.e.,

efficient communication might compensate low effective commitment and high effective commitment might compensate scarce communication.

The SMART-INDEX measures the commitment in a more comprehensive way, considering not only the effective commitment (traditional indicator) but also the effort in online CSR communication (social media indicator).

Figure 3: SMART-INDEX construction strategy.

## 3. Emerging aspects of quality evaluation

Evaluating the quality of the resulting innovative indicators (SM-INDEX and SMART-INDEX) is crucial. The quality of a composite indicator depends on several factors, including the quality of the underlying data, construction procedures and its ability to produce stable and accurate measurements (robustness).

In the field of composite indicators, robustness is evaluated through uncertainty (UA) and sensitivity analyses (SA) [13]. UA focuses on how uncertainty in the input factors propagates through the structure of the composite indicators and influence its value. SA studies how much each individual source of uncertainty contributes to the output variance. For a general discussion of the procedures, please refer to [13].

However, considering the multi-source nature of the process, the evaluation of data quality and robustness is needed at each of the three layers and new aspects related to digital trace data should be considered [12]. These aspects include data retrieval (e.g., selection of social media accounts), data preprocessing and analytical methods to transform unstructured data into structured. Therefore, careful consideration of these factors is crucial to evaluate the quality of the innovative indicators.

In the next sections we present some preliminary analysis related to the influence of data retrieval and analytical choices on the SMART-INDEX. Additional results will be available in a forthcoming paper.

DRAFT

## 3. Sensitivity Analyses - Preliminary Results

In this section, we present the initial findings from a portion of our sensitivity analyses. It could be argued that the uneven selection of accounts (CSR, news, and multipurpose) may impact the results. Therefore, we conducted the first part of our sensitivity analyses to investigate the effects of various data selection strategies. Specifically, we compared the indexes derived from the complete dataset to those obtained from three distinct scenarios.

1. *Scenario 1:* all accounts but restricted to tweets for which the CSR text proportion (according to the STM model results) is higher than 50%.
2. *Scenario 2:* all accounts but restricted to tweets for which the CSR text proportion (according to the STM model results) is higher than 80%.
3. *Scenario 3:* only CSR accounts.

To facilitate summary and illustration, we limit our analysis to firms with more than 500 tweets in all scenarios (i.e., we only consider Microsoft, Verizon, Salesforce, Amgen and Cisco). Table 1 provides the values of SM-INDEX, TRAD-INDEX and SMART-INDEX considering all data and in the three scenarios considered above.

In general, we can see that the TRAD-INDEX is very similar across all companies. A rational behind this similarity is that the index is constructed considering mainly compliance to laws and regulation with respect to CSR reporting that, nowadays, is a common practice for most companies. The SM-INDEX

allows to discriminate better the communication about CSR commitment among firms. The combination of the two indicators provides an innovative measure of CSR commitment and communication effectiveness, giving additional insights to researchers.

Table 1: Sensitivity analyses results.

| | | All data | | | Scenario 1: CSR >= 50% | | |
|---|---|---|---|---|---|---|---|
| Firm Name | TRAD INDEX | SM INDEX | SMART INDEX | Ranking | SM INDEX | SMART INDEX | Ranking |
| Verizon | 107,30 | 89,61 | 96,87 | 5 | 91,26 | 97,86 | 5 |
| Salesforce | 99,36 | 98,25 | 98,80 | 4 | 96,45 | 97,99 | 4 |
| Microsoft | 104,44 | 95,18 | 99,38 | 3 | 94,51 | 98,98 | 3 |
| Amgen | 107,30 | 103,40 | 105,28 | 2 | 104,11 | 105,66 | 2 |
| Cisco | 107,30 | 110,22 | 108,72 | 1 | 110,24 | 108,73 | 1 |
| | | Scenario 2: CSR >= 80% | | | Scenario 3: Only CSR Accounts | | |
| Firm Name | TRAD INDEX | SM INDEX | SMART INDEX | Ranking | SM INDEX | SMART INDEX | Ranking |
| Verizon | 107,30 | 90,26 | 97,31 | 5 | 91,63 | 98,23 | 4 |
| Salesforce | 99,36 | 97,43 | 98,37 | 4 | 97,40 | 98,36 | 3 |
| Microsoft | 104,44 | 95,00 | 99,27 | 3 | 92,99 | 98,05 | 5 |
| Amgen | 107,30 | 103,44 | 105,30 | 2 | 104,46 | 105,84 | 2 |
| Cisco | 107,30 | 110,52 | 108,86 | 1 | 109,32 | 108,29 | 1 |

To the purpose of the sensitivity analyses, we can compare the general results with the three scenarios. When ranking the companies based on the value of the smart INDEX, it is evident that there are no differences with Scenario 1 and 2. However, the ranking in scenario 3 is different although the value of the indexes slightly differs. More precise measures of the size of the difference can be computed through the difference of the ranks and the calculation of Spearman correlation coefficient.

## 4. Conclusions and Future Developments

This paper demonstrated that augmenting traditional data with digital trace data offers new opportunities to better understand phenomena. In particular, the proposed SMART-IDEX allows to measure sustainability commitment in a more comprehensive way.

However, an important point to stress is the quality of the resulting indicator. This paper represents a first attempt to understand the stability of the results to different methodological choices. Future developments for the final paper include a comprehensive classical sensitivity analysis (e.g. aggregation strategies, weighting) when constructing both the SM and the SMART indexes and a more in-depth understanding of the impact of analytical choices when using innovative and unstructured data sources.

DRAFT

## References


[1] Araujo, Theo; Kollat, Jana. Communicating effectively about CSR on Twitter: The power of engaging strategies and storytelling elements. Internet Research, 2018.

[2] Biemer, Paul P. "Total survey error: Design, implementation, and evaluation." Public opinion quarterly 74.5 2010: 817-848.

[3] Bernal, Irina; Sejersen, Tanja. Big data for economic statistics. 2021.

[4] Bianchi, Annamaria, Salvatore, Camilla, and Biffignandi, Silvia. "Using Social Media to Enhance Survey Data." The Survey Statistician. Vol. 87. 2023: 27-34.

[5] Carroll, Archie B. "The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders." Business horizons 34.4 1991: 39-48.

[6] Costanzo, Luigi. "Use of Administrative Data and Use of Estimation Methods for Business Statistics in Europe: an Overview." Admin Data ESSnet Workshop “Using Admin Data-Estimation approaches”(Vilnius). 2011.

[7] Imaz, Oier, and Andoni Eizagirre. "Responsible innovation for sustainable development goals in business: An agenda for cooperative firms." Sustainability 12.17 2020: 6948.

[8] Mazziotta, Matteo, and Adriano Pareto. "Methods for constructing composite indices: One for all or all for one." Rivista Italiana di Economia Demografia e Statistica 67.2 2013: 67-80.
[9] Joint Research Centre-European Commission. Handbook on constructing composite indicators: methodology and user guide. OECD publishing, 2008.
[10] Ricciato, Fabio, Albrecht Wirthmann, and Martina Hahn. "Trusted Smart Statistics: How new data will change official statistics." Data & Policy 2 2020: e7.
[11] Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. "Stm: An R package for structural topic models." Journal of Statistical Software 91 2019: 1-40.
[12] Rocci, Fabiana, Roberta Varriale, and Orietta Luzi. "Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes." Journal of Official Statistics 38.2 2022: 533-556.
[13] Saisana, Michaela, Andrea Saltelli, and Stefano Tarantola. "Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators." Journal of the Royal Statistical Society: Series A (Statistics in Society) 168.2 2005: 307-323.
[14] Salvatore, Camilla, Silvia Biffignandi, and Annamaria Bianchi. "Corporate Social Responsibility Activities Through Twitter: From Topic Model Analysis to Indexes Measuring Communication Characteristics." Social Indicators Research 2022: 1-32.
[15] Salvatore, Camilla, Silvia Biffignandi, and Annamaria Bianchi. Augmenting Business Statistics Information by Combining Traditional Data with Textual Data: A Composite Indicator Approach, submitted 2023.

DRAFT

# Enhancing Principal Components by a Linear Predictor: an Application to Well-Being Italian Data

Laura Marcis[a], Maria Chiara Pagliarella[a], and Renato Salvatore[a]

[a]University of Cassino and Southern Lazio (Italy); laura.marcis@unicas.it, mc.pagliarella@unicas.it, rsalvatore@unicas.it

**Abstract**

We consider the case of a multivariate random vector that obeys a linear mixed model when the vector itself lies in a lower dimensional subspace. This situation suggests that this subspace can be modeled by the probabilistic (random-effects) principal components. By reason of this, the random vector follows at the same time two different models. We employ a linear predictor adjusted by the residual part of the probabilistic principal components that results not explained by the linear model. The new predictor can be considered as the vector of scores that comes from that principal components, enhanced by the linear mixed model. The application to the official Italian well-being data shows some features of the method.

***Keywords:*** multivariate random vector, principal components, linear mixed model, well-being

DRAFT

## 1. Introduction

Principal component analysis (PCA) is recognized as one of the most employed methods to reduce dimensionality, by means of the projection of a set of variables in a subspace of them. By summarizing and allowing to visualize data, and, at the same time, minimizing the loss of information in the lower dimensional space, in many cases principal components (PCs) lead to a better assessment of the bundled statistical information, seized by the original variables [2]. Because PCs are linear combinations, the interpretation of the scores by these new "data-dependent" variables is hard to give some time. In particular situations, the contribution improve understanding of some case studies may be poor and may lead to misguided or unclear findings. Furthermore, in the great majority of cases, when the interpretation stills on loadings that exceed a threshold, the linking of the variables hardly fails to provide some explanation or a bit of insight. Because of using of the common practice of ignoring the PCs affected by lower loadings, the focus shifts to the first PCs, which arise from the most correlated variables in the original set. The issue of resting on the main linearly-dependent variables is particularly relevant and becomes crucial in several instances. We may come across the trade-off between considering the retained PCs as highlighting latent phenomena, or in reproducing similar information unnecessarily. One of the recurrent ways of approaching redundancy and, in general, the recursive informative content of multivariate sample data, is given by considering a common subset of covariates that the population obeys. Two main cases in the literature are deemed representative of the joint dependence on a multivariate vector, the PCs "with covariates" or Partial PCA, and the redundancy analysis. Given a subspace spanned by the sample vectors of predictor variables, both of them rely on the common baseline of splitting the sample variance as the sum of the variance "explained" by a multivariate linear regression model, and the variance due to the regression residuals. Although these last represent a very useful tool in some cases, the deployment

of linear models to explain part of the sample variability has had a remarkable development in the last years. One of these studies brings into play the role of prediction by linear statistical models.

Tipping and Bishop [7] had already introduced the notion of prediction for PCs. They called "Probabilistic PCA" (PPCA) the model behind the PCA, which parameters are estimated with the Expectation - Maximization algorithm. The "noisy" PC model (nPC), proposed by Ulfarsson and Solo [8] has a quite similar formulation with respect to the PPC model, providing - in a similar way - the nPC prediction after giving the model estimates. Instead of a "fixed effects PCs", as the traditional linear regression PCA model, the PPC (or nPC) are random variables. This condition suggests, on the one hand, the Bayesian approach to handle the estimates for the PPC linear model and, on the other hand, to predict PCs under its meaning within random linear models theory [4]. The Bayesian approach to the estimation requires an expectation of some model parameters that are random, conditional to the observed data. Given normality of the error $\varepsilon \sim N(0, \sigma^2 I)$, for a linear model $\tau = B\lambda + \varepsilon$ - in case of the vector $\lambda$ random - the likelihood is based on the conditional distribution $\lambda|\tau \sim N[E(\lambda|\tau), var(\lambda|\tau)]$.

Moreover, it is known [6] that $E(\lambda|\tau) = \widetilde{\lambda}$ is the "Best Prediction" (BP) estimate, with $var(\widetilde{\lambda} - \lambda) = E_\tau[var(\lambda|\tau)]$. This is somewhat different from the standard linear regression model prediction by $E(\tau|\lambda)$. Therefore, given a linear mixed model (LMM) [1] for $\tau$, with $E(\tau|\lambda) = \lambda$, the model parameters are the realizations of random variables. The BP of a linear combination of the LMM fixed and random effects (i.e. linear in $\tau$, with $E[E(\tau|\lambda)] = 0$) gives the "Best Linear Unbiased Prediction" (BLUP) estimates [5]. LMM's are particularly suitable for modeling with covariates (fixed and random) and for specifying model covariance structures [1]. They allow researchers to take in account special data, such as hierarchical, time-dependent, correlated, covariance-patterned models. Thus, given the BP estimates of the nPC $\lambda$, $\widetilde{\lambda} = E(\lambda|\tau)$, the vector $\widetilde{\tau} = B\widetilde{\lambda}$ represents the BP of the $p$-variate vector.

In the present paper, we introduce a multivariate LMM that considers the dependent random vector effectively represented by the subspace of the PPCs. The new predictor combines the linear model and the PPC's, carrying simultaneously the Best Linear Predictor, and the contribution given by the PPCs not "explained" by the linear predictor itself. An application to the official Well-being Italian indicators shows some of the features of the method.

DRAFT

## 2. Theory

In the sequel we report the following symbols, giving the model specification.

- $n =$ the number of subjects in the LMM model $(i = 1, ..., n)$;
- $N = \Sigma n_i =$ total sampling units considered;
- $p =$ the number of the response dependent variables;
- $l =$ the number of the linear model covariates;
- $j = 1, ..., n_i$ the within-subject (groups) units;
- $s =$ the dimension of the effective PC subspace.

Consider $\Theta$ as the $N \times p$ sample matrix of the $p$-variate $p \times 1$ random vector $\theta$, with $N$ as the total number of the units given by the sample. Moreover, consider that the vector $\theta$ obeys the linear model:

$$\theta = \beta' x + u', \tag{1}$$

where $x$ is the $l \times 1$ vector of covariates, $\beta$ is the $l \times p$ matrix of the regression effects, $u$ is the vector of the $p$-variate random effect, with $u \sim N(0, \Sigma_u)$, $\Sigma_u = cov(u)$. Furthermore, we consider at the same time that the multivariate random vector $\theta$ obeys the following linear model:

$$\theta = Ab + \epsilon, \tag{2}$$

in which $A$ is $p \times s$ a loading matrix of eigenvectors, $b$ is the random vector of PPCs, and $\epsilon$ is a vector of isotropic error, with $\theta \sim N(\mu, A\Psi A' + \sigma_\epsilon^2 I)$, $b \sim N(0, \Psi)$, $\Psi = diag(\psi_1, \ldots, \psi_s)$, $s < p$, and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$. When a sample of $N$ observations is given, an $N \times p$ matrix $Y$ of observations from the random vector $\theta$ is simply modeled as $Y = \Theta + E$, with the "sampling error" $Np \times Np$ covariance matrix

$cov(vec(E)) = (\Sigma_e) \otimes I_N$ ($\otimes$ is the Kronecker product), $e \sim N(0, \Sigma_e)$, $\Sigma_e = var(e)$. Thus, models (1) and (2) are rewritten as $Y = \Theta + E = X\beta + ZU + E$, with the (1) that becomes $\theta = \beta' x + u' + e'$, and $Y = \Theta + E = BA' + \Xi + E = BA' + \Gamma$, with the (2) is $\theta = Ab + \epsilon + e = Ab + \gamma$, respectively. The model errors $u$, $\epsilon$, and $e$, are mutually independent. The matrix $Z$ represents the $N \times n$ design matrix of random effects and $E$ is the $N \times p$ matrix of the residual errors of the multivariate LMM, $B$ is the $N \times s$ matrix of the PPCs that lie in the $s$-dimensional subspace, $\Xi$ is the $N \times p$ matrix of the isotropic errors of the model (2). The models (1) and (2) have the following conditional expectations and variances:

$$\begin{aligned} E(\theta|y) &= \widetilde{\theta}_y = y - E(e|y) = y - cov(e,y)var(y)^{-1}y = y - var(e)Py, \\ var(\theta|y) &= var(\theta) - cov(\theta,y)var(y)^{-1}cov(y,\theta) \\ &= var(e) - var(e)Pvar(e), \end{aligned}$$

for the model in (1), where $P = \Sigma_y^{-1}(I - P_X)$, $\Sigma_y = var(y)$, and $P_X$ is the projection matrix. For the model in (2):

$$\begin{aligned} E(b|\theta) &= E(b) + cov(b,\theta)var(\theta)^{-1}(\theta - \mu) \\ &= cov(b, \mu + Ab + \epsilon)C^{-1}(\theta - \mu) = \Psi A' C^{-1}(\theta - \mu), \\ var(b|\theta) &= var(b) - cov(b,\theta)var(\theta)^{-1}[cov(b,\theta)]' \\ &= \Psi - \Psi A' C^{-1} A\Psi \\ C &= A\Psi A' + \sigma_\epsilon^2 I. \end{aligned}$$

Based on some results on linear projections, i.e., given the random variable $y$, and the $1 \times j$, $1 \times k$ random vectors $x, z$, with positive definite covariance matrix of $(y, x, z)'$, then for the linear projection $L(y|x,z)$:

$$L(y|x,z) = L(y|x) + [z - L(z|x)]\gamma,$$

where $\gamma = var(z|x)^{-1}[cov(y,z|x)]'$, we get the following:

**Proposition 1.** *Given the model (2) for the $p$-dimensional random vector $\theta$, with $b = \overline{F}'(\theta - \epsilon)$, and under the models in (1) and (2), the multivariate Best Predictor based on $(y,b)$, $E(\theta|y,b)$, is:*

$$E(\theta|y,b) = \widetilde{\theta}_{y,b} = E(\theta|y) + cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \tag{3}$$

*with $\widetilde{b} = E(b|\theta)$, $\overline{F}$ the $sN \times pN$ matrix $(\overline{A}'\overline{A})^{-1}\overline{A}$, and $\overline{A}$ is the $pN \times sN$ matrix $A \otimes I_N$. Then, $var(\theta|y,b) = var(\theta|y) - cov(\theta,b|y)var(b|y)^{-1}[cov(\theta,b|y)]'$.*

The "hybrid" predictor $\widetilde{\theta}_{y,b}$ in (3) gives the Best Linear Unbiased Predictor $E(\theta|y)$, "embedding" the PPCs through an adjoint component. The last is due to knowing that the random vector $\theta$ lies in the $s$-dimensional subspace of the PPCs. In particular, the difference $\widetilde{b} - E(b|y)$ gives the multivariate vector of the PPCs "not explained" by the estimation of the linear model $E(\theta|y)$. The matrix $var(\theta|y,b)$ has rank $s$, and, consequently, there are $(p-s)$ linear combinations of $\theta$ for which their respective variances do not depend on the PPCs.

**Proposition 2.** *Given the $p$-dimensional random vector $\theta$, under the models in (1) and (2), and the Best Predictor $E(\theta|y,b)$ in (3), we get:*

$$\begin{aligned} \overline{F}E(\theta|y,b) &= \overline{F}\widetilde{\theta}_{y,b} = \widetilde{b}^* \\ &= \overline{F}E(\theta|y) + \overline{F}cov(\theta,b|y)var(b|y)^{-1}\left\{\widetilde{b} - E(b|y)\right\}, \end{aligned} \tag{4}$$

*where $\widetilde{b}^*$ is the $s$-dimensional vector of the PPCs "enhanced" (ePCs) by the linear predictor $E(\theta|y)$. As a particular case, when $\sigma_\epsilon^2 \longrightarrow 0$, $var(\epsilon) \longrightarrow 0$, $var(\gamma) \longrightarrow var(e)$, and $\widetilde{b} \longrightarrow \overline{b}$. Therefore, $\overline{F}\widetilde{\theta}_{y,b} = \widetilde{b}^* \longrightarrow \overline{b}$ with $\overline{b}$ the sample PCs $\overline{b} = A'\theta$.*

The ePCs $\widetilde{b}^*$ are then the PPCs "adjusted" by the Linear BP $E(\theta|y)$, and $\widetilde{b}^*$ is then the vector of the ePC scores. Note that the ePC scores give a non-orthogonal matrix. Given $\sigma_\epsilon^2 = 0$, the vector $\theta$ in the model (2) lies in the $s$-dimensional subspace of the sample PCs $\overline{b}$. In fact, in this case $\widetilde{b} \equiv \overline{b}$, $cov(\theta,b|y) = var(\theta|y)\overline{F}'$, $var(b|y) = \overline{F}var(\theta|y)\overline{F}'$, $E(b|y) = \overline{F}E(\theta|y)$, and then $\overline{F}E(\theta|y,b) = \overline{b}$.

DRAFT

## 3. Application

In accordance with the recent law reforms in Italy, the Equitable and Sustainable Well-being indicators (in Italian, BES) [3] - annually provided by the Italian Statistical Institute (ISTAT) - are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. In order to highlight the result of the proposed method we use 12 BES indicators relating to the years 2013-2016, collected at NUTS-2 (Nomenclature of Territorial Units for Statistics 2 level). The variables employed in the application study are in Table 1. We use the per capita adjusted disposable income variable (its logarithm, as is usually done in economics studies) - indicated with BE1 - as a unique covariate in the LMM model, while the remaining 11 variables are dependent variables (Table 1 reports the description and acronyms used for the variables). The application uses the Restricted Maximum Likelihood estimation, a Sas/IML code, and a sequence of Sas-HPMixed procedures. Table 2 shows the slope parameter estimates from the multivariate regression, with their significance level. Table 3 reports the MANOVA multivariate test statistics, based on the characteristic roots. These are the eigenvalues of the product of the sum-of-squares matrix of the regression model and the sum-of-squares matrix of the regression error. The null hypothesis for each of these tests is the same: the independent variable (LBE1) has no effect on any of the dependent variables. The four tests are all significant.

Figure 1 shows the application of Proposition 1, where all the measures are plotted in the space of the sample PCs. This plot reports simultaneously the factorial coordinates of the original variables, of the linear predictor $E(\theta|y)$, and of the hybrid predictor $\widetilde{\theta}_{y,b} = E(\theta|y,b)$. The dependent criterion variables in the application can be split into two main groups, starting from both an analysis of the plot and the correlation matrix between the sample PCs, the LMM predicted values, and the hybrid predictor values. Moreover, we have eight dependent variables for which there is accordance in terms of their mutual correlation inside the original variables, as well as the LMM predicted and the hybrid predictor. This means that the hybrid predictor $\widetilde{\theta}_{y,b}$ does not change significantly the mutual correlations, substantially because the component of the PPCs not explained by the linear predictor $E(\theta|y)$ is relatively small. For the remaining three variables - with the acronyms REL4, Q2, and BS3 - the correlation changes: in some cases it changes sign, going from positive correlation values by the sampling and predicted values, to negative correlation values between the predictor $\widetilde{\theta}_{y,b}$, and vice versa. Therefore the predictor (3) highlights the major influence of the component of the PPCs not explained by the linear predictor $E(\theta|y)$. The latter is in accordance with the sample PCs in the mutual correlations between these three criterion variables. Since the classical predictor matches the mutual correlation inside the original variables - meaning that these mutual relations in the sample are due to the disposable income (BE1), the covariate in the mixed model regression - then the different mutual correlation values of the hybrid predictor $\widetilde{\theta}_{y,b}$ can be interpreted as relationships not captured by the model. For instance, the correlation between $Q2_{E(\theta|y)}$ and $INN1_{E(\theta|y)}$ is $-0.82$, and becomes $+0.30$ between $Q2_{\widetilde{\theta}_{y,b}}$ and $INN1_{\widetilde{\theta}_{y,b}}$, meaning that conditionally to the model - thus looking at the correlation between the attendance of childhood services (Q2) and the percentage of R&D (INN1) in the space orthogonal to per capita income - the correlation is positive. It could be interpreted as saying that the Regions with a greater investment in R&D have a higher benefit of childhood services. In our opinion, this highlights how the "hybrid" predictor is able to grasp the relationship that actually exists between investment in research and development and the importance given to training starting from the earliest years of life, regardless of per capita income.

DRAFT

## 4. Discussion

The introduced predictor (3) can be viewed from two different perspectives. An "adjusted" linear predictor by the PPCs, and, by relation (4), as the enhanced PCs (ePCs) that modify the probabilistic PCs (PPCs) to accommodate the mixed model regression predicted values. The present work considers the probabilistic principal components like a "constraint" model, that links together the components of the multivariate random vector in a lower dimensional subspace.

While the estimation of the PPC model requests a quite simple procedure, one of the causes of concern in the estimation of the parameters of a multivariate mixed model is the number of covariance

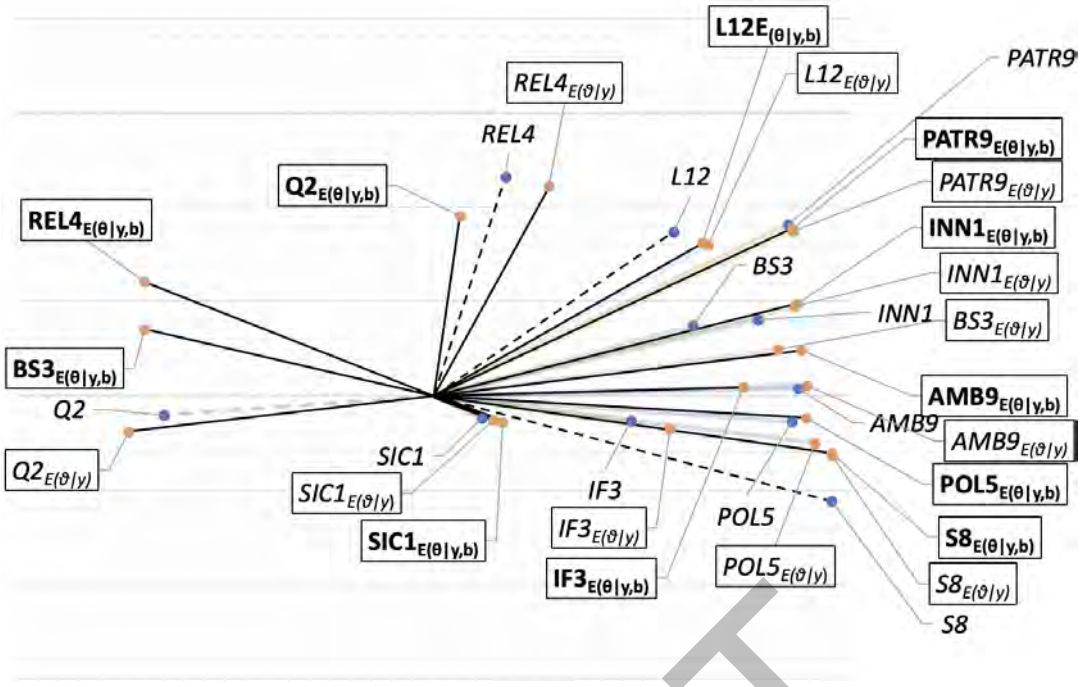



Figure 1: Plot of the application of the model (1) in the space of the sample PCs. There are represented the factorial coordinates of the original variables, of the linear predictor $E(\theta|y)$, and of the hybrid predictor $\widetilde{\theta}_{y,b} = E(\theta|y,b)$.

parameters, which may be too high for a speed software computation. Like the application presented, we suggest estimating the model covariance parameters under a uniform correlation structure among the multivariate components of the random effects. This structure is equivalent to the compound-symmetry covariance structure, with a better numerical property in terms of optimization. Indeed, some studies highlight that using uniform correlation matrices reduces the estimation noise. The model covariance matrix of random effects is then a generalized uniform correlation matrix, and works with two parameters. The "hybrid" multivariate linear predictor (3), by adjusting its standard formulation through the sample parameter vector scores in a convenient subspace, is designed to accommodate not only PPCs, but also factor models based on a random structure. In the present work, the components of the multivariate parameter among the subjects are linked by a principal components model. In order to overcome convergence problems, the method introduced can be extended to include multidimensional information from the data, through a reduced number of dependent variables in the linear model.

DRAFT

## References


[1] Demidenko, E.: Mixed Models: Theory and Applications. Wiley, New York (2004)

[2] Hardle, W. K., Simar, L.: Principal Components Analysis. In Hardle, W. K., Simar, L. (eds.) Applied Multivariate Statistical Analysis, 319-358. Springer (2015)

[3] ISTAT: BES project www.istat.it/en/well-being-and-sustainability

[4] Longford N.T.: Random Coefficient Models. In: Lovric M. (eds.) International Encyclopedia of Statistical Science. Springer, Heidelberg (2011)

[5] McCulloch, C.E., Searle, S.R.: Generalized Linear and Mixed Models. Wiley, New York (2001)

[6] Timm, N. H.: Applied Multivariate Analysis. Springer, New York (2002)

[7] Tipping, M.E., Bishop C.M.: Probabilistic principal component analysis. J. R. Stat. Soc., Ser. B (Stat. Methodol.) **61**(3), 611–622 (1999)

[8] Ulfarsson, M.O., Solo, V.: Sparse variable PCA using geodesic steepest descent. IEEE Trans. Signal Process **56**(12), 5823–5832 (2008)

Table 1: Description of the variables used for the application

| Variables | Description |
|---|---|
| S8 | Age-standardised mortality rate for dementia and nervous system diseases |
| IF3 | People having completed tertiary education (30-34 years old) |
| L12 | Share of employed persons who feel satisfied with their work |
| REL4 | Social participation |
| POL5 | Trust in other institutions like the police and the fire brigade |
| SIC1 | Homicide rate |
| BS3 | Positive judgment for future perspectives |
| PATR9 | Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest |
| AMB9 | Satisfaction for the environment - air, water, noise |
| INN1 | Percentage of R&D expenditure on GDP |
| Q2 | Children who benefited of early childhood services |
| BE1 | Per capita adjusted disposable income |
| LBE1 | Logarithm of Per capita adjusted disposable income |

DRAFT

Table 2: The slope parameters by the multivariate regression with the LBE1 covariate

| Dependent variable | Slope parameter (LBE1) | STD Error | t | Pr >t |
|---|---|---|---|---|
| AMB9 | 0.9802 | 0.3255 | 3.01 | 0.0035 |
| BS3 | 0.9330 | 0.0891 | 10.47 | 0.0001 |
| IF3 | -0.3166 | 0.1673 | -1.89 | 0.0621 |
| INN1 | -0.0433 | 0.0170 | -2.54 | 0.0130 |
| L12 | 0.0016 | 0.0107 | 0.15 | 0.8786 |
| PATR9 | 0.0975 | 0.0756 | 1.29 | 0.2007 |
| POL5 | -0.0036 | 0.0085 | -0.42 | 0.6775 |
| Q2 | 0.2031 | 0.1762 | 1.15 | 0.2526 |
| REL4 | 0.5602 | 0.1690 | 3.31 | 0.0014 |
| S8 | -0.0506 | 0.0293 | -1.73 | 0.0879 |
| SIC1 | -0.0072 | 0.0150 | -0.48 | 0.6314 |

Table 3: MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall LBE1 Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks Lambda | 0.0566 | 102.98 | 11 | 68 | <.0001 |
| Pillai's Trace | 0.9434 | 102.98 | 11 | 68 | <.0001 |
| Hotelling-Lawley Trace | 16.6590 | 102.98 | 11 | 68 | <.0001 |
| Roy's Largest Root | 16.6590 | 102.98 | 11 | 68 | <.0001 |

# Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data

Farah Naz[a] and Elena Ballante[b,c]

[a]Department of Mathematics, University of Pavia, Italy ,
farah.naz01@universitadipavia.it
[b]Department of Political and Social Sciences, University of Pavia, Italy
[c]IRCCS Mondino Foundation, Pavia, Italy

**Abstract**

The health sciences often involve survival data that may be censored and can contain correlated covariates. While there has been some research on the impact of correlated variables on survival models, there is a need for further investigation of how bootstrap methods can be used to handle correlation in survival analysis. In fact, if the variables are strongly correlated, the bootstrap samples from the prior may mask the effect of each other, making it difficult to discern the true relationship between the variables and the response, which can ultimately lead to unrealistic estimates. This article aims to extend the Proper Bayesian bootstrap ensemble tree model for analyzing survival data with highly correlated covariates. The model's performance was assessed through a simulated study, demonstrating better results compared to traditional survival models, such as the Cox model and survival random forest, with greater stability in terms of the integrated Brier score, particularly with smaller sample sizes.

***Keywords:*** Survival analysis, Bootstrap, Bayesian nonparametric learning, Ensemble models, Correlated data

DRAFT

## 1. Introduction

Multivariate responses that are highly correlated are frequently encountered in health science studies. In such cases, a subject may have multiple related responses, and the presence of these correlated variables can have a significant impact on survival models. The adverse effects of highly correlated variables include overfitting, reduced model performance, and biased estimation of individual predictor effects.

In cases where the data comprises highly correlated variables, the survival model may exhibit bias towards splitting one of the correlated variables while disregarding information from the other correlated variables (1). This can result in overfitting of the model, leading to a reduction in its predictive performance, as the model may fail to accurately capture the true underlying relationship between the variables and the time-to-event outcome (such as the expected survival time or the probability of survival beyond a specific landmark time-point).

This paper aims to enhance the use of the Proper Bayesian bootstrap ensemble tree model for analyzing survival data by incorporating a pre-processing stage. This stage involves carefully managing the data by either removing or transforming highly correlated variables, to reduce the detrimental impact that high correlation can have on the analysis.

The model performs Bootstrap resampling techniques to approximate the posterior distribution

of a statistical function of a decision tree $\phi(F)$, where $F$ is a random distribution function as defined in (2).
The proposed method is rooted in the family of Bayesian bootstrap procedures, starting with Rubin (1981) (3), followed by Muliere and Secchi's proper Bayesian bootstrap (1996) (2), and Lo's Bayesian bootstrap for censored data (1993) (4). The algorithm also draws upon Efron's classical bootstrap technique (5) for bagging survival trees and survival forests.
The approach inherits the advantages of Bayesian non-parametric learning such as flexibility and computational strength and considers prior opinions to overcome the drawbacks of traditional bootstrap procedures used in classical ensemble decision tree models.
The structure of the paper is as follows: Section 2. details the proposed methodology, while Section 3. and 4. provide information on the computational environment and present the initial findings, respectively. Finally, Section 5. summarizes the main outcome of the study and discusses the potential areas for future research.

## 2. Proposed Model

This paper aims to extend the idea of a Proper Bayesian Bootstrap Ensemble Tree model, specially tailored for survival data analysis. The model is designed to address the challenges posed by highly correlated variables and their impact on the model's performance. The proposed model is based on the initial outcomes of the Proper Bayesian Bootstrap concept, which was previously introduced by (2). However, it has been adapted and modified to suit the specific requirements of survival data analysis.
The primary metric for evaluating ensemble tree models is the decision tree $\phi(F, X)$, which depends on the underlying distribution $F$ and the observed data $X$. According to a study by (6), the posterior distribution of $\Phi$ can be estimated using bootstrap procedures. This is achieved by fitting the model to a weighted dataset generated from the bootstrap process, and the resultant predictions provide an estimate of the posterior mean.
We set a prior distribution $D(kF_0)$ for $F$ using a Dirichlet process to apply the Proper Bayesian bootstrap. To explain the response variable $y$ based on a list of highly correlated covariates $x_1, ..., x_P$, the parameter $F_0$ of the Dirichlet process is established as a joint distribution that depends on both $x$ and $y$.
The bootstrap sampling method from the posterior of $\phi(F, X)$ is taken from (6), where each bootstrap resample $(x_1^*, y_1^*), ..., (x_m^*, y_m^*)$ is created by combining distributions from the prior estimate $F_0$ and the empirical distribution $F_n$. Since the covariates are highly correlated, in the bootstrap resampling process, a new sample is generated from the original distribution $F_0$ and a new vector of covariates is produced from the original prior distributions $F_0(x_k)$ that accounts for the dependence among the covariates. The resampling process continues iteratively until the desired number of bootstrap samples is generated. This approach can help to account for the dependence among the covariates and produce more accurate estimates of the parameters of interest.
The response variable $y$ is linked to the vector of covariates generated from the prior $F_0$ estimate obtained using an appropriate survival model to perform survival analysis. The method for combining the predictions should be based on the characteristics of the time-to-event data. Such as suggested in (7) the output of the model is a bootstrap aggregated version of the estimated conditional survival function $S$ for a new observation $X_{new}$ computed by $\widehat{S}_A^B(\cdot|x_{new}) = \widehat{S}_{L^B_{A(}x_{new})}(\cdot)$.
The distinguishing feature of the proposed method from traditional ensemble methods is its ability to generate new observations through the prior distribution $F_0$, which improves the prediction of the model.

DRAFT

Moreover, as a novelty element with respect to (8), we incorporate the correlation structure between covariates into the bootstrap resampling technique. With this modification, it takes into account the covariance matrix of the covariates which prevents the bootstrap samples coming from the prior distribution from having an unrealistic joint distribution. Incorporating this feature has the potential to produce more robust and reliable results, particularly in situations where there are strong correlations among the covariates.

## 3. Experimental setting

The data simulation is done using the flexible-hazard method, as outlined in (9). Also, to demonstrate the capabilities of the proposed method, we investigated the result of different weights assigned to the prior distribution $F_0$ with sample size $N = 50$. The simulated dataset generated for the time-to-event target variable, including $10\%$ censored observations, comprises five highly correlated numerical covariates that are sampled from a multivariate normal distribution with adjusted parameters of mean and standard deviation. The correlation coefficients between covariates are higher than 0.85.
A total of $100$ simulated datasets were produced, each consisting of $N = 50$ samples. The survival time values for observations sampled from the prior were estimated using an exponential regression model.
The proposed model was contrasted against two of the most prevalent models in survival analysis, the Survival Random Forest, and the Cox model. The performance of the predictions was evaluated using the Integrated Brier Score (IBS) through a 5-fold cross-validation process.

## 4. Preliminary results

The average values and nonparametric confidence intervals of the IBS results for the comparison with classical models are displayed in Figure 1. Figure 2 shows the performances of multivariate sampling with respect to the indipendent one.

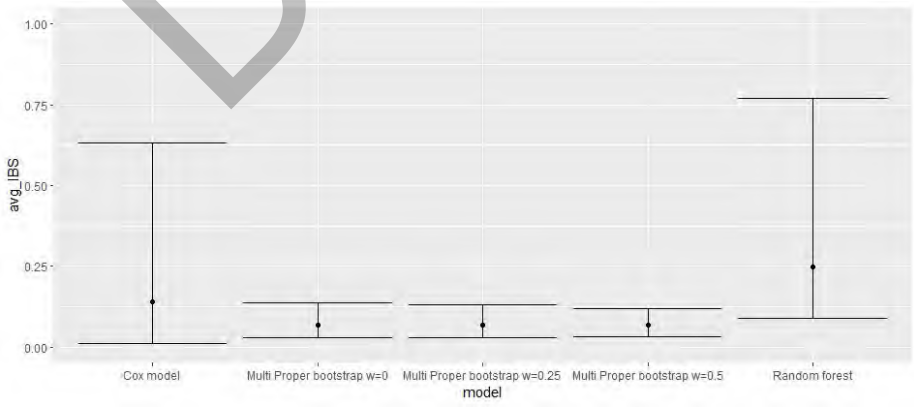



Figure 1: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 100 simulated datasets of highly variable covariates of sample size 50. The proposed model is compared with Random Survival Forest and Cox Model

DRAFT

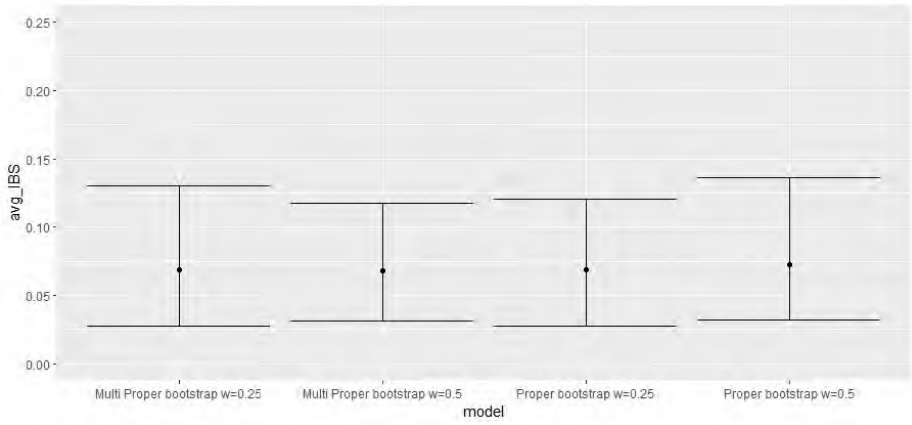



Figure 2: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 100 simulated datasets of highly variable covariates of sample size 50. Multidimensional prior sampling is compared to the independent one

The results in Figure 1 showed that the average IBS score for the proposed Proper bootstrap model was lower, but not significantly lower when considering the width of the confidence interval for the traditional Cox and Random Forest models which is evidently large. Observing the confidence intervals of the proposed model which are significantly smaller compared to those of the traditional models, indicates that the proposed method is more consistent in its prediction performance.
The results in Figure 2 show that the application of multivariate prior sampling, that takes into account the covariance structure of the covariates, leads to slightly better results than the independent sampling when the covariates are highly correlated.

## 5. Conclusion

The paper introduces a novel ensemble tree modeling approach that utilizes Proper Bayesian Bootstrap, to analyze survival data while mitigating the effect of highly correlated variables to obtain increased stability and comparable results in a simulated environment.
The use of synthetic data, derived from prior distributions that are not present in the original dataset, overcomes the drawbacks of classical ensemble models which only consider the data without any prior opinion. This as a result enhances the stability of the final ensemble model, particularly for datasets with limited sample sizes. Taking into account the covariance structure of the data at hands prevent the proper Bayesian bootstrap from generating unrealistic data that could lead to more noisy results.
Further research is planned to evaluate the model's sensitivity towards varying degrees of censored data, categorical variables, and more effective techniques for sampling correlated data variables, derived from non-normal distributions.

## References

[1] X.-R. Liu, Y. Pawitan, and M. S. Clements, "Generalized Survival Models for Correlated Time-to-event Data," *Statistics in Medicine*, vol. 36, no. 29, pp. 4743–4762, 2017.
[2] P. Muliere and P. Secchi, "Bayesian Nonparametric Predictive Inference and Bootstrap Techniques," *Annals of the Institute of Statistical Mathematics*, pp. 663–673, 1996.

DRAFT

[3] D. B. Rubin, “The Bayesian Bootstrap,” *The annals of statistics*, vol. 14, no. 3, pp. 130–134, 1981.
[4] A. Y. Lo, “A Bayesian Bootstrap for Censored Data,” *The Annals of Statistics*, pp. 100–123, 1993.
[5] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics*, vol. 7, no. 1, pp. 569–593, 1992.
[6] M. Galvani, C. Bardelli, S. Figini, and P. Muliere, “A Bayesian Nonparametric Learning Approach to Ensemble Models Using the Proper Bayesian Bootstrap,” *Algorithms*, vol. 14, no. 1, pp. 1999–4893, 2021.
[7] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Troger, “Bagging Survival Trees,” *Statistics in Medicine*, vol. 23, no. 1, pp. 77–91, 2004.
[8] E. Ballante, “An extension of proper bayesian bootstrap ensemble tree models to survival analysis,” in *Book of Short Papers of the 51th Scientific Meeting of the Italian Statistical Society* (A. Balzanella, M. Bini, C. C., and R. Verde, eds.), pp. 1766–1770, Pearson, 2022.
[9] J. J. Harden and J. Kropko, “Simulating Duration Data for the Cox Model,” *Political Science Research and Methods*, vol. 7, no. 4, p. 921â928, 2019.

DRAFT

# ROBOUT: a step-wise methodology for conditional outlier detection

Matteo Farnè[a] and Angelos Vouldis[b]

[a]University of Bologna; Via delle Belle Arti 41; Bologna; Italy matteo.farne@unibo.it
[b]European Central Bank; DG Statistics; Frankfurt am Main; Germany
angelos.vouldis@ecb.europa.eu

**Abstract**

We present a step-wise methodology, called ROBOUT, to recover outliers in a dependent variable conditional on its predictors, to be identified from a large number of potential predictors. ROBOUT entails a preliminary imputation procedure to identify potential leverage outliers, a robust variable selection (via LASSO-penalized Huber loss regression), a robust regression (via MM) and an outlier detection step. We show in an ad-hoc simulation study that ROBOUT is the most effective methodology for what concerns outlier detection, predictor recovery, and coefficient estimation, even compared to existing integrated procedures like SPARSE-LTS and RLARS.

***Keywords:*** conditional outlier, leverage outlier, high dimension

DRAFT

## 1. Introduction

We propose a sequential methodology, named ROBOUT, to identify the most suitable regression model for a target variable from a high number of potential predictors, and to recover anomalies in a target variable conditionally on the identified predictors. ROBOUT comprises two consecutive robust steps, one performing a variable selection and the second performing coefficient estimation. The method is shown through simulations to outperform competitive approaches in the estimation of coefficients and the identification of outliers, when the dataset is contaminated with outliers.

Although integrated methods that simultaneously perform predictor recovery, outlier detection and coefficient estimation exist in the literature, such as RLARS (2) and SPARSE-LTS (1), they may be computationally expensive or systematically ineffective when the ratio $p/n$ is large. Differently, SNCD (Semismooth Newton Coordinate Descent) algorithm (5) provides a fast and reliable solution to the recovery of predictors, by minimizing a Huber or a least absolute deviation (LAD) loss of the residuals penalized by an elastic net (7). Its optimization problem is solved by explicitly deriving Karush-Kuhn-Tucker (KKT) conditions. At the same time, SNCD has two significant drawbacks: it provides no residual scale estimate, therefore no outlier detection can be conducted, and it is not robust to outliers in the predictors, as shown in (1).

For this reason, we provide a robust preliminary imputation procedure, that replaces anomalous values in the predictors by the median of the $s$ closest non-outlying neighbours, identified by looking at the most (robustly) correlated variable with the predictor in question. This procedure is designed to cope with two relevant confounding factors for predictor recovery such as a high level of multicollinearity, and a high number of variables $p$ compared to the sample size $n$, which actually positively impact on our leverage outlier identification procedure. In this way, SNCD may be applied with Huber loss on the clean

imputed dataset to identify the *right* predictors, on which a robust regression model like MM regression (6) can be finally employed to spot conditional outliers, by means of the robustly estimated residual scale. ROBOUT is therefore a doubly robust post-selection method, both with reference to outliers in $y$ and in its predictors.

In order to summarize the task which we address, let us consider $n$ numerical observations of one response variable $y$ and $p$ additional variables. We call the unknown set of conditional outlier indices $O$ with $|O| = [\alpha n]$ and $\alpha \in [0, 0.5]$. The response variable vector $\mathbf{y}$ is expressed in terms of the following regression model

$$\mathbf{y} = a + \mathbf{X}_D\beta + \epsilon, \tag{1}$$

where $\mathbf{y}$ is the $n\times 1$ vector of the response variable, $a$ is the intercept, $\mathbf{X}_D$ is the $n\times K$ matrix of predictors (whose columns are indexed by $D$, with $K \ll p$), $\beta$ is the $K \times 1$ vector of regression coefficients and $\epsilon$ is the $n \times 1$ vector of residuals. In this context, the statistical challenge lies in identifying the set $D$ of true predictors and correctly estimating the relative regression coefficients $\beta$ in a robust way, while simultaneously recovering the outliers in the set $O$.

## 2. ROBOUT methodology

We present here our proposed conditional outlier detection method, ROBOUT, consisting of four robust steps, namely, preliminary imputation, variable selection, low-dimensional regression and conditional outlier detection.

### 2.1 Preliminary imputation

Let us consider the $n \times p$ matrix of potential predictors $\mathbf{X}$. First, for each $j = 1, \ldots, p$, we compute $x_{med,j} = \text{med}(x_j)$ and $x_{mad,j} = 1.4826\text{mad}(x_j)$. Then, relying on $x_{med,j}$ and $x_{mad,j}$, we derive a robust z-score for each entry $ij$ of $\mathbf{X}$: $z_{ij} = \frac{x_{ij}-x_{med,j}}{x_{mad,j}}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. In the end, we flag as outliers the entries that present an outlying $z_{ij}$, i.e. we set $w_{ij} = 0$ if $|z_{ij}| > \phi^{-1}(0.995)$, where $\phi^{-1}$ is the inverse standard normal distribution function, and $w_{ij} = 1$ otherwise.

At this stage, we calculate robust rank-based pairwise correlations for each pair $j'j''$ of variables, $j', j'' = 1, \ldots, p$, $j' \neq j''$. If $p$ is large, this can be done for a restricted number $p'$ of randomly chosen variables, with $p' = \min(p, [300^2/p])$, for instance. We can use Spearman's rho, that is $\hat{\rho}_{j'j''}$, but Kendall's tau, $\hat{\tau}_{j'j''}$, could similarly be used, because both measures are robust to leverage outliers. Then, for each entry $ij$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, such that $w_{ij} = 0$, we apply Algorithm 1.

**Algorithm 1** Algorithm for preliminary imputation.

For each $ij$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, such that $w_{ij} = 0$:

1. derive the set of ordered variable indices $D_j$ by sorting $\hat{\rho}_{j'j}$ (or $\hat{\tau}_{j'j}$), $j' \neq j$, in decreasing order;
2. take the first index in the ordered set $D_j$, $j^{top}$, such that $w_{ij^{top}} = 1$;
3. identify among all the points $i' \neq i$ the set $I_{i,s}$ of the $s$ closest neighbours of $x_{ij^{top}}$ (in absolute norm) with $w_{i'j^{top}} = 1$;
4. derive the set $I_{i,3} \in I_{i,s}$ that contains the three closest neighbours of $x_{ij^{top}}$ (in absolute norm) with $w_{i'j} = 1$, $i' \in I_{i,s}$;
5. impute $x^*_{ij} = \text{med}(\mathbf{X}_{I_{i,3}j})$.

Finally, we can standardize each column of the imputed matrix $\mathbf{X}^*$, by calculating the standardized imputed values $x^{**}_{ij} = \frac{x^*_{ij}-\overline{x}^*_j}{\text{sd}(x^*_j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$.

DRAFT

## 2.2 Variable selection

We aim to consistently select the relevant set of the predictors of our target response variable $y_i$, $i = 1, \ldots, n$, from a large set of variables under the presence of conditional outliers in mean.

The first method (5) that we consider is to use the objective function

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \rho_{H,\delta}(\epsilon_i) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2}$$

where $\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}^{**}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, $\lambda$ is a penalization parameter, and

$$\rho_{H,\delta}(t) = \begin{cases} \frac{t^2}{2\delta}, & |t| \leq \delta \\ |t| - \frac{\delta}{2}, & |t| > \delta \end{cases}$$

is the Huber weight function, where $\delta$ is a tuning parameter. Henceforth, we call (2) SNCD-H objective function.

The other robust alternative that we consider in the simulation study substitutes $\rho_{H,\delta}(\epsilon_i)$ in (2) with the absolute loss $\rho_L(\epsilon_i)$, where

$$\rho_L(t) = t\left(\frac{1}{2} - \mathrm{I}(t < 0)\right), \; t \in \mathcal{R}.$$

We call that version the SNCD-L objective function.

Minimization 2 with $\rho_{H,\delta}(t)$ or $\rho_L(t)$ is practically performed for a decreasing sequence of values $\lambda = \lambda_t$, $t = 1, \ldots, T$, such that $\lambda_0 = \lambda_{\max}$ and $\lambda_T = \lambda_{\min}$, where $\lambda_{\max}$ returns no predictors, and $\lambda_{\min}$ returns $K_{\max}$ predictors. Predictor selection is performed by the adaptive version of the strong rule of (4), proposed in (5). At each value of $\lambda_t$, that rule exploits the coefficient estimates at $\lambda_{t-1}$. The optimal $\lambda$ is then selected by cross-validation, using as Out-Of-Sample (OOS) metric the loss $\rho_{H,\delta}$ for SNCD-H and the loss $\rho_L$ for SNCD-L.

Let us denote by $\widehat{\beta}_{\min}$ and $\widehat{\beta}_{1SE}$ the coefficient vectors obtained by setting $\lambda = \lambda_{\min}$ or $\lambda = \lambda_{1SE}$ respectively. In general, $\widehat{\beta}_{\min}$ is composed by $\widehat{K}_{\min}$ coefficients indexed by the set $D_{\min}$ and $\widehat{\beta}_{1SE}$ is composed by $\widehat{K}_{1SE}$ coefficients indexed by the set $D_{1SE}$, with $\widehat{K}_{1SE} \leq \widehat{K}_{\min}$ and $D_{1SE} \subseteq D_{\min}$. Let us define the difference set $D_{diff} = D_{\min} \setminus D_{1SE}$ and the corresponding estimated coefficient vector $\widehat{\beta}_{diff}$. Then, it is sufficient to formally test the hypothesis $H_0 : \beta_{diff} = \mathbf{0}_{K_{\min} - K_{1SE}}$ by a robust ANOVA test between the two nested MM regression models. If the resulting p-value is smaller than $5\%$, we reject $H_0$ and we select $\widehat{D} = D_{\min}$ as predictor set, and $\widehat{K} = \widehat{K}_{\min}$ as the number of predictors. Otherwise, we select $\widehat{D} = D_{1SE}$ and $\widehat{K} = \widehat{K}_{1SE}$.

## 2.3 Robust regression

In this step, we apply a robust regression method to the recovered predictors stored in matrix $\mathbf{X}_{\widehat{D}}$, in order to robustly estimate coefficients, residuals, and residual scale. To this aim, we utilise MM regression (6). In the simulation study, we also test the performance of least trimmed squares (LTS) (3). At the end of this step, we get the estimated $\widehat{K} \times 1$ vector of coefficients $\widehat{\beta}$, the estimated $n \times 1$ vector of residuals $\widehat{\epsilon}$, and the residual scale estimate $\widehat{\sigma}$.

## 2.4 Outlier detection

As a last step, we recover the vector $\widehat{O}$ of conditional outlier indices as the set of all the points $i \in \{1, \ldots, n\}$ with robustly rescaled residuals $\widehat{\epsilon}_i / \widehat{\sigma}$ larger than $\phi^{-1}(0.995)$ in absolute value.

DRAFT

## 3. Simulation study

In this section, we test the four possible variants of ROBOUT methodology based on the different variable selection and robust regression estimation options presented in Section 2. In the first step, either the SNCD-H or the SNCD-L objective function can be used, while in the second step either LTS or MM can be used to estimate the regression equation. We call the ensuing four versions of ROBOUT as H+LTS, L+LTS, H+MM and L+MM, where H and L refer to SNCD-H and SNCD-L, respectively.

Our simulation study aims both to compare ROBOUT to competitor methods but also to identify the optimal design of ROBOUT with respect to its constituent components. The competitor methods against which the ROBOUT versions are tested are SPARSE-LTS and RLARS.

All the methods are run with the default preprocessing step for all the potential predictors: standardization for SNCD, unit-norm normalization for SPARSE-LTS, robust standardization for RLARS. The intercept is included for all estimations.

### 3.1 Settings

The set of outliers $O$ is randomly generated with outlier proportion $\alpha = 0.1$. The set of predictors $D$ is randomly generated with $K = 3$. The intercept is $a = 10$. The coefficients are generated in the following way: $\beta_1, \beta_3 \sim U(10, 20)$, $\beta_2 \sim U(-20, -10)$, with residual variance $\sigma^2 = 1$. For each non-outlier index $i \notin O$, we assume by simplicity that $\epsilon_i \sim N(0, \sigma^2)$, $\mathbf{x}_{D,i} \sim N(\mathbf{0}_K, \mathbf{I}_K)$ and $\mathbf{x}_{D',i} \sim N(\mathbf{0}_{p-K}, \mathbf{I}_{p-K})$, where $\mathbf{x}_{D',i}$ is the vector of non-predictors ($D'$ stores the indices of non-predictor variables). For each outlier index $i' \in O$, we assume the following:

1) conditional outliers in mean are generated as $\epsilon_{i'} \sim N((m-1)a, \sigma^2)$, $m \in \mathcal{R}$, $m > 1$;
2) the vector of predictors is generated as $\mathbf{x}_{D,i'} \sim N(\mathbf{0}_K, \mathbf{I}_K)$, then $y_{i'}$ is generated by model 3, and in the end, leverage outliers are generated by replacing *a posteriori* $x_{D,i'k}$ with the perturbed values $x_{D,i'k} + (m-1) \times \text{sgn}(x_{D,i'k})$, $k = 1, \ldots, K$, $m \in \mathcal{R}$, $m > 1$.

The parameter $m > 1$ represents the degree of perturbation. We then generate the regression model for the single observation $i \notin O$ as

$$y_i = a + \mathbf{x}'_{D,i}\beta + \epsilon_i, \tag{3}$$

where, for each non-outlier index $i \notin O$, we assume by simplicity that $\epsilon_i \sim N(0, \sigma^2)$, $\mathbf{x}_{D,i} \sim N(\mathbf{0}_K, \mathbf{I}_K)$ and $\mathbf{x}_{D',i} \sim N(\mathbf{0}_{p-K}, \mathbf{I}_{p-K})$, while, for each outlier index $i' \in O$, we consider **case 1**, with conditional outliers in mean and leverage outliers, where the outliers are generated in $y_{i'}$ according to 1) and in $x_{D,i'k}$ according to 2). We consider $m = 1, 5, 9, 13, 17, 23, 29, 37, 45, 55$. In addition, we also allow for multicollinearity by setting $COV(x_{j'}, x_{j''}) = \rho_{j'j''}$, $\forall j', j'' \in \{1, \ldots, p\}$, $j' \neq j''$, where $x_j$ is the $j$-th variable in the $n \times p$ matrix $\mathbf{X}$, representing the available information set $\Omega$.

In this paper, we consider the following scenarios by the degree of multicollinearity in the matrix $\mathbf{X}$ and the relative size of the matrix ($p/n$ ratio). As regards the degree of multicollinearity, we set $\rho_{j'j''} = \rho$, $\forall j', j'' \in \{1, \ldots, p\}$, $j' \neq j''$, with $\rho = 0.3, 0.7$. As regards the relative size of $\mathbf{X}$, we examine two different cases for the $p/n$ ratio:

- case a: $p = 60$, $n = 300$ (i.e. $p/n = 0.2$, $p \ll n$);
- case b: $p = 100$, $n = 200$ (i.e. $p/n = 0.5$, $p < n$).

Henceforth, we refer to scenarios by combining the number and the letter of the above cases e.g. "scenario 1a, $\rho = 0.7$" refers to a scenario with case 1, case a for the $p/n$ ratio i.e. $p = 60$ and $n = 300$, $\rho = 0.7$. Each scenario is run 100 times for each value of $m$.

For each scenario, each value of parameter $m$ and each method, we derive the $F_1$ score for outlier detection, the empirical probability to recover all true predictors, the minimum recovery rate of single predictors across the true ones, and the maximum mean square error across the estimated coefficients of true predictors, whenever retrieved.

DRAFT

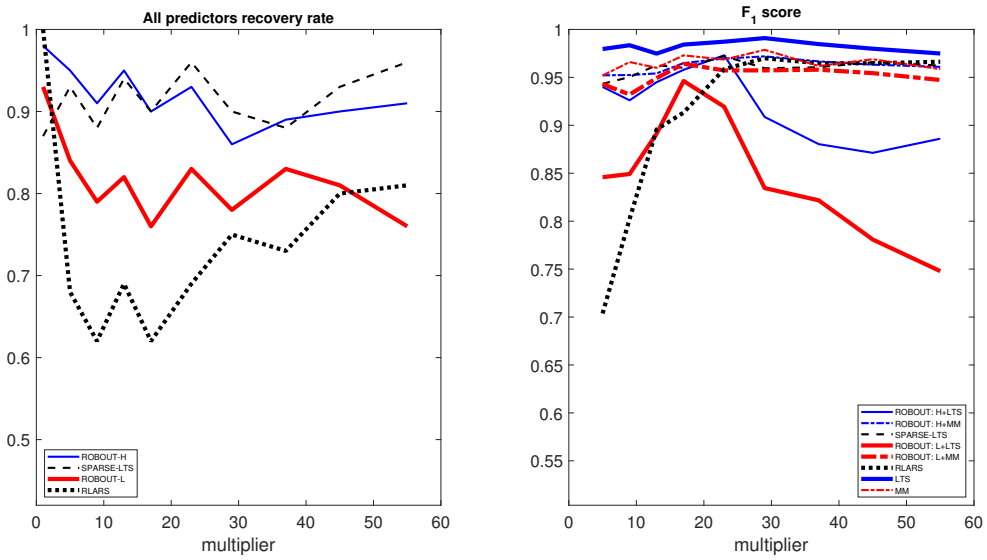



Figure 1: Average recovery proportion of all predictors (left panel) and $F_1$ score (right panel) of all considered methods under scenario 1a, $\rho = 0.7$.

## 3.2 Results

We note in Figure 1 that a clear hierarchy exists among methods under the most challenging scenario, 1a, $\rho = 0.7$. RLARS is very imprecise both concerning predictor recovery and outlier detection at small values of $m$. The same holds for the SNCD-L method on predictor recovery, and for H+LTS and L+LTS on outlier detection. On the contrary, H+MM, L+MM and SPARSE-LTS are doing well at outlier detection, and SPARSE-LTS and SNCD-H methods are doing reasonably well at predictor recovery.

Figure 2 shows that there are also clear levels of coefficient estimation quality: after the benchmarks, we find H+LTS and H+MM, L+LTS and L+MM, RLARS and SPARSE-LTS, which encounters difficulties in spite of the good performance in outlier detection and predictor recovery.

Under scenario 2a, $\rho = 0.7$ all ROBOUT options are doing very well both at outlier detection and predictor recovery, in comparison with competitors like SPARSE-LTS and RLARS, which are slightly less effective at small values of $m$. SPARSE-LTS coefficients are the worst by far; RLARS is doing reasonably well, approximately as ROBOUT options apart from H+MM, which works like the benchmarks in this case across all values of $m$.

Scenarios 1a, $\rho = 0.3$ and 2a, $\rho = 0.3$ present qualitatively similar results compared to their counterparts with $\rho = 0.7$.

## 4. Conclusions

In this paper, we have proposed a new step-wise methodology, called ROBOUT, to recover outliers in a dependent variable conditionally on its most relevant predictors retrieved from a high-dimensional dataset. ROBOUT comprises a preliminary imputation procedure, a robust predictor selection, a robust regression and an outlier detection step. The first step is designed to ensure that ROBOUT is robust against leverage outliers in the predictors.

We have shown in an ad-hoc simulation study that the combination of LASSO-penalized Huber regression to select predictors and MM regression to estimate coefficients and recover outliers is the most effective concerning predictor recovery, outlier detection, and coefficient estimation, also considering LASSO-penalized LAD regression and LTS as possible alternatives to Huber and MM regression, respectively.

DRAFT

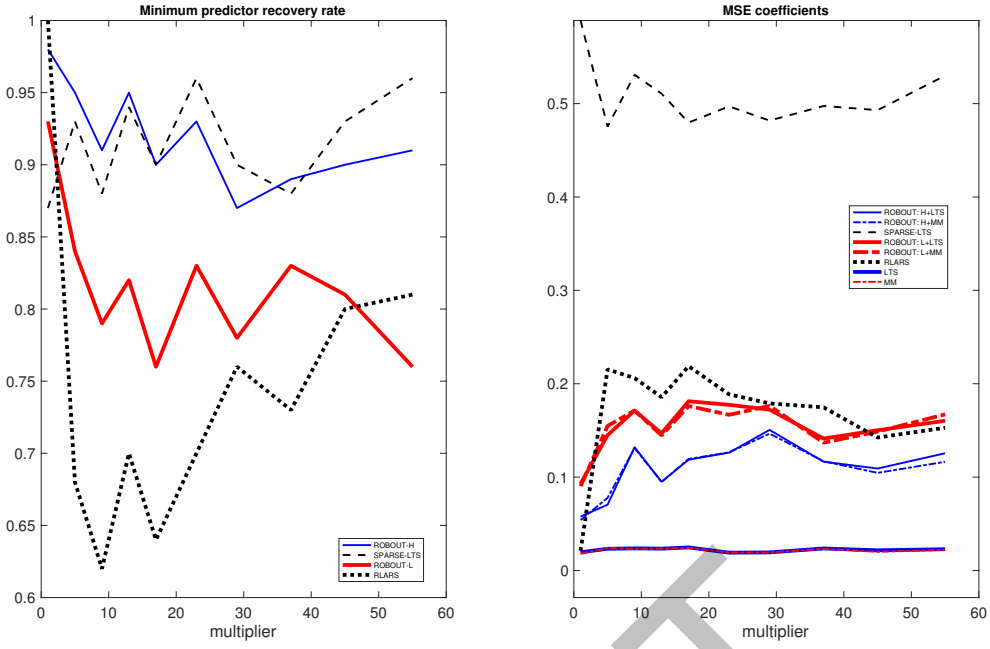



Figure 2: Minimum recovery rate of single predictors (left panel) and maximum mean square error of estimated coefficients (right panel) of all considered methods under scenario 1a, $\rho = 0.7$.

DRAFT

## References


[1] Alfons A., Croux C., Gelper S.: Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics. **1**, 226–48 (2013)

[2] Khan J.A., Van Aelst S., Zamar R.H.: Robust linear model selection based on least angle regression. Journal of the American Statistical Association. **102**(480), 1289–99 (2007)

[3] Rousseeuw P.J., Van Driessen K.: Computing LTS regression for large data sets. Data mining and knowledge discovery. **12**(1), 29-45 (2006)

[4] Tibshirani R., Bien J., Friedman J., Hastie T., Simon N., Taylor J., Tibshirani, R.J.: Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology). **74**(2), 245–266 (2012)

[5] Yi C, Huang J.: Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. Journal of Computational and Graphical Statistics. **26**(3), 547–557 (2017)

[6] Yohai V.J.: High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics. 642-56 (1987)

[7] Zou H., Hastie T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). **67**(2), 301–320 (2005)

# Robustness of the Efficient Covariate-Adaptive Design for balancing covariates in comparative experiments

Rosamarie Frieri[a], Alessandro Baldi Antognini[a], Maroussa Zagoraiou[a], and Marco Novelli[a]

[a]Department of Statistical Sciences, University of Bologna; rosamarie.frieri2@unibo.it, a.baldi@unibo.it , maroussa.zagoraiou@unibo.it, m.novelli@unibo.it

**Abstract**

Obtaining comparable groups in terms of the important covariates is a fundamental concern in comparative experiments. Indeed, potential imbalances of the covariates distribution across the groups could jeopardize the final inferential analysis and the validity of the study. In the sequential setting, especially within the field of clinical trials, Covariate-Adaptive randomization procedures are being increasingly used in practice, but most of them can accommodate only qualitative factors. The recently proposed Efficient Covariate-Adaptive Design [2] can incorporate also quantitative factors and it is high-order balanced. This study has the objective of exploring the robustness of this design, also in comparison with other procedures, in particular when i) some important covariates are omitted from the randomization and ii) when quantitative factors are transformed into categorical ones.

***Keywords:*** Loss of information, Mahalanobis distance, Treatment comparisons, Unobserved covariates, Discretized variables

DRAFT

## 1. Introduction

In the field of comparative experiments, especially in clinical trials, obtaining comparable experimental groups with respect to some important covariates represents a fundamental issue in order to guarantee reliable inference about the treatment effects [14]. In this regard, several Covariate-Adaptive (CA) designs have been proposed in the literature. These are sequential randomized procedures that, by considering the allocations and the characteristics of the past subjects, as well as those of the current one, change at each step the treatment assignment probabilities with the aim of reducing possible sources of heterogeneity between the experimental arms. The so-called minimization method [18; 13], for example, is aimed at minimizing a weighted sum of the marginal imbalances of allocations, while the stratified randomization approach exploits a separate randomization sequence within each covariate stratum [4; 17].

However, most of the proposed procedures can only deal with categorical variables, so quantitative covariates are at best discretized and at worst completely discarded. This strongly affects the inferential precision of the clinical trial [15; 5]. The few available exceptions (see, e.g., [1] or [11]) present a slow rate of convergence to balance and/or exhibit poor performances as the number of considered covariates grows. This clearly conflicts with the recent increase in data availability combined with the advances in biomarkers-based personalized medicine, which has made it increasingly common to include several covariates in the analysis, especially of continuous nature [9; 16]. Recently Baldi Antognini et al. [2]

introduce a new class of CA randomization called the Efficient Covariate-Adaptive Design (ECADE), which it is high-order balanced (namely, it converges to balance with the fastest available rate among other CA procedures) and can manage both quantitative and qualitative factors with a potentially complex interaction structure. The object of the present work is to explore in depth the operating characteristics of the ECADE proposed in [2] in several 'real-life' scenarios. Moreover, through an extensive comparison with other well-known CA rules proposed in the literature, the aim is to better understand i) the impact of discretizing continuous factors and ii) the effect of unobserved covariates in terms of information loss.

## 2. Model and Imbalance measures

Consider an experiment in which $n$ statistical units come sequentially and are assigned to one of two competing treatments, say $A$ and $B$. Suppose that for each experimental unit a $p$-dimensional vector of (qualitative and/or quantitative) covariates, $\boldsymbol{X} = (X_1, \ldots, X_p)^t$, can be observed before the treatment allocation assignment is made. In what follows $\delta_i$ is the treatment indicator variable of the $i$-th unit, with $\delta_i = 1$ if he/she is assigned to $A$ and 0 if $B$. After $n$ steps, $\boldsymbol{\delta}_n = (\delta_1, \ldots, \delta_n)^t$ is the vector of the treatment allocations and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the $p$-dimensional vectors of observed covariates. The outcomes $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^t$ are assumed to follow, at least approximately, a linear model

$$\mathbb{E}[\boldsymbol{Y}_n] = \boldsymbol{\delta}_n \mu_A + (\mathbf{1}_n - \boldsymbol{\delta}_n)\mu_B + \mathbb{Z}_n \boldsymbol{\beta}, \quad \text{var}(\boldsymbol{Y}_n) = \sigma^2 \mathbb{I}_n$$

where $\mathbf{1}_n$ is the $n$-dimensional vector of ones and $\mathbb{I}_n$ the $n$-dimensional identity matrix, $\mu_A$ and $\mu_B$ the treatment effects, $\boldsymbol{\beta}$ a $q$-dimensional vector of covariate effects, $\mathbb{Z}_n = [\boldsymbol{f}(\boldsymbol{X})^t]$ a $n \times q$ matrix of covariate effects with $\boldsymbol{f}$ being a known vector function that may include interactions among the covariates, (i.e., when $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{X}$, $p = q$ but generally $q > p$) and $\sigma^2$ is the common variance. Let us also define $\mathbb{Q}_n = n^{-1} \begin{bmatrix} n & \mathbf{1}_n^t \mathbb{Z}_n \\ \mathbb{Z}_n \mathbf{1}_n^t & \mathbb{Z}_n^t \mathbb{Z}_n \end{bmatrix}$, a square matrix of size $q+1$.

In this set-up, Atkinson in [1] proposed a widely used measure of covariate imbalance, the so-called loss of information,

$$L_n = n^{-1} \boldsymbol{b}_n^t \mathbb{Q}_n^{-1} \boldsymbol{b}_n, \tag{1}$$

where $\boldsymbol{b}_n^t = (D_n; (2\boldsymbol{\delta}_n - \mathbf{1}_n)^t \mathbb{Z}_n)$ is the so-called *imbalance vector* and $D_n = 2\boldsymbol{\delta}_n^t \mathbf{1}_n - n$ is the difference between the assignments in the two arms. Here, $L_n$ measures the loss of estimation precision induced by the covariate imbalance after $n$ patients, whereas $L_n/n$ is the corresponding loss of estimation efficiency. Under this framework, both the estimation accuracy and the power of the Wald test are strictly related to the loss and are maximized when $L_n$ vanishes [4; 2].

By letting $\bar{\boldsymbol{z}}_{An} = \mathbb{Z}_n^t \boldsymbol{\delta}_n / \mathbf{1}_n^t \boldsymbol{\delta}_n$ and $\bar{\boldsymbol{z}}_{Bn} = \mathbb{Z}_n^t (\mathbf{1}_n - \boldsymbol{\delta}_n)/(n - \mathbf{1}_n^t \boldsymbol{\delta}_n)$ be the $q$-dimensional vectors collecting the sample means in the two experimental arms, in the causal inference framework, Morgan and Rubin [12] proposed the Mahalanobis distance between $\bar{\boldsymbol{z}}_{An}$ and $\bar{\boldsymbol{z}}_{Bn}$ as an alternative measure of covariate imbalance, namely

$$M_n = (\bar{\boldsymbol{z}}_{An} - \bar{\boldsymbol{z}}_{Bn})^t \, \text{var}(\bar{\boldsymbol{z}}_{An} - \bar{\boldsymbol{z}}_{Bn})^{-1} (\bar{\boldsymbol{z}}_{An} - \bar{\boldsymbol{z}}_{Bn}).$$

## 3. The Efficient Covariate-Adaptive Design (ECADE) and other CA procedures

The ECADE is a CA randomization procedure based on the sequential minimization of the weighted Euclidean norm of the imbalance vector $||\boldsymbol{b}_n||_{\mathbb{W}} = \sqrt{\boldsymbol{b}_n^t \mathbb{W} \boldsymbol{b}_n}$ with respect to a weight matrix $\mathbb{W}$ (symmetric and positive-definite). The $n$-th subject, with covariate profile $\boldsymbol{x}_n$, is assigned to treatment $A$ with probability

$$P(\delta_n = 1 | \boldsymbol{\delta}_{n-1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \eta((1; f(\boldsymbol{x}_n)^t) \mathbb{W} \boldsymbol{b}_{n-1}) \tag{2}$$

DRAFT

where $\eta : \mathbb{R} \longrightarrow (0,1)$ is a decreasing (strictly decreasing at 0) symmetric function such that $\eta(x) = 1-\eta(-x)$. Here the argument of $\eta(\cdot)$ in (2) is proportional to the difference between the squared norm of the potential imbalances that one would obtain by assigning subject $n$ to $A$ and $B$ (see [2]).

Under the ECADE, for qualitative covariates, $\boldsymbol{b}_n = O_p(1)$ and the loss and the Mahalanobis distance are asymptotically equivalent with an order $o_p(1)$ of convergence to balance. The same result holds for quantitative and mixed covariates, provided that $\lim_{x\to\infty}\eta(x) = \varepsilon \in (0,1/2)$.

Several choices for the weight matrix can be adopted. The ECADE encompasses also weights that change at each step $n$, adopting $\mathbb{W}_n$ instead of $\mathbb{W}$ in (2). All the theoretical properties of the ECADE are preserved provided that $\mathbb{W}_n$ is a sequence of symmetric and positive-definite matrices such that $\mathbb{W}_n \longrightarrow \mathbb{W}$ almost surely, as $n$ grows [2]. In the rest of this paper, we implement the ECADE by setting $\mathbb{W}_n = \mathbb{Q}_n^{-1}$.

With regards to the function $\eta$, we will consider i) $\eta_E(x) = 1/2 - sign(x)(\rho - 1/2)$ with $\rho \in (1/2,1)$, namely Efron's allocation function [6] and ii) $\eta_\varepsilon(x) = \varepsilon + (1-2\varepsilon)[1-\Phi(x)]$ where $\Phi(\cdot)$ is the standard normal cdf.

In the following analysis, among the CA designs that can be implemented for both qualitative and quantitative covariates, we consider the $D_A$-BCD by Atkinson (ATK) [1] and the CA randomization based on kernel densities by Ma and Hu (KER) [11]. With regards to the procedures that apply only to categorical covariates we consider the minimization method by Pocock and Simon [13] (PS) as well as the design by Hu and Hu [8] (HH), whose allocation probability is based on the Efron's function.

## 4. Numerical Comparisons

To illustrate the finite sample properties of the ECADE and to compare it to other CA procedures, we base our simulation study on the experimental setting in [7]. They report the results of the randomized clinical trial NIDA-CSP-1019 with the aim of testing the treatment effect of the *selegiline transdermal system*, a treatment of cocaine dependence, against a placebo. The trial has enrolled 300 patients and, as reported in the paper, the treatment groups were balanced with a *biased coin* procedure with respect to gender (GEN: male, female), presence of attention deficit disorder (DEFDIS: 0 absence, 1 presence), historical self-report of cocaine use in the past 30 days (COCUSE: quantitative discrete with values in 1-30) and severity of depression calculated by Hamilton Depression Rating Scale (DEPRE: quantitative discrete with values 0-44, i.e., not depressed to very severe depression). We re-design this study by also including in the randomization procedure the age of patients (AGE: measured in years) as well as the study center (CLIN, discrete with 16 levels). Thus, six covariates are considered: three qualitative, two of which are dichotomous and one with 16 levels, and three quantitative factors.

The first simulation has the aim of studying the effect of balancing the observed covariates in mitigating the potential imbalance of other covariates not observed before randomization. This issue has been recently discussed by [10], but only for categorical factors. Indeed, while for categorical covariates a possible measure to assess unobserved covariates imbalances can be based, for instance, on the marginal imbalance, when also continuous variables are included another strategy has yet to be defined [10]. In this regard, here we propose a criterion that is based on a two-step procedure described as follows. In the first step, the trial is simulated including all the six covariates in the CA randomization, obtaining $L_n$ and $M_n$. In the second step, the same trial is simulated excluding CLIN and COCUSE from the randomization procedure. This results in a new treatment allocation vector $\tilde{\boldsymbol{\delta}}_n$ and then in an imbalance vector $\tilde{\boldsymbol{b}}_n$. Thus the loss in the case of unobserved covariates is $\tilde{L}_n = n^{-1}\tilde{\boldsymbol{b}}_n^t\mathbb{Q}_n^{-1}\tilde{\boldsymbol{b}}_n$; the Mahalanobis distance $\tilde{M}_n$ is computed accordingly. The rationale is that CLIN was not included in the original study and the data on the cocaine use could be unreliable or faked by the patients. For the covariate effect we restrict our attention to the case $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{X}$. The results on the losses and Mahalanobis distances of the competing procedures, based on 10000 simulations, are displayed in Figure 1. Here ECADE$_E$ and ECADE$_{0.1}$ refer to the adoption of $\eta_E$ (with $\rho = 0.8$) and $\eta_\varepsilon$ (with $\varepsilon = 0.1$), respectively and the biasing probability for the KER procedure is 0.8. In addition, the superscript $u$ refers to the unobserved covariates case and thus to $\tilde{L}_n$ and $\tilde{M}_n$ of the corresponding CA design. We first notice the severe inflation in the loss of estimation precision when two covariates are excluded from the CA randomization. More specifically,

DRAFT

the curves of ATK$^u$ and KER$^u$ do not seem to have a decreasing trend with $n$, while a slightly decreasing behavior is observed for the ECADE$^u$, which is associated with smaller values of the loss ($\tilde{L}_n$ is around 16 for the ECADE while is around 22-23 for ATK and KER). No evident differences as $\eta$ changes are observed. When all the covariates are included in the randomization procedure, the KER induces a loss that essentially doubles the one under ATK for $n = 300$. The ECADE is the best performing CA procedure (for $n > 100$) with a slight difference in ECADE$_E$ and ECADE$_{0.1}$ whose $L_{300}$ are equal to 2.2 and 1.3, respectively, against the loss of ATK which is 4.9. Similar considerations can be drawn by looking at the Mahalanobis distance.

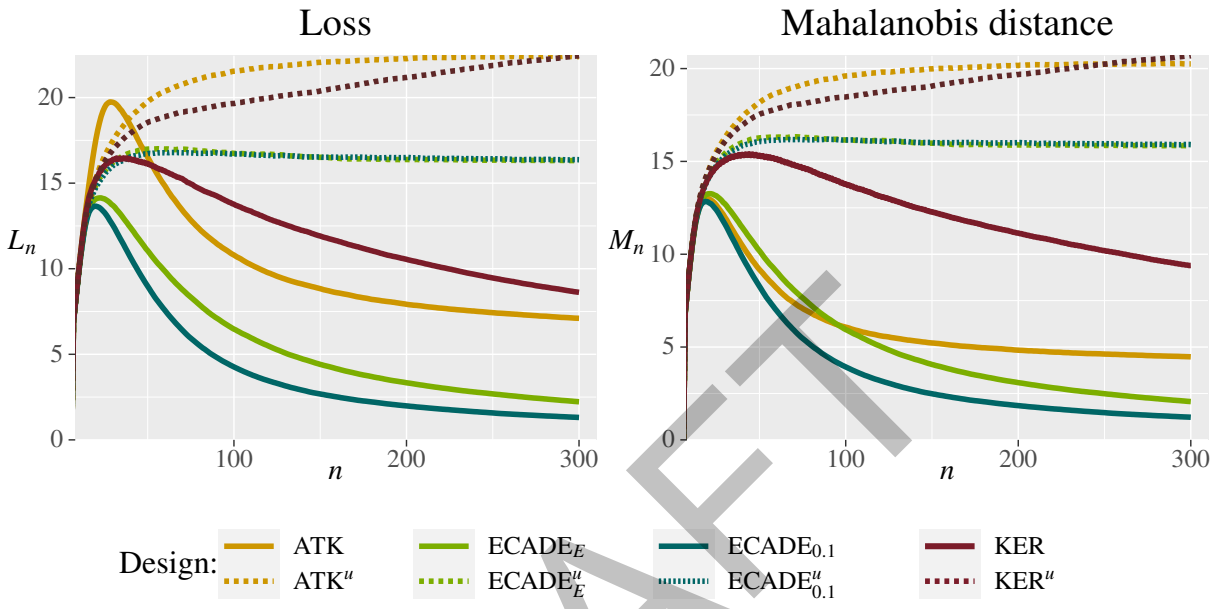



Figure 1: Loss of estimation precision $L_n$ and Mahalanobis distance $M_n$ for cocaine dependence clinical trial as $n$ varies. The superscript $u$ refers to the procedures in which two covariates have been omitted from the randomization procedure.

DRAFT

As most of the CA randomization procedures only deal with qualitative factors, a standard practice is to break down the numerical covariates into two or more categories. However, as it is well known [15; 5; 2], this operation does not come without a cost, as it induces a loss of information. To illustrate such an effect, we focus on the inflation in the loss of estimation precision and Mahalanobis distance induced by discretizing COCUSE into 3 categories (1:[0-10], 2:[11-20], 3:[20-30]), DEPR into 5 levels (1:[0-7], 2:[8-13], 3:[14-18], 4:[19-22], 5:[$>$ 23]) and AGE into 4 levels (1: [18-30], 2: (30-40], 3: (40-50], 4:($>$ 50)), as described in [10]. In this study, besides the $D_A$-BCD, the ECADE$_E$ and KER (both with $\rho = 0.85$) which can be implemented with any kind of covariates, we consider PS design and HH (with biasing probability equal 0.85). The simulation results (10000 iterations) are reported in Figure 2, where the superscript $d$ refers to CA procedures that are implemented by discretized covariates. The PS method which is known to be strongly inadequate in the presence of a large number of covariates, has even worst performance in this setting. The HH$^d$ presents an almost constant trend as $n$ increases, with very large values of $L_n$. For discretized factors the ECADE$^d$ is still the best choice for $n > 60$, inducing a loss even smaller than that of ATK, for $n$ greater than 180. Finally the best procedure is the ECADE in which no subjective operation of categorization is performed and the nature of the covariates is preserved. Also in this case, $L_n$ and $M_n$ present very similar behavior for any procedure.

## 5. Discussion

Since the true relationship between study outcome and covariates is not generally known, it's fundamental to assess the performance of Covariate-Adaptive procedures when some important covariates are

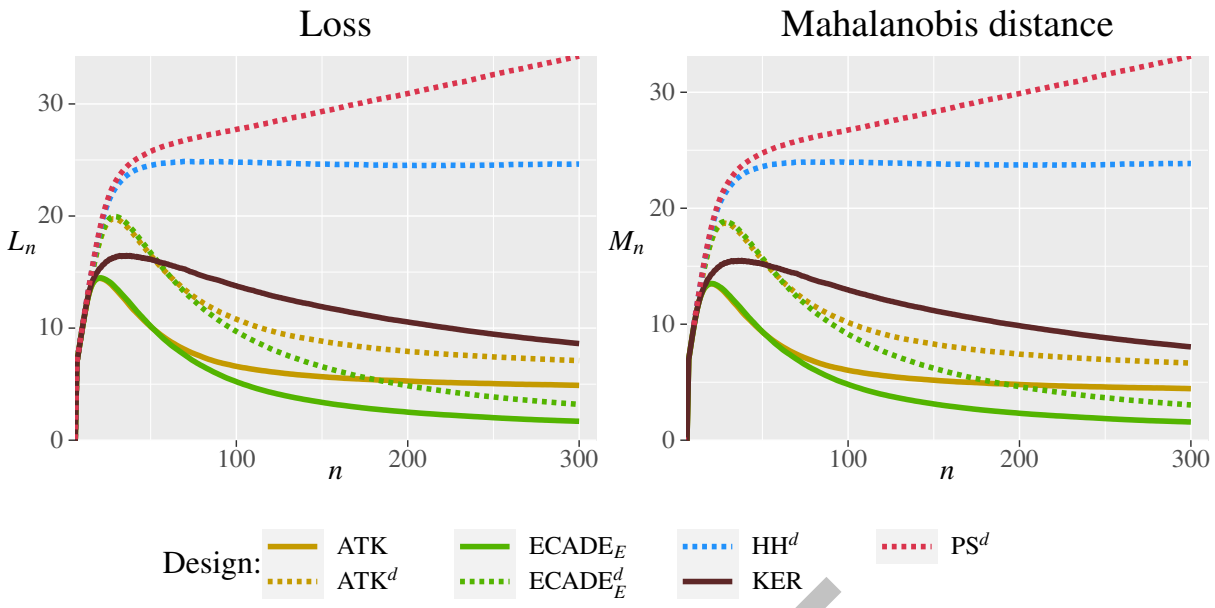



Figure 2: Loss of estimation precision $L_n$ and Mahalanobis distance $M_n$ for cocaine dependence clinical trial as $n$ varies. The superscript $d$ refers to CA procedures that are implemented by discretized quantitative covariates.

DRAFT

not included in the randomization procedure. In practice, this is a measure of robustness of the randomization procedure to a form of misspecification. Other forms of misspecifications can occur, for instance, when some covariates effects (e.g., quadratic or interaction terms) are not included in the randomization procedures (see [2]).

In addition, note that the impact of categorizing a continuous variable strongly depends on the chosen cutoff (see [2]). On this purpose, it is customary to take quantiles of the covariate-generating distribution to break down continuous covariates. However, we wish to point out that such distribution is generally unknown in sequential studies, making even more serious the potential consequences of categorizing quantitative variables.

Finally, it is worth noticing that methods for balancing covariates could be also successfully employed in enrichment designs, as biomarkers can be treated as covariates from a mathematical viewpoint. In an effort towards optimal designs for enrichment trials, Covariate-Adaptive randomization procedures could be suitably adjusted to fit an enrichment clinical study [3].

## References

[1] Atkinson, A.C.: Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* 69, 61–67 (1982).

[2] Baldi Antognini, A., Frieri, R., Zagoraiou, M. and Novelli, M.: The Efficient Covariate-Adaptive Design for high-order balancing of quantitative and qualitative covariates. *Stat Papers*, (2022). DOI: 10.1007/s00362-022-01381-1.

[3] Baldi Antognini, A., Frieri, R., Zagoraiou, M.: New insights into adaptive enrichment trials. *Stat Papers*, (2023). DOI: 10.1007/s00362-023-01433-0.

[4] Baldi Antognini, A. and Zagoraiou, M.: The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika* 98, 519–535 (2011).

[5] Ciolino J., Zhao W., Martin R. et al. Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. *Contemp Clin Trials* 32(2):250–259 (2011).

[6] Efron B.: Forcing sequential experiments to be balanced, *Biometrika*, 58: 403–417 (1971).

[7] Elkashef A., Fudala P.J., Gorgon L., Li S., Kahn R., Chiang N., Vocci F., Collins J., Jones K.,

Boardman K., Sather M.: Double-blind, placebo-controlled trial of selegiline transdermal system (STS) for the treatment of cocaine dependence *Drug and Alcohol Dependence* 85(3): 191–197 (2006).

[8] Hu Y. and Hu F.: Asymptotic properties of covariate-adaptive randomization *Ann Stat* 40:1794–1815 (2012).

[9] Karczewski, K. J. and Snyder, M. P.: Integrative omics for health and disease. *Nat Rev Genet* 19.5, 299 (2018).

[10] Liu Y, and Hu F.: Balancing unobserved covariates with covariate-adaptive randomized experiments, *JASA*, 117(538): 875–886 (2022).

[11] Ma, Z. and Hu, F.: Balancing continuous covariates based on kernel densities. *Contemp Clin Trials* 34(2):262–269 (2013).

[12] Morgan K.L. and Rubin D.B.: Rerandomization to improve covariate balance in experiments. *Ann Stat* 40(2): 1263–1282 (2012).

[13] Pocock, S. J. and Simon, R.: Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Clin Cancer Res* 31, 103–115 (1975).

[14] Rosenberger, W. F. and Sverdlov, O.: Handling covariates in the design of clinical trials. *Stat Sci*, 23, 404–419 (2008)

[15] Royston P., Altman D.G. and Sauerbrei W.: Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med* 25(1): 127–141 (2006).

[16] Rudrapatna, V.A., Butte, A.J. and others: Opportunities and challenges in using real-world data for health care. *J Clin Invest*, 130(2), 565–574 (2020).

[17] Scott, N.W., McPherson, G.C., Ramsay, C.R. and Campbell, M.K.: The method of minimization for allocation to clinical trials: a review. *Control Clin Trials* 23.6, 662–674 (2002).

[18] Taves, D. R.: Minimization: a new method of assigning patients to treatment and control groups. *J Clin Pharm Ther* 15, 443–453 (1974).

DRAFT

# Separation scores: a new statistical tool for scoring and ranking partially ordered data

Marco Fattore[a]

[a]Department of Statistics and Quantitative Methods - University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1 - 20126; marco.fattore@unimib.it

**Abstract**

We introduce a new statistical procedure for scoring and ranking partially ordered data derived from ordinal multi-indicator systems, improving over approaches based on so-called "dominance scores". The procedure draws upon a novel matrix representation of partial order relations ("separation matrix") that provides a fuzzy measure ("separation score") of the number of elements in-between two given elements of a partially ordered set. We introduce the new procedure through a real example pertaining to Political Pluralism and Participation in Middle East countries, showing how the new scores effectively reflect the differences in the political freedom levels of the statistical units.

***Keywords:*** Multi-indicator system, Partial order, Politics; Ranking, Synthetic indicators.

DRAFT

## 1. Introduction

The problem of computing synthetic indicators and building rankings from multidimensional systems of ordinal indicators is still an open issue, although some advances have been made in the last years, in particular by using partially ordered sets (*posets*) as the basic data structure and by exploiting concepts and tools from *order theory* (5; 11), to design and implement proper scoring algorithms. Most of the proposals available in the literature employ, in different ways, so-called *dominance scores*, which can be seen as a way to measure to what extent an element of the input poset "tends" to globally dominate the other elements (2; 8; 9). Dominance scores are usually computed as the *mutual ranking probabilities* (6) between the poset elements; these, in turn, require the linear extensions (11) of the poset to be built, somehow limiting the complexity of the partially ordered sets that can be effectively treated, although some smart algorithms and approximation formulas, reducing the computational burden, are available in the literature (3; 4; 6; 7). Although the final ranking produced by dominance scores is in many cases reasonable and reliable, the numerical differences between the ranking scores are not that effective in revealing the corresponding differences between the overall achievement levels of the compared units. This motivates the construction of *separation* scores which exploit more effectively the information comprised in the indicator system, incorporating a kind of "metric" information (speaking with some language abuse) on the relative positions of the elements of the poset. Interestingly, separation scores can be easily obtained from the same inputs used to build the dominance scores, so adding just a small computational overhead to the scoring and ranking process. The short paper is organized as follows. Section 2 describes the data on Political Pluralism and Participation, used to introduce the new scores; Section 3 briefly recalls the essentials behind the computation of dominance scores; Section 4 motivates and develops the computation of separation scores; Section 5 concludes and provides some hints for future research.

## 2. The Political Pluralism & Participation poset

We introduce separation scores by working out a real example, based on data extracted from the *Freedom in the World 2022* dataset (available on https://freedomhouse.org) on the rights and freedoms enjoyed by individuals, across the world. The dataset assesses 195 countries and 15 territories on three pillars relative to political rights, namely (A) *Electoral Process*, (B) *Political Pluralism and Participation* and (C) *Functioning of Government*, and on four pillars relative to civil liberties, i.e. (D) *Freedom of Expression and Belief*, (E) *Associational and Organizational Rights*, (F) *Rule of Law* and (G) *Personal Autonomy and Individual Rights*. Here we focus on the *Political Pluralism and Participation* (PP&P) pillar, which is evaluated by considering the following four variables, each scored on a 0-4 ordinal scale: (V1) *Right of people to organize in different political parties*, (V2) *Opportunity for the opposition to increase its support or gain power through elections*, (V3) *Freedom of people's political choices from domination by external forces* and (V4) *Full political rights and electoral opportunities for relevant social groups and population segments*. For the sake of simplicity, and to allow for simpler graphical representations, here we consider only the 15 Middle East countries/territories. Each of them is attached a 4-component PP&P profile, comprising the scores on the V1-V4 variables; the set of profiles is then turned into a partially ordered set, denoted by $\pi$, by comparing each of them to the others componentwise. The Hasse diagram depicted in Figure 1 provides a visual representation of the resulting partially ordered set, together with country/territory labels. As it can be seen, some countries share the same score configuration, so that only 11 unique profiles are realized in the dataset, out of 15 units. The PP&P poset has both *top* (profile 3442, that dominates all of the other profiles) and *bottom* (profile 0000, which is dominated by all of the other profiles), comprises 42 comparabilities and 13 incomparabilities (i.e. pairs of profiles that cannot be reciprocally ordered, due to conflicting scores); its *height* (number of edges in the longest *chain*) is equal to 6 and its *width* (cardinality of the largest *antichain*) is equal to 3.

DRAFT

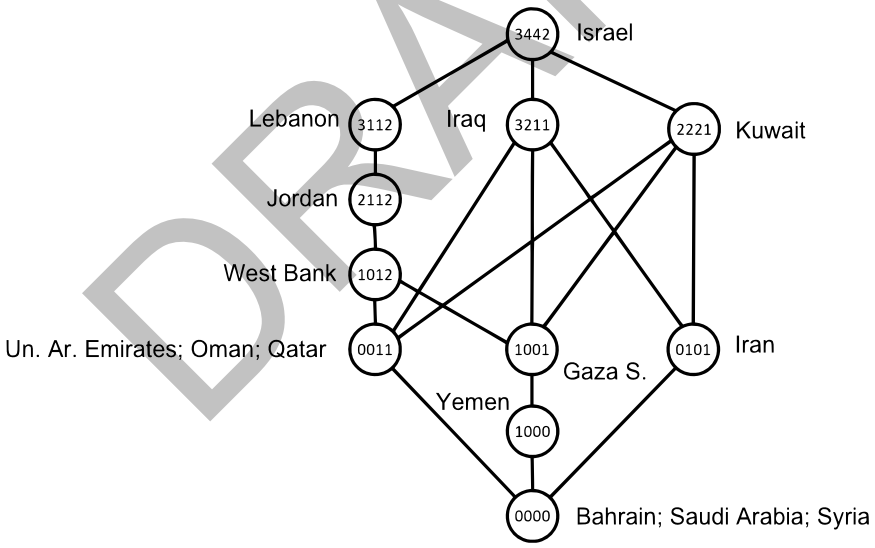



Figure 1: Hasse diagram of the PP&P profiles ordered componentwise, with country labels. Profile components refer to the variables V1-V4, each scored on a 0-4 ordinal scale, introduced in the main text (the higher the score, the better the achievement on the corresponding variable).

## 3. Dominance and incomparability scores

Our goal is to turn the PP&P partial order into a ranking, i.e. into a *linear order* $\ell$ respecting the dominances of the input poset $\pi$, so that $x_i < x_j$ in $\pi \Rightarrow x_i < x_j$ in $\ell$, for any pair of comparable profiles $x_i$ and $x_j$. The procedure to achieve this, fully respecting the ordinal nature of the variables and avoiding the inconsistent aggregation of ordinal scores, is described in (8; 9) and will be only sketched

here. At its heart is the computation of so-called *pairwise dominance scores* ($pdom_{ij}$) between all pairs of profiles $x_i$ and $x_j$ expressing, in a fuzzy spirit, the degree of dominance of $x_j$ over $x_i$. In principle, these scores are computed as the fraction of linear extensions of $\pi$ (i.e. linear orders preserving the dominances in $\pi$ (11)), where $x_i$ is dominated by $x_j$, i.e. as

$$pdom_{ij} = \frac{|\{\lambda \in \Omega(\pi) : x_i <_\lambda x_j\}|}{|\Omega(\pi)|}. \tag{1}$$

where $\Omega(\pi)$ is the set of linear extensions of $\pi$. In practice, however, this is computationally unfeasible, but for very simple posets, and alternative formulas must be employed, to compute the dominance scores directly based on some simpler quantities describing the mutual "relational positions" of the compared elements, in the poset. Here, in particular, we use the Brueggemann-Lerche-Sørensen (*bls*) formula proposed in (1), defined by:

1. For $x_i < x_j$, $pdom_{ij} = 1$ and $pdom_{ji} = 0$;
2. For $x_i$ incomparable with $x_j$, $pdom_{ij} = a_{ij}/(a_{ij} + a_{ji})$ where

$$a_{ij} = \frac{|Up(i,j)| + 1}{|Down(i,j)| + 1} \tag{2}$$

with

$$Up(i,j) = \{x_h : x_i < x_h\} \cap \{x_h : x_j < x_h\}^c \tag{3}$$

and

$$Down(i,j) = \{x_h : x_h < x_i\} \cap \{x_h : x_h < x_j\}^c. \tag{4}$$

Pairwise dominance scores lies in the interval $[0, 1]$, equal 1 if and only if $x_i < x_j$ in the input poset and satisfy the equality $pdom_{ij} = 1 - pdom_{ji}$. They are collected in the *dominance* matrix $D$, whose main diagonal is by definition set to 0. The pairwise dominance scores associated to each element $x_j$ can then be aggregated (e.g. by simple or weighted means) into a synthetic dominance score $dom_j$, used to build the final ranking. Looking at the Hasse diagram of $\pi$, however, one notices that profile 3211 directly dominates (technically, *covers*) profile 1001, since no intermediate profiles are observed in the dataset, and that this occurs although the former represents a much better configuration score than the latter. An analogous situation involves, for example, profiles 2221 and 0101. To avoid making the dominance scores of such pairs of profiles too similar, we embed the observed PP&P poset into the componentwise poset $\pi^{all}$ composed of all possible $5^4 = 625$ profiles, built upon the variables V1-V4, and compute the pairwise, and the synthetic dominance scores, taking into account the existence of both observed and non-observed profiles, so better discriminating the units. Clearly, turning a partially ordered set into a linearly ordered one, i.e. turning incomparabilities into comparabilities, can indeed be forcing because incomparable profiles get put on the same "low-high" axis, notwithstanding their inherently different structure. To account for this, dominance scores are complemented with incomparability scores $inc_j$ in the $[0, 1]$ interval, assessing to what extent profiles tend to be non-comparable to the others, in the input partial order (10). These additional scores are computed by first defining *pairwise incomparability scores* as $inc_{ij} = min(dom_{ij}, dom_{ji})$ and then aggregating and normalizing them, analogously to the dominance case. Table 1 reports the synthetic incomparability and dominance scores for the PP&P profiles, also graphically shown in the left panel of Figure 2. As it can be seen, PP&P profiles are spread along the vertical dominance axis, with Israel at the top and Bahrain, Saudi Arabia and Syria at the bottom, as expected. Notice also that profiles are in general only "moderately" incomparable (the highest incomparability score is 0.235), so that the final ranking can be taken as quite meaningful.

DRAFT

## 4. Separation matrix and separation scores

The generic entry $D_{ij}$ of the dominance matrix $D$ can be interpreted as a measure of the strength of the statement $x_i < x_j$ i.e., in a fuzzy spirit, as the membership of element $x_i$ to the downset $x_j{\downarrow}$ of $x_j$,

Table 1: Synthetic incomparability, dominance and separability scores, for the PP&P profiles (aggregated by simple averages and normalized to [0,1]).

| Profile | 3442 | 2221 | 3211 | 3112 | 2112 | 1012 | 0101 | 1001 | 0011 | 1000 | 000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *inc* | .000 | .231 | .231 | .199 | .058 | .015 | .235 | .200 | .226 | .052 | .000 |
| *dom* | 1.000 | .783 | .784 | .801 | .629 | .498 | .292 | .300 | .287 | .126 | .000 |
| *sep* | .783 | .245 | .247 | .247 | .164 | .050 | .007 | .007 | .007 | .001 | .000 |

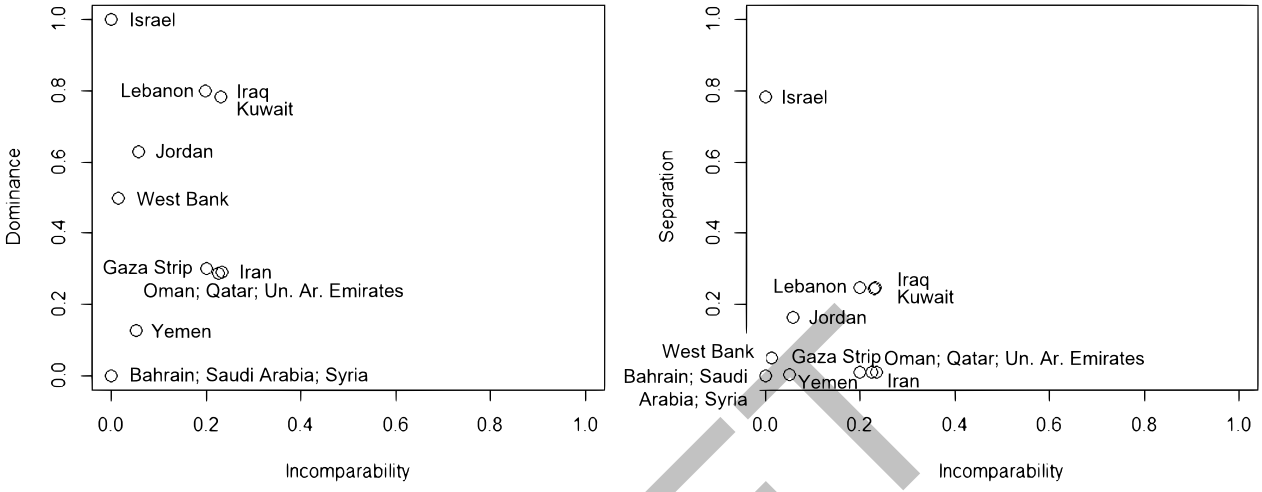



Figure 2: Scatterplot of the incomparability and dominance (left panel) and separation (right panel) scores, given in Table 1, with country labels

in the poset $\pi$. As a consequence, the sum of the entries of the $j$-th column of $D$ can be interpreted as the fuzzy cardinality of $x_j\downarrow$ and the dominance score $dom_j$ as a measure of the "effective" number of elements below $x_j$, normalized to [0,1]. As it can be observed in the left panel of Figure 2, the distribution along the vertical axis of the dominance scores of the countries is quite homogeneous, notwithstanding the fact that some profiles, although "near" in the Hasse diagram of the input poset, express very different PP&P levels. This is the case, for example, of Israel and Kuwait which are both placed in the upper part of the scatterplot. The non-complete effectiveness of the dominance scores to fully account for the "distance" (informally speaking) between observed PP&P profiles, even if embedded in the complete poset $\pi^{all}$, can be understood by observing that the pairwise scores $pdom_{ij}$ do not account, explicitly, for the number of profiles of $\pi^{all}$ in-between $x_i$ and $x_j$ in the linear extensions of $\pi^{all}$, but only for the number of these where $x_i$ is dominated by $x_j$. This "metric" information is recovered, only partially, when pairwise dominance scores $pdom_{ij}$ are summed (or averaged) into $dom_j$ and in fact the latter can be proved to be linked to the average height of $x_j$, in the set of linear extensions of $\pi^{all}$ (6). To increase the discriminating power of the final ranking scores, we thus want them to directly incorporate the information on the quantity of profiles separating the compared elements. In a poset, however, the concept of "in-betweeness" is not trivial to define, since two incomparable elements $x_i$ and $x_j$ have no element $x_k$ such that $x_i < x_k < x_j$ (otherwise, by transitivity, $x_i < x_j$ would hold, as well). In a fuzzy spirit, however, we can attach to any $x_k$ a score $inb_{ikj}$ of in-betweeness given $x_i$ and $x_j$, by observing that $dom_{ik}$ and $dom_{kj}$ express the strengths of the statements $x_i < x_k$ and $x_k < x_j$ respectively, and by taking their product, i.e. by putting $inb_{ikj} = dom_{ik}dom_{kj}$. As a consequence, the *pairwise separation score*

$$psep_{ij} = 1 + \sum_{k=1}^{n} inb_{ikj} = 1 + \sum_{k=1}^{n} dom_{ik}dom_{kj} \tag{5}$$

(where $n$ is the number of poset elements) can be seen as the "effective" quantity of elements in-between $x_i$ and $x_j$ (the added 1 is necessary, since by definition the diagonal pairwise dominance scores $pdom_{hh}$

DRAFT

are null). Notice that $psep_{ij}$ is not symmetric in $i$ and $j$ and that, if $x_i$ and $x_j$ are incomparable, both $psep_{ij}$ and $psep_{ji}$ are non-null. By computing $psep_{ij}$ for each pair of poset elements and arranging them into a matrix, we get the *separation matrix* $S$, easily computed from the dominance matrix $D$ as

$$S = (I + D)D. \tag{6}$$

By summing the entries of the generic $j$-th column of $S$, and normalizing, we finally get the *separation score* $sep_j$ ($j = 1, \ldots, n$) which lies in $[0, 1]$. Formally separation scores are defined as

$$sep_j = \frac{\sum_{i=1}^{n} psep_{ij}}{(n-1) \cdot psep_{\perp\top}}. \tag{7}$$

where $psep_{\perp\top}$, used for normalization, is the pairwise separation score between the top $\top$ and the bottom $\perp$ of the complete poset $\pi^{all}$ (respectively, profiles 4444 and 0000, in the PP&P example). The right panel of Figure 2 shows the scatterplot incomparability vs. separability scores (reported in Table 1). Compared to the left plot, it is now evident how Israel is neatly separated from the other countries, reflecting its much better achievements on the PP&P variables. At the same time, the other countries are "compressed" into the lower part of the plot, since their achievement profiles are more similar and less discriminating, in terms of political freedom. Notice also that in the right plot Israel has a score lower than 1, because the theoretical maximum separation score is assigned to the top of the *complete* poset $\pi^{all}$, i.e. to profile 4444. This shows how the information provided by the evaluation context, i.e. by the full poset of all possible PP&P profiles, anchors the ranking scores to a kind of measurement scale, although (and interestingly) this one does not come as a linear order of scores (i.e. as a unidimensional ordinal variable), but as a partially ordered set of profiles. In summary, incomparability, dominance and separation scores provide a comprehensive toolbox for investigating ordinal multi-indicator systems, for scoring and ranking purposes, accounting for both the "horizontal" and the "vertical" dimension of the evaluation process, when set in a partially ordered context.

DRAFT

## 5. Conclusion

In this paper we have proposed a new tool for scoring and ranking units assessed against ordinal multi-indicator systems, based upon the separation matrix representation of finite partial orders. The tool allows for incorporating into the scoring and ranking process a kind of "metric" information, extracted from the evaluation context where the observed data are embedded, so as to better reflect the difference in the achievement patterns of the input profiles, complementing the information provided by the incomparability and the dominance scores. Importantly, the computation of all of these scores can be made quite light, by using proper formulas for the construction of the pairwise dominance matrix, based on simple features of the order relation, with no need to generate the set of linear extensions of the complete evaluation poset. The toolbox is currently under software implementation and will be made available, both in the `R` and Python programming languages. Some further research is indeed due. Firstly, the computations introduced in the previous paragraphs must be modified to take into account the frequency distribution of the population or sample, on the set of profiles, so providing a "weighted version" of incomparability, dominance and separation scores. Secondly, a way to introduce variable importance, without using numerical weights in an ordinal context, must be worked out, to improve the flexibility of the approach and its fit to the practical needs of decision-makers. Both research avenues are currently under investigation and their results will be integrated in the toolbox and its software implementation.

## References

[1] Brueggemann, R., Lerche, D.B., Sørensen, P.B.: First attempts to relate structures of Hasse diagrams with mutual probabilities. In Sørensen, P.B., Brueggemann, R., Lerche, D.B., Voigt, K. Welzl, G., Simon, U., Abs, M., Erfmann, M., Carlsen, L., Gyldenkærne, S., Thomsen, M., Fauser,

P., Mogensen, B.B., Pudenz, S., Kronvang, B. (eds.) Order Theory in Environmental Sciences. Integrative approaches. The 5th workshop held at the National Environmental Research Institute (NERI), Roskilde, Denmark, November 2002. National Environmental Research Institute, Denmark. **161** NERI Technical Report no. 479 (2003)
[2] Brueggemann R., Patil, G.P.: Ranking and Prioritization for Multi-indicator Systems. Springer (2011)
[3] Brueggemann, R., Sørensen, P.B., Lerche, D., Carlsen, L.: Estimation of averaged ranks by a local partial order model. J. Chem. Inf. Comput. Sci.. **44**(2), 618–625 (2004)
[4] Bubley, R., Dyer, M.: Faster random generation of linear extensions. Discrete Math. **201**, 81–88.
[5] Davey, B.A., Priestley, B.H.: Introduction to Lattices and Order. CUP (2002)
[6] De Loof, K.: Efficient computation of rank probabilities in posets. Ph.D dissertation (2010)
[7] Fattore, M., Arcagni, A.: A reduced posetic approach to the measurement of multidimensional ordinal deprivation. Soc. Ind. Res. **136**(3), 1053–1070 (2018)
[8] Fattore, M., Arcagni, A., Maggino, F.: Optimal scoring of partially ordered data, with an application to the ranking of smart cities. Book of Short Papers SIS 2019. Pearson (2019)
[9] Fattore M., Arcagni, A. Ranking extraction in ordinal multi-indicator systems. Book of Short Papers SIS 2020. Pearson (2020)
[10] Rimoldi, S.L., Arcagni, A., Fattore, M., Terzera, L.: Social and Material Vulnerability of the Italian Municipalities: Comparing Alternative Approaches. Soc. Ind. Res **161**, 523–540 (2022)
[11] Schröder, B. S. W.: Ordered sets. Birkäuser (2002)

DRAFT

# Community detection analysis with robin on hashtag network

Valeria Policastro[a], Francesco Santelli[b], and Giancarlo Ragozini[a]

[a]University of Naples Federico II; valeria.policastro@gmail.com, gragoz@unina.it
[b]University of Trieste; fsantelli@units.it

**Sommario**

In Social Network science, and especially in the Social Media field, the research of communities is still an open and challenging task, mostly for what concerns the reliability of the results obtained. When dealing with hashtag networks, the research of communities is related to the identification of topics, which is a challenging achievement. Moreover, when dealing with political debates, which is our study's aim, it is even more complex. In this work, we aim to look for reliable communities on a co-occurrence hashtag network related to the Italian Political campaign (2022). To achieve this goal, we applied two different procedures to compare and validate different community detection algorithms.

***Keywords:*** social media analytics, community detection comparison, hashtag network, topic detection, political tweets, robustness

DRAFT

## 1. Introduction

In network science, the research of communities is a considerable challenge [3; 4]; moreover, there is a relevant debate around the meaning of communities, especially concerning the Social Media field [5]. In such types of networks, sometimes, not only the edge-nodes positions are used to analyze the network structure in terms of communities. Other information could be used, such as the type of link or the content itself [6; 7]. These additional attributes can lead to obtaining detailed network insights.

This contribution aims to find reliable communities of hashtags in data retrieved from Twitter concerning the Italian Political electoral campaign (2022). When dealing with political debate speeches, researching topics is an arduous task. We investigate them through a hashtags co-occurrence network. Even if many methods have been implemented, the problem of choosing the best algorithm should be addressed, especially considering how different networks can be concerning density, structure, node, and edge definitions.

To answer the problem using only edges position in the network, we used two procedures implemented in the R package `robin` [1] for comparing and validating the communities.

## 2. Aim and Methods: Comparison and validation of communities with `robin`

In network science, many methods for community detection have been developed. However, the statistical validation of the results still needs to be addressed. The main issue is whether the detected communities are significant or merely a result of chance due to the edge positions in the network. `Robin`

(Policastro et al., 2021 [1]) assesses the robustness of the community structure of a network found by one or more methods to give indications about their reliability. The basic idea is that if a partition is significant, it will be recovered even if the graph's structure is modified. In contrast, if a partition is not significant, a minimal modification of the graph will be sufficient to change the partition.

To address this issue, in `robin`, the network structure is studied through a perturbation strategy to assess the robustness of a community structure and to compare two selected detection algorithms on the same graph.
More precisely, two procedures are implemented:

- The first one tests the stability of the partitions found by a single community detection algorithm against random perturbations of the original graph structure named "null model" (function `robinRobust`).
- The second method helps to choose among different community detection algorithms the one that best fits the network of interest, comparing their robustness in pairs (function `robinCompare`).

The first procedure finds two partitions one for the real network $C_1$ and the other for the null network $C_2$ then it perturbs both networks and finds two new partitions $C_{1(p)}$ and $C_{2(p)}$, at this step it calculates two stability measures (e.g. VI [9], NMI [11], ARI [10]..) between the original partitions and the ones obtained from the perturbed network as $M(C_{1(p)}, C_1)$ and $M(C_{2(p)}, C_2)$. This process is repeated many times at different perturbation levels $p \in [0 : 0.05 : 0.6]$; such procedure leads to the construction of two stability curves, one for the observed network and the other for the null model. After the construction of the two stability curves to test their behaviour, two tests are defined. The first test which is implemented in the function `robinGPTest`, gives a more global information testing if the two curves came from the same process or two different processes. While the second test implemented in the function `robinFDATest` gives a more precise information as it analyses the curves by intervals.

*The comparison between the two curves enables us to reconsider the problem regarding the significance of the retrieved community structure in the context of the robustness of the recovered partition against perturbations.*

Moving to the second procedure, the one for the comparison of different community detection algorithms, is similar to the one just explained. However, instead of comparing the real network with the null network, it applies two different community detection algorithms on the same graph to compare them. Moreover, the R [2] package `robin` (available on CRAN: https://CRAN.R-project.org/package=robin) gives the flexibility to use every kind of detection algorithms, even custom external functions, and every kind of null models. For further methodological details, see Policastro et al. (2021) [1].

DRAFT

## 3. From the Data to the Network

The Data referring to Italian political leaders involved in the 2022 campaign have been collected via API Twitter in the R software environment, using packages `rtweet` and `twitteR`. We have done some qualitative assessments to select the most relevant national leaders with active Twitter accounts. In the end, we selected 15 leaders which belong to different political parties, from left to right wing. The temporal window in which tweets have been retrieved ranges from 2022-08-11 to 2022-09-20. In total, the tweets collected are 5969, as depicted in the table on the right side in 1.
From these data, we focused on the (hashtags $\times$ hashtags) relationships, excluding all the information related to the leaders. Overall, the complete network of hashtags has 1538 nodes and 4146 edges. Isolated hashtags (i.e., never appear with other hashtags in the same tweet) have been dropped. Furthermore, many co-occurrences have a weight equal to 1 (the $81.86\%$). Thus, such links have been deleted as well, leading to a reduced and simplified network that is easier to interpret and more informative, with fewer *noisy* links.
The final network is conceived using edge for the hashtags that appeared together at least two times, obtaining a network of 405 nodes (the hashtags) and 752 edges that now are considered binary (presence

of link or absence of link). Weights are excluded for simplicity for the community detection algorithm comparisons. The following results will refer to this *reduced* binary network.

| Leaders | N of tweets | N of hashtags |
|---|---|---|
| Giorgia Meloni | 172 | 186 |
| Matteo Salvini | 607 | 621 |
| Silvio Berlusconi | 194 | 22 |
| Giovanni Toti | 322 | 287 |
| Luigi Brugnaro | 344 | 477 |
| Maurizio Lupi | 42 | 57 |
| Enrico Letta | 493 | 693 |
| Nicola Fratoianni | 1465 | 2207 |
| Emma Bonino | 41 | 21 |
| Luigi Di Maio | 37 | 36 |
| Giuseppe Conte | 193 | 135 |
| Matteo Renzi | 126 | 135 |
| Carlo Calenda | 1215 | 477 |
| Gianluigi Paragone | 718 | 1567 |

**Figura 1:** *Reduced* Hashtags adjacency matrix as network structure-visualization; Fruchterman-Reingold Layout and node (hashtag) size proportional to degree. Table representing leaders involved in the Twitter analysis and their Social activities.

DRAFT

## 4. Analysis

The analysis of the communities will include different steps that will be described in detail below. After the construction of the hashtag network, we wondered which community detection algorithm was adequate, and to achieve this goal, we applied on it the `robinCompare` function comparing the Louvain algorithm with all other algorithms. Figure 2 shows the plots of all the comparisons in pairs with the Variation of Information measure. Analyzing the curves, we can say that the most stable algorithm, i.e., the one that creates the lowest curve, is the Infomap algorithm.

Once the most stable algorithm was found, the next step was to validate it, so we analysed the behaviour of the Infomap algorithm on a random graph and on the real graph. To do so, we first constructed the random graph with the function *random*, which rewires the edges while preserving the original graphâs degree distribution, i.e., applying the *Configuration Model* (Bender and Canfield, 1978 [8]), and second applied the function *robinRobust*. In Figure 3, it is reported the plot of the robustness procedure. Just watching the plot, we could already see by eye that Infomap behaved in the same way with the real and the random network, but also the two tests confirm that the clustering was not statistically significant. Such a result makes us reflect that even if the partitioning is stable, such partitioning is probably random at the beginning.

Since Infomap did not give reliable partitions, we went to investigate the second most stable algorithm for our network which was the Louvain algorithm (in Figure 2) as in the first steps of perturbation was the most stable algorithm (had the lowest Variation of Information, i.e. VI, measures). So we analysed the robustness of the Louvain algorithm, (see the plot in Figure 4), and we could see from the curves that the real data and the null model behaved differently.

The last step was to test the two curves with the two tests. For the Gaussian Process test, the output was a Bayes Factor equal to 82.14 which gave us strong evidence for H1, i.e., that the two curves came from two different processes. While for what concerns the Functional Data Analysis, the output was the plot in Figure 5, which explains that the two curves are statistically different as the p-values are all under

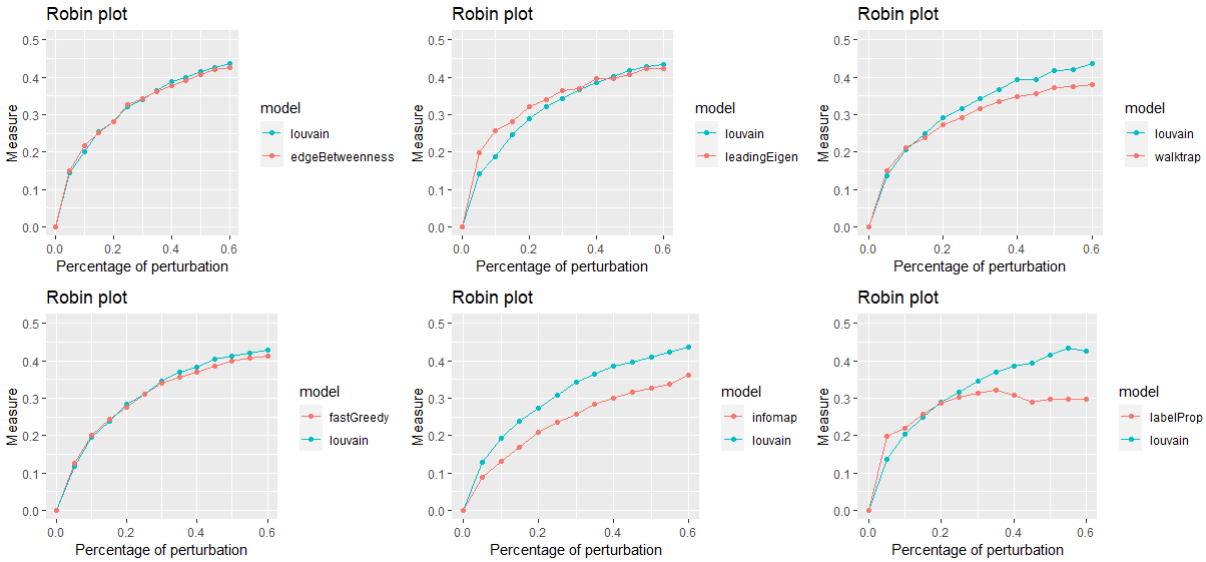



**Figura 2:** Comparison of all community detection algorithms versus Louvain algorithm with VI measure

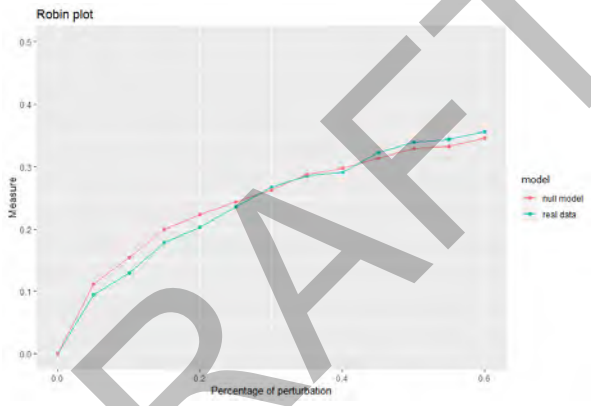



**Figura 3:** Infomap Validation curves

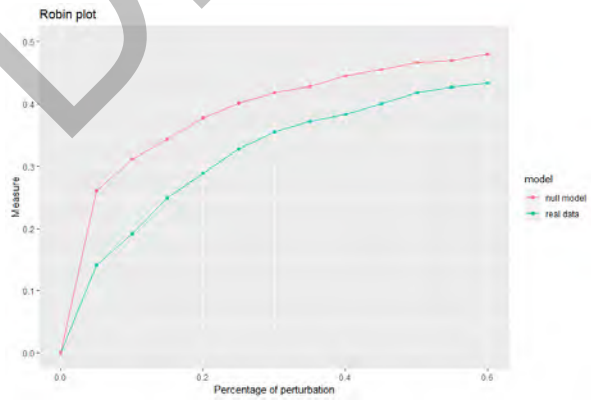



**Figura 4:** Louvain Validation curves

the red line (p-value=0.05) except for the first point which is only due to the constrain that both curves start from the 0.

We could conclude that Louvain algorithm had solid evidence to form the most stable and reliable clustering.

Analysing the different clusters coming from the Louvain algorithm see Figure 6, we found reliable

DRAFT

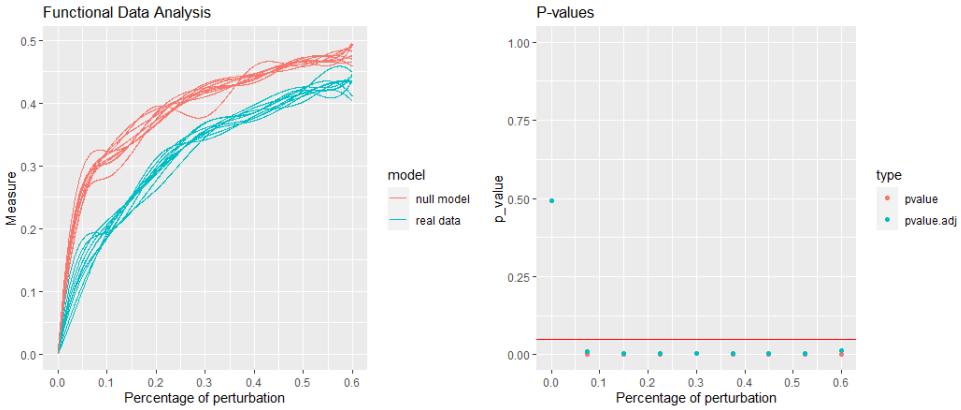



**Figura 5:** Functional Data Analysis Test of Louvain validation procedure

and interpretable clusters related to leaders and political parties, for example:

- pink cluster with hashtags: salvini, flattax, cristiani and lampedusa can be related to *Lega party - Matteo Salvini*
- blue cluster with hashtags: italexit, noeuro, novax and greenpass can be related to *Italexit - Gianluigi Paragone*
- grey cluster with hashtags: alleanzaverdisinistra, climateemergency, giustiviasociale, sinistra and crisiclimatica can be related to *Verdi sinistra italiana - Nicola Fratoianni*
- yellow cluster with hashtags: volontaripd, giovani, 25settembrevotopd and lettameloni can be related to *Democratic Party - Enrico Letta*

DRAFT

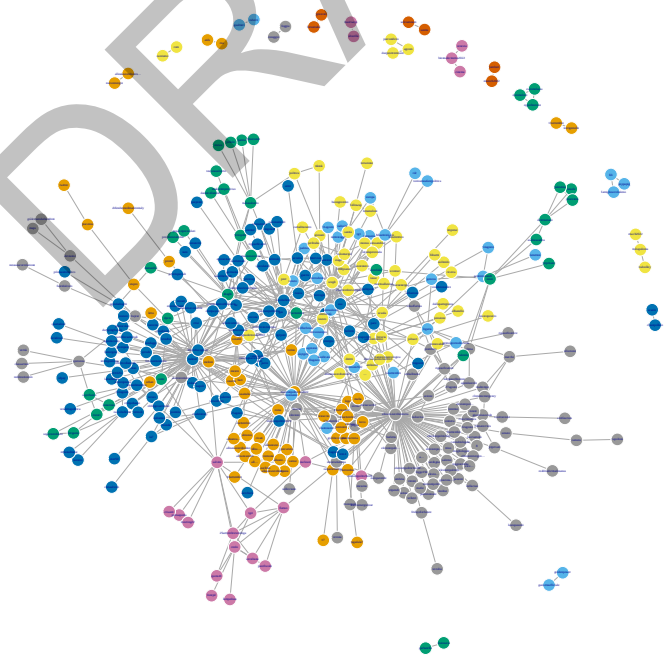


**Figura 6:** Louvain communities

## 5. Conclusions and remarks

To sum up, for the difficulty of researching topics in the political debate, we applied two procedures to our hashtag network, giving different insights into the network structure. Even if Infomap seems to be the most stable algorithm for our network (Fig:2) it has a very similar behaviour (in terms of communities discovered) when considering the random graph, indicating that the communities found are random and so not reliable. Therefore, Louvain algorithm, which is the second best for stability (as well as other methods Fig:2) identifies communities related to the real clusterization based on the network's structure, as the validation process gives significant results.

We confirmed the reliability of the partitioning of the hashtags with the Louvain algorithm as it founds different significant groups related to leaders and political parties, for example, the cluster related to *PD* or the one related to the *Lega party*.

The method applied can help to find reliable communities in social networks. However, there is not one algorithm which is the best for all types of networks, so each network must be compared and validated to find the best algorithm and both procedures must be applied to define the most stable and reliable communities.

## Riferimenti bibliografici


[1] Policastro, V., Righelli, D., Carissimo, A., Cutillo, L., and Feis, I. D. (2021). ROBustness In Network (robin): an R Package for Comparison and Validation of Communities. The R Journal, 13(1).

[2] R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[3] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America, 99 (2002), pp. 7821-7826.

[4] S. Fortunato, Community detection in graphs. Physics reports, 486(3-5), (2010), pp. 75-174.

[5] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media: Performance and application considerations. Data mining and knowledge discovery, (2012). 24, pp. 515-554.

[6] Qi, G. J., Aggarwal, C. C., and Huang, T. (2012, April). Community detection with edge content in social media networks. In 2012 IEEE 28th International conference on data engineering (pp. 534-545). IEEE.

[7] De Stefano, D., and Santelli, F. (2019, July). Combining sentiment analysis and social network analysis to explore twitter opinion spreading. In 2019 $28^{th}$ International Conference on Computer Communication and Networks (ICCCN) (pp. 1-6). IEEE.

[8] E. A. Bender and E. R. Canfield, The asymptotic number of labeled graphs with given degree sequences, Journal of Combinatorial Theory A, 24 (1978), pp. 296-307

[9] Meila, Comparing clusterings an information based distance, Journal of Multivariate Analysis, 98 (2007), pp. 873â895.

[10] L. Hubert and P. Arabie, Comparing partitions, Journal of Classification, 2 (1985), pp. 193â218

[11] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, Comparing community structure identification, Journal of Statistical Mechanics: Theory and Experiment, 9 (2005), pp. 219â228


DRAFT

# Film Tourism Motivation through the lens of Trip Advisor data

Nicolò Biasetton[a], Marta Disegna[a], Girish Prayag [b], Elena Barzizza [a]

[a] [1]Department of Management and Engineering, University of Padova
Stradella S. Nicola, 3, 36100 Vicenza, Italy;
nicolo.biasetton@phd.unipd.it; marta.disegna@unipd.it; elena.barzizza@phd.unipd.it;
[b] Department of Management, Marketing and Tourism, University of Canterbury Business School, Christchurch , New Zealand; girish.prayag@canterbury.ac.nz

**Abstract**

Motivation remains a topical issue in film tourism. However, there is no longitudinal assessment of motivation of film tourists using either survey methods or user generated comments (UGC). Online reviews, i.e., UGC contain great value as a way to understand tourist behaviour in an unstructured format and providing insights into destination marketing and management. In this study, through webscraping we extract comments from a popular global movie "The Hobbit" and "The Lord of the Rings", from Tripadvisor, and focus on selecting reviews that are related to activities and experiences available in Hinuera, New Zealand where some scenes of "The Lord of the Rings" and "the Hobbit" movies were shot. We analyse this data through Text mining techniques, Sentiment and Emotion analysis and LDA topic modelling to analyse and describe the tourist motivation to visit those locations.

**Keywords:** Movie set tourism, Textual data, Sentiment Analysis, Text mining, LDA

DRAFT

## 1. Introduction

Visiting a destination after watching a film is called film-induced tourism. Film-induced tourism is a highly personal experience in which each individual gives its own interpretation of media images whether they are authentic or based on fantasies. From a consumer perspective, film-induced tourism is based on different tourist behavioral aspects. For example, people may be motivated to travel to a destination they gazed on a television screen [1]. In the recent years, there has been a growing interest in understanding how movies shape visitation, destination images and tourists' expectations [2]. For example, several high-profile films have been shot in New Zealand including the Lord of the Rings (LOTR) trilogy and the Hobbit trilogy. These films have created a strong New Zealand's international image. These movies repositioned New Zealand as a Middle Earth and each trilogy publicized the destination, New Zealand, as a distinctive location, different from other landscapes in other countries. According to Connell [2], watching films such as the LOTR gives tourists an intangible experience, which can then motivate them to visit the real location. Visiting such locations where favorite films were shot gives tourists a sense of involvement and can create strong emotional bonds with the place [3]. Both quantitative and qualitative research methodologies have been used to examine the characteristics of film-induced tourists visiting different destinations from different populations and cultures ( [4], [5], [6], [7]). In this study, we focus on examining both the push and pull factors for visitors wanting to visit film locations. Existing studies suggest that the push factors revolve around relaxation and fun, increasing knowledge about the film location, exemplary qualities of the scenery and attractions related to the film, cultural and social characteristics and imagery [8]. Pull factors including quality of the film locations, tourist infrastructure around the location, and other attractions in the vicinity ( [8], [9]). In another study, Roesch [10] suggests that the characteristics of film locations are not the only appeal to viewers, but other factors such as film-related performances and experiences such as romance, nostalgia, and fantasy play an important role in captivating viewers to consider film destinations. Yet, only recently social media comments have been employed to understand and analyze film tourism (see [11]). For example, Gomez-Morales et al. [11] using Instagram posts for Game of thrones, found that tourists post more photographs on social media of locations that are shown longer on screen, that have more sequences of importance to the narrative structure and that characters interact with more intensely. These findings demonstrate the importance of the audiovisual text to the potential role of film in tourist destination promotion.

Social media are interactive computer-mediated tools that permit users to create or share content, such as information, ideas, career interests, and other forms of expression among virtual communities quickly and in real-time. Networks, such as Facebook, Twitter, and TripAdvisor are some of the most popular platforms among people worldwide. In this research, we focus on the analysis of reviews written on TripAdvisor, a popular American travel platform that return useful information to tourists who wants to organize their holidays, trips, and business meetings.
In this paper we apply different Natural Language Processing techniques aiming to deeply understand motivations and behaviours leading tourists to visit film locations. The key questions that this research aim to address analysing tourists' reviews written on TripAdvisor are: what tourists most frequently talk about when they visit a film location? What are they interested in? How do they feel about different topics?
Furthermore, tourists' level of satisfaction with the visited film location is investigated by means of Sentiment analysis, Emotional analysis, and Latent Dirichlet Allocation (LDA) model.
These methodologies are briefly introduced in the following section while in section 3 the case study is presented and discussed.

## 2. Method

### 2.1 Text mining

Text mining in its general definition concerns the extraction of structured and high-quality information from textual data. It usually involves the process of structuring the input text into databases and deriving patterns through statistical analysis and finally to evaluate and interpret the output.
In this work we pre-process the textual reviews to analyse them in terms of term-frequencies, bigram frequencies, terms co-occurrence and term frequencies through time, searching the reviews for some particular words, combination of words or bigrams defined by experts.
Before underlying such analysis, textual reviews need to be pre-processed, therefore the following steps have been performed:

- text lowering,
- punctuation removal,
- numbers removal,
- English stopwords removal (stopwords is a list of words that show up frequently but do not bring any valuable information as for example determiners (e.g. the, a, an) or prepositions (e.g. above, across, before)),
- white space stripping,
- tokenization: divide the document into single unit - in our case we tokenize into unigram (each single element from a string) and bigram (each sequence of two adjacent elements from a string),
- stemming: bringing each token to its root form (e.g. learning, learned, learns are stemmed into learn).

Once reviews are cleaned, a Document-Term-Matrix (DTM) can be constructed: the DTM is made up by all the documents (i.e. tourists' review) in rows and all the tokens (i.e. pre-processed words) extracted from the documents in the columns. Each cell of the matrix contains the frequency with which a token appears in a document.
This process allows us to identify the more frequent words used in the reviews, the presence of the experts' identified words, and to analyse associations among words.

### 2.2 Sentiment and Emotion analysis

Sentiment analysis (SA) is the process of computationally identifying and categorizing opinions expressed in a piece of text, with particular interest in determining whether writer's attitude towards a particular item (i.e. topic, product or service) is positive, neutral or negative. Through the last 15 years, different methodologies that can be grouped into 3 main approaches have been proposed in the literature to conduct a SA [12]: 1) ML based sentiment analysis; 2) lexicon-based sentiment analysis; 3) hybrid approaches.
In the present work a Lexicon based sentiment analysis has been adopted. This approach to SA requires the identification of a lexical resource called sentiment lexicon, or opinion lexicon, which is a predefined list of sentiment words associated to their semantic orientation with a positive or negative score [13], [14]. A sentiment word can be any word, i.e. adjective, noun, verb or adverb; a score can be a sentiment polarity value (for instance +1, 0 or -1 for positive, neutral or negative respectively), or a numerical value expressing

DRAFT

the sentiment intensity. According to Lexicon-based SA method a document is pre-processed and then a sentiment value is assigned to each token for which a match with words in the lexicon exists. In its base form, the final sentiment orientation of a document is obtained by combining (sum and average are the most frequent formulas) the semantic orientation values associated to the words that compose the whole document. The document is then classified as positive, negative, or neutral depending on whether the final score is positive negative or zero respectively.
Considering the characteristics of SA, Emotion analysis (EA) represents an extension of SA, allowing to identify emotions and not only the polarity of a text. EA allows to deeply describe the underlying consumers' feeling and emotions expressed in the text, using a much more complex system of analysis. Unlike SA, EA considers the subtleties within human emotions. In this context, lexicon-based approach to EA have been developed and its functioning is the same described for SA but the emotion lexicons are lists of words and their associated emotions (e.g. joy, sadness, fear, anger, surprise etc.). Being interested in extracting both sentiments and emotions from the reviews we adopt the NRC Word-Emotion Association Lexicon[1][2] (aka EmoLex). The NRC (National Research Council Canada) Emotion Lexicon is a list of English words and for each word two sentiments - negative and positive - and the eight basic emotions (proposed by Plutchik [15], [16], [17]) - anger, fear, anticipation, trust, surprise, sadness, joy, and disgust - are associated.

## 2.3 Topic modelling

Latent Dirichlet Allocation (LDA) [18], belonging to the family of topic models developed to extract topics from textual data, is an unsupervised learning topic model assuming that each document consists of a mixture of topics and each topic is a probability distribution over words. It is a document-generative model that specifies a probabilistic procedure by which documents are generated. The input to the model consists in a set of documents $D$ while the outputs are a distribution for each document over topic (*document-topic distribution* $\theta$), and a distribution for each topic over words (*topic-word distribution* $\phi$) both following a multinomial distribution. To simplify statistical inference, it is also assumed that the 2 outputs have Dirichlet priors with $\alpha$ and $\beta$ hyperparameters. Considering that:

- $D$ denotes the number of documents,
- $T$ is the number of
- topics, usually defined by the user,
- $V$ is the number of unique words in the entire document corpus,
- $N$ is the number of words in each document (document $d$ is a sequence of $N_d$ words),
- $w$ is the bag of all observed words with cardinality,
- $|w| = \sum_d N_d$ $z$ denotes the topic assignment of all words in all documents,
- $z_i$ denotes the topic assignment of $i$ -th word $w_i$ in document $d$

the graphical model representation of LDA in plate notation, where $\theta$, $\phi$ and $z$ are latent variables and word $w$ is observed is reported in Figure 1. To obtain the two distributions $\theta$ and $\phi$, two main algorithms have been proposed: variational inference and Gibbs sampling. Using these algorithms, we are able to obtain the estimated distributions that are the final output of LDA model. In particular $\phi_{v,t}$ represent the predictive probability or distribution of sampling a new instance of vocabulary word $v$ from topic $t$ and it gives a list of words under the topic $t$ ranked according to their probability values: these often provide good indication to label the topics and for this purpose they are adopted in the present work.

[1] https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/sentiment-emotion-lexicons
[2] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

DRAFT

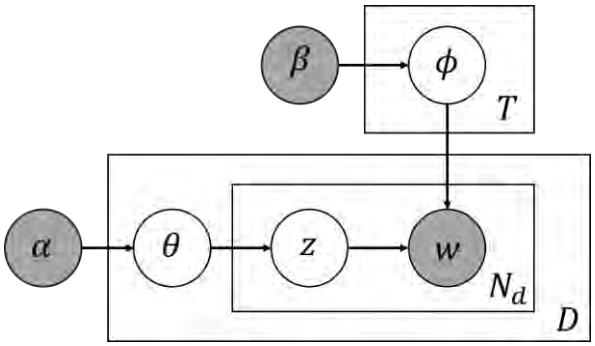


Figure 1: LDA model.

## 3. Case study and conclusion

Over the years, TripAdvisor became popular among tourists because it is free and easy to engage with. As a result, the number of user reviews and opinions on TripAdvisor have been increased rapidly from 200 million to 884 million from 2014 to 2020 [19]. Therefore, manually extracting emotions, sentiments, and behaviour from each review is an impossible task. To overcome the complexity of the process of extracting content and data from TripAdvisor, or any other websites, the technique called web crawling or web scraping [20] can be used.

In this study we will analyse tourists' reviews who visited New Zealand to see one of the film movie location of this destination. The web scraping process started from a web page collecting all activities available in Hinuera, New Zealand, a settlement in the Waikato Region of New Zealand's North Island, where the hobbits' village was settled for the Lord of the Rings (LOTR) and Hobbit movies, all pages of different activities have been scraped, collecting for each one all the reviews written by tourists that choose that experience, tour, or location. An automatic python script collected a total amount of 3285 textual reviews.

Experts identified a list of 8 basic motives to travel and visit such a destination: landscapes and scenery, learning about the movies and learning about the film shooting locations, technology used to make the film, fantasy world, destination image, film sets, fun and excitement and opportunity to socialise with others that have similar interests in films. Therefore, the words related to these motives that have been searched within the reviews are the following: "landscapes", "scenery", "learning", "movies", "film", "shooting", "locations", "technology", "fantasy", "world", "destination", "set", "fun", "excitement", "socialise", "interests", "experience", "memorable".

According to what explained in Section 2, term frequencies, bigram frequencies and term co-occurrence have been extracted from the reviews. The final list of unigrams has been filtered keeping only the list of interesting words listed above (once they have been properly stemmed the same way as the review, to match the unigrams). Using these words, a new DTM has been computed and analysed.

The LDA model has been applied to the pre-processed reviews, identifying 4 topics about which tourists talk the most. Observing words distribution over topics it was possible to label the topics as follows: topic 1 "On-site experience"; topic 2 "Packaged tour activities"; topic 3 "Other regional activities"; topic 4 "Movie set details". To describe the kind of emotions underlying each review, both sentiment and emotion analysis have been performed.

A lexicon-based Sentiment Analysis has been performed using the NRC Emotion lexicon which allows to extract both polarity of the comments (positive/negative) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). For each review, the SA procedure matches the words inside the review (properly pre-processed) with a list of words defining a lexicon, to which a score on each polarity and emotions is associated. In this way, a final score is computed to each review as the sum of the scores of the words belonging to the review.

To offset the possible influence of the review's length (number of words), the final review score has been weighted by the total number of words in the review.

Finally, each review has been associated to a single topic, i.e. the topic with the highest score, that represent the more likely topic the reviewer wanted to express in his/her review.

The analyses reveal that a combination of both push and pull factors over time affect decision-making to visit film locations. Tourists use both cognitive (landscape, scenery, destination) and emotive (fun, excitement) cues as motivational drivers. While previous studies [8] have examined different types of motives, they fail to identify whether these motives persist. Thus, we found through topic modelling that motives are primarily

DRAFT

related to the need to experience the "real" through on-site experiences. Not only tourists want to see the movie set details but they also combine their visit with other regional activities in the area. Thus, while film locations by themselves are a drawcard to the area, complementarity with other attractions is critical for destination marketing purposes. Tourists want to see a several attractions in one visit which requires destination management organizations to position film locations in relation to other attractions to visit in the region. The experience itself is also related to what is included or excluded in the package. For example, visiting Hobbiton and other areas of interest related to the films require a coordination of elements such as transport and complementary tours, which highlight value for money as being critical as well in driving visitation to the site.

**Acknowledgments**

We would like to thank Peter Fieger from Federation University, Australia, for the initial idea on this project. This research is part of the project titled "Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis" funded by University of Padova.

## References

[1] S. Beeton, H. E. Bowen and C. A. Santos, "State of knowledge: Mass media and its relationship to perceptions of quality," *Quality tourism experiences,* pp. 25-37, 2006.

[2] J. Connell, "Film tourism–Evolution, progress and prospects," *Tourism management,* vol. 33, no. 5, pp. 1007-1029, 2012.

[3] S. Kim, "Audience involvement and film tourism experiences: Emotional places, emotional experiences," *Tourism management,* vol. 33, no. 2, pp. 387-396, 2012.

[4] B. Chan, "Film-induced tourism in Asia: A case study of Korean television drama and female viewers' motivation to visit Korea," *Tourism Culture & Communication,* vol. 7, no. 3, pp. 207-224, 2007.

[5] S. S. Kim, H. Lee and K. S. Chon, "Segmentation of different types of Hallyu tourists using a multinomial model and its marketing implications," *Journal of Hospitality & Tourism Research,* vol. 34, no. 3, pp. 341-363, 2010.

[6] S. Kim and N. O'Connor, "Film tourism locations and experiences: A popular Korean television drama production perspective," *Tourism Review International,* vol. 15, no. 3, pp. 243-352, 2011.

[7] S. Kim and G. Assaker, "An empirical examination of the antecedents of film tourism experience: A structural model approach," *Journal of Travel & Tourism Marketing,* vol. 31, no. 2, pp. 251-268, 2014.

[8] S. Nunes, A. D. M. Agúndez, J. F. D. Fonseca, A. Sejdini, S. Chemli and K. J. Seo, "The importance of film-induced tourism as a motivational influence on travel decisions: analysis of push and pull factors from the perspective of Portuguese consumers," *International Journal of Tourism Policy,* vol. 12, no. 4, pp. 392-410, 2022.

[9] B. Rittichainuwat and S. & Rattanaphinanchai, "Applying a mixed method of quantitative and qualitative design in explaining the travel motivation of film tourists in visiting a film-shooting destination," *Tourism Management,* vol. 46, pp. 136-147, 2015.

[10] S. Roesch, "The experiences of film location tourists," *Channel View Publications,* vol. 42, 2009.

[11] B. Gómez-Morales, J. Nieto-Ferrando and S. Sánchez-Castillo, "Visiting game of thrones: film-induced tourism and television fiction," *Journal of Travel & Tourism Marketing,* vol. 39, no. 1, pp. 73-86, 2022.

[12] M. Birjali, M. Kasri and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends.," *Knowledge-Based Systems,* pp. 226: 107-134, 2021.

[13] A. Jurek, M. Mulvenna and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics.," *Security Informatics,* vol. 4, no. 1, pp. 1-13, 2015.

[14] M. Hu and B. Liu, "Mining and summarizing customer reviews.," *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 168-177, 2004.

[15] R. Plutchik, The Emotions: Facts, Theories, and a New Model, New York: Random House, 1962.

[16] R. Plutchik, "A general psychoevolutionary theory of emotion.," *Theories of emotion Elsevier,* pp. 3-

DRAFT

33, 1980.

[17] R. Plutchik, "On emotion: The chicken-and-egg problem revisited," *Motivation and Emotion,* vol. 9, pp. 197-200, 1985.

[18] D. Blei, A. Ng and M. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* pp. 993-1022, 2003.

[19] Statista, "Number of user reviews and opinions on TripAdvisor worldwide 2014–2020.," Statista, 21 Febbraio 2022. [Online]. Available: https://www.statista.com/statistics/684862/TripAdvisor -number-of-reviews/. [Accessed 13 Febbraio 2023].

[20] B. Batrinca and P. Treleaven, "Social media analytics: a survey of techniques, tools and platforms.," *AI & SOCIETY,* vol. 30, no. 1, pp. 89-116, 2015.

DRAFT

# Life satisfaction and social activities in later life in Italy: a focus on the Internet use

Claudia Furlan[a] and Silvia Meggiolaro[a]
[a] Dept. of Statistical Sciences, University of Padova, Italy
claudia.furlan@unipd.it; meg@stat.unipd.it

**Abstract**

The study aims to provide new insights regarding the role of social activities, and in particular those connected with the use of Internet, for the wellbeing of older people in Italy, taking into account potential differences according to the living arrangements of older people and their gender. Data come from the survey "Aspects of Daily Life", undertaken in Italy in 2020 by ISTAT. Results suggest that both for older men and women, and independently from the living arrangement, the social relations have a role for the life satisfaction, but some aspects of socialization present associations with life satisfaction that are differentiated according to gender and living arrangement. In particular, even if direct social relations maintain their important role for wellbeing of older individuals, a positive association between social relations connected with Internet use and life satisfaction are emerging except for older men living alone: Internet can thus be considered a tool to strengthen the social network.

**Keywords:** older people, life satisfaction, social relations, Internet use, logit model

DRAFT

## 1. Introduction

With the progressive ageing of societies, increasing attention in social research and among policy makers has been directed towards older adults and the ageing process (Wilhelmson et al. 2005). Two particular areas of interest are the so-called "successful ageing" and older people's wellbeing. The term "successful ageing" was popularized by Rowe and Kahn (1987), who defined the term as including three main components: "low probability of disease and disease-related disability, high cognitive and physical functional capacity, and active engagement with life" (Rowe and Kahn 1997, p. 433). Older people's "wellbeing," meanwhile, is said to be the the subjective counterpart of successful ageing (Stanley and Cheek 2003; Adams, Leibbrandt and Moon 2011), and this explains the attention that the wellbeing of older people receives in literature as well as in policy-making processes.

Specifically, among the different aspects of subjective wellbeing (SWB) that can be distinguished (see Steptoe, Deaton, and Stone, 2015 for a discussion), the evaluative dimension measured by life satisfaction (LS) is usually considered a key indicator of wellbeing and is one of the most commonly used measures of successful aging too (Tate, Lah, and Cuddy 2003). This explains the interest in life satisfaction by social research.

Previous studies have suggested that SWB is influenced by many factors (see, for example, Gaymu and Springer, 2010) and, in particular, it is positively associated with social contact and social integration (Golden et al., 2009; Meggiolaro and Ongaro, 2015); however, social networks in later life decrease due to retirement, potential health problems and death of family member and friends (Bergland et al., 2016). As a consequence, social isolation and loneliness can increase, leading to a deterioration of wellbeing.

In recent years, Internet use has become an important tool for social activities and leisure time among people, and it may be important also among older people to reduce loneliness and increase their wellbeing (Boz and Karatas, 2015). In fact, research on the association between Internet use and wellbeing among older people showed that two opposite effects can be observed (see studied cited by

Khalaila and Vitman-Schorr, 2018): Internet use can be associated, indeed, with decreased time spent with friends, and in social activities, thus decreasing wellbeing (Veena et al., 2012, Yang et al., 2021); but, on the other hand, it can be associated with social involvement in new personal friendships, and with family and friends, but also with new forms of entertainments (Chen and Schulz, 2016; Chopik, 2016; Du and Wang, 2020).

Thus, even if the literature emphasized the importance of Internet use for wellbeing among older adults, empirical results are mixed. The aim of the current study is to provide new insights regarding the role of social activities, and in particular those connected with the use of Internet, taking into account potential differences in these roles according to the living arrangements of older people and their gender. The reference is a country such as Italy, which represents an interesting case study for various reasons. Italy's population is ageing at an unprecedented rate, with a rapid rise in the numbers of older people, along with persistently low fertility and high life expectancy (Tommasini and Lamura 2009). Since the Italian context is characterized by strong family ties (Dalla Zuanna and Micheli 2004), social relationships may play a crucial role in the SWB of the elderly. Italy also offers an interesting context to study the gender differences, being characterized by a still relatively unbalanced gender system (Anxo et al. 2011), and thus differences in socialisation of old men and women could be particularly strong in shaping expectation and behaviour in later life (and thus, SWB). Lastly, in Italy the percentage of individuals using the Internet is above the EU average[1], suggesting that potential role of social activities in Internet may be particularly strong.

## 2. Data and methods

We use data from the survey "Aspects of Daily Life", undertaken in Italy in 2020 by ISTAT; it is based on a nationally representative sample of approximately 25,000 households, with a total of 42,831 individuals. In our study, we focused on 10,946 individuals aged 65 and above. The survey collected information on several dimensions of life. Besides socio-demographic and socio-economic characteristics, information on health, lifestyle, religious practices, and social integration for each household member were recorded.

As regards the focus of our study, life satisfaction was assessed with the question: "How satisfied are you with your life on the whole at present?", with answers ranging from 0 (not satisfied at all) to 10 (very satisfied). The main explanatory variables of interest for our study are those describing social relations, and we consider both direct and Internet social relations. Direct social relations were referred to social network integration and active lifestyles. Social network integration was measured on how often individuals met their friends (every day, at least once a week, less often than once a week, never or without friends) and whether they participated at meetings of cultural association. Moreover, the social network was measured by considering if individuals can count on friends, relatives, or neighbours. Active lifestyles were measured considering physical activity, volunteering, and attendance at cultural activities. Physical activity can be, at least indirectly, connected with a form of social integration and, in the following analyses, we distinguished individuals who played sports regularly or occasionally or engaged in physical activity at least once a week, those who were rarely physically active and those without any physical activity. Volunteering is measured considering whether individuals carried out free activities for voluntary or not-voluntary associations. Cultural activities considered whether individuals had gone to some cultural or entertainment venues (theatre, cinema, museum or exhibitions, classical music concerts, other music concert, monuments, or archaeological sites) in the last year before the interview. In particular, one variable was created reporting the highest frequency of attendance among the six considered activities.

As regards Internet social relations, the survey investigates if the individuals who have used Internet in the last three months, have had video calls or have texted through Internet (for example with WhatsApp or Skype), participated or expressed opinions in social networks, or uploaded self-created content on web sites or through apps. The use of social media was included in the models as the number of these activities than the individual has used.

[1] https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ifp_iu/default/table?lang=en

DRAFT

Another important variable being connected with social relations is the family in which older people live. In fact, in our study, separate analyses by gender and living arrangement are performed to take into account potential differences in the roles of social relations for different groups of individuals. In this context, the variable describing the family in which individuals live is differentiated for those living alone and those living with other people: specifically, for the former group, the marital status is considered (never married, separated or divorced, widowed); for individuals living with other people, we distinguished those living with partners without children, those living with partners and at least one child, those living without a partner but with at least one child, and those living with other persons.

Several variables are considered in the analyses as control and they may be grouped into four main domains: socio-demographic characteristics, socio-economic conditions, health status, and contextual aspects.

Socio-demographic characteristics controlled for in the analyses are the age (65-74 and 75 or more years) and religiosity (attendance at religious services: at least once a week, sometimes in a month, sometimes in a year, never).

Education and a subjective evaluation of family's economic resources were used to describe the socio-economic background of individuals. Education has four categories: low (no schooling or primary school), middle low (junior high school), middle (secondary school), high (degree or more). Economic situation is determined through a subjective evaluation of the family's economic resources: a dichotomous variable was built distinguishing whether the family had poor or insufficient resources (as opposed to very good or good).

Health was described by three covariates. The first refers to a subjective perception of health. Individuals were asked how their health is, in general (excellent or good, fair and poor health). A second variable referring to the self-reported presence of limitations in usual activities distinguishes three categories: severe limitations, only mild limitations, no limitations. The third variable refers to an indicator of mental health, where bigger values (from 0 to 100) indicate a better wellbeing.

Contextual aspects are measured considered the geographical area of residence (northern, central, and southern Italy), and some variables on the characteristics of the neighbourhood (presence of problems and availability of services).

Multivariate analyses were applied to study the association between older people's life satisfaction and their living circumstances. Life satisfaction is measured through a 0-10 anchored scale with fixed endpoints (Saris, 1988; Corbetta, 2003), which generates an "almost cardinal" variable. Since treating it as a cardinal variable is a controversial issue (Corbetta, 2003), we estimated logistic regression models dichotomizing the scores in life satisfaction. Specifically, we considered one binary categorization, using a threshold of 7, which corresponds to the mean and the median satisfaction score in our sample of the older people[2].

DRAFT

## 3. Results

Table 1 present the coefficients of the logistic regression models used to estimate the determinants of the probability of being satisfied with life (PLS) for men and women aged 65 and over in different living arrangements by gender. Results suggest that both for older men and women, and independently from the living arrangement, the social relations have a role for the life satisfaction of individuals, but some aspects of socialization present associations with the PLS that is differentiated according to gender and living arrangement.

Specifically, whereas an active lifestyle represented by physical activity has a positive role for both men and women and both for those living alone or not, other forms of active lifestyles are connected with higher PLS only for some individuals. In particular, volunteering is connected with a higher PLS for men living with others and for women living alone: for the women, this is probably due because it is a good way to find new social networks to reduce the sense of loneliness, and to satisfy the need of caring typical among lone women. Instead, older people participating in cultural activities are more likely to be satisfied with their life and this is true for all groups of older people except for men living alone.

[2] The results provided by the logistic model with the threshold of 7 are similar to the ones obtained with a threshold of 8, which is the threshold value used by Istat (2022) to identify the satisfied people.

As regards the role of social network integration, meeting friends have a positive association with the PLS again for all groups of older people, expect for men living alone. Only among the latter ones, those who can count on relatives are more likely to be satisfied with their life, probably because for them having someone to can count in case of need might be particularly important being alone.

Table 1: Results of logit models for being satisfied with life for men and women aged 65 and over, living alone or with others. β coefficients.

| | **Living alone** | | **Living with others** | |
|---|---|---|---|---|
| | *Men* | *Women* | *Men* | *Women* |
| Marital status (Ref. never married) | | | | |
| Separated/Divorced | 0.809** | -0.330 | | |
| Widowed | 0.750** | 0.102 | | |
| Living arrangement (Ref. couple with children) | | | | |
| Couple without children | | | 0.069 | 0.442*** |
| Lone parent with children | | | -0.369’ | -0.030 |
| With others | | | -0.140 | 0.268’ |
| Count on relatives (Ref. no): Yes | 0.445* | 0.066 | -0.125 | -0.126 |
| Count on friends (Ref. no): Yes | -0.072 | -0.045 | -0.004 | 0.193’ |
| Count on neighbours (Ref. no): Yes | 0.054 | 0.086 | 0.088 | -0.023 |
| Meet friends (Ref. never or without friends) | | | | |
| Less often than once a week | 0.240 | 0.142 | 0.164 | 0.173 |
| At least once a week | 0.291 | 0.299’ | 0.363* | 0.439*** |
| Every day | -0.301 | 0.464* | 0.274 | 0.633** |
| Cultural meetings (Ref. no): Yes | -0.001 | 0.001 | 0.001 | 0.003 |
| Physical activity (Ref. sedentary) | | | | |
| Rarely | 1.141*** | 0.463** | 0.143 | 0.336** |
| More than once a week | 0.750*** | 0.523*** | 0.317** | 0.452*** |
| Volunteering (Ref. no): Yes | -0.664* | 0.847** | 0.332* | 0.170 |
| Cultural activities in one year (Ref. never) | | | | |
| 1-3 times | 0.027 | 0.068 | 0.268* | 0.521*** |
| 4 or more times | 0.504 | 0.801** | 0.377* | 0.553** |
| Number of Social networks (Ref. none) | | | | |
| 1 | -0.695 | -0.096 | 0.408* | -0.155 |
| 2 | 0.247 | 0.315 | 0.475** | 0.160 |
| 3 or more | 0.041 | 0.577* | 0.504** | 0.346* |

*Significance level: ’ p<0.1, * p<0.05, ** p<0.01, *** p<0.001.*
*The logistic regression models also control for age, religiosity, education, family’s resources, health, presence of limitations, mental health, geographical area of residence, characteristics of neighbourhood.*

DRAFT

The social relations defined by the family is important among men living alone, and, in particular, those having had a marriage are more likely to be satisfied with life than their never married counterpart, but the same variable has no role for the PLS of women living alone. For women living with others, those in a couple without children are more likely to be satisfied with life than those living in couple with children.

As regards Internet social relations, the Internet use has a positive role on LS for women both living alone and with others to maintain (or create) connections with people of their social network. The Internet use has a positive effect as well for men living with other persons; instead, it is not significant for men living alone, confirming how the connection between social relations and the PLS is weak for this group, as before seen with the direct social relation.

It is undoubtedly that the direct social relations maintain their important role for the PLS of older people; however, some signals of positive effects of the Internet social relations are emerging and suggest that Internet social relations can be seen as a tool to strengthen the social network for some group

of people. This can suggest health campaigns and social programs to encourage older people to learn computer skills and use the Internet in their daily life as a mean to reduce loneliness and increase their wellbeing.

## References


Adams, K.B., Leibbrandt, S., and Moon, H.: A critical review of the literature on social and leisure activity and wellbeing in later life. Ageing Soc **31**, 683–712 (2011)

Anxo, D., Mencarini, L., Pailhé, A., Solaz, A., Tanturri, M. L., and Flood, L.: Gender differences in time use over the life course in France, Italy, Sweden, and the US. Fem Econ **17(3)**, 159–195 (2011)

Bergland, A., Meaas, I., Debesay, J., Brovold, T., Jacobsen, E. L., Antypas, K., Bye, A.: Associations of social networks with quality of life, health and physical functioning. Eur J Physiother **18(2)**, 78–88 (2016)

Boz, H., Karatas, S.E.: A review on Internet use and quality of life of the elderly. Cypriot J Educ Sci **10(3)**, 182–191 (2015)

Chen, Y.R.R., Schulz, P.J.: The effect of information communication technology interventions on reducing social isolation in the elderly: A systematic review. J Medical Internet Res **18**, 1–11 (2016)

Chopik WJ: The benefits of social technology use among older adults are mediated by reduced loneliness. Cyberpsychol Behav Soc Netw **19**, 551–6 (2016)

Corbetta P: La ricerca sociale: metodologia e tecniche. Vol. 2: Le tecniche quantitative. Il Mulino (2003)

Dalla Zuanna, G., and Micheli, G.: Strong family, familism, and lowest-low fertility. Dordrecht: Kluwer Academic Press (2004)

Du P, Wang B.: How does internet use affect life satisfaction of the Chinese elderly?. Popul Res. **44**, 3–17 (2020)

Gaymu, J. and Springer, S.: Living conditions and life satisfaction of older Europeans living alone: a gender and cross-country analysis. Ageing Soc **30(7)**, 1153-1175 (2010)

Golden, J., Conroyb, R. M., Lawlora, B.A.: Social support network structure in older people: Underlying dimensions and association with psychological and physical health. Psychol Health Med **14**, 280–290 (2009)

Khalaila, R., Vitman-Schorr, A.: Internet use, social networks, loneliness, and quality of life among adults aged 50 and older: mediating and moderating effects. Qua Life Res **27**, 479–489 (2018)

ISTAT: Rapporto Bes 2021: Il Benessere equo e sostenibile in Italia. Roma: Istat (2022)

Meggiolaro S., Ongaro F.: Life satisfaction among older people in Italy in a gender approach. Ageing Soc **35**, 1481-1504 (2015)

Rowe, J.W., Kahn, R.L.: Human aging: Usual and successful. Science **237**, 143–149 (1987)

Rowe, J.W., Kahn, R.L.: Successful aging. Gerontologist **37(4)**, 433–40 (1997)

Saris, W.E.: Variation in response functions: a source of measurement error in attitude research. Amsterdam: Sociometric Research Foundation (1988)

Stanley, M., Cheek, J.: Well-being and older people: a review of the literature. Can J Occup Ther **70(1)**, 51–9 (2003)

Steptoe, A., Dwaton, A., Stone, A.A.: Subjective wellbeing, health, and ageing. The Lancet **385**, 640–648 (2015)

Tate, R.B., Lah, L., Cuddy, T.E. Definition of successful aging by elderly Canadian males: The Manitoba follow-up study. Gerontologist **43**, 735–744 (2003)

Veena, C., Kwon, W., Juan, G.: Internet use and perceived impact on quality of life among older adults: a phenomenological investigation. Int J Health **2(3)**, 1–13 (2012)

Tomassini, C., Lamura, G.: Population Ageing in Italy and Southern Europe. In: P. Uhlenberg, (Eds), International Handbook of Population Aging. International Handbooks of Population, Dordrecht: Springer, vol 1, pp. 69–89 (2009)

Wilhelmson, K., Andersson, C., Waern, M., Allebeck, P.: Elderly people's perspectives on quality of life. Ageing Soc **25(4)**, 585–600 (2005)

Yang H.L., Wu Y.Y., Lin X.Y., Xie L., Zhang S., Zhang S.Q., Ti S.M., Zheng X.D. Internet use, life satisfaction, and subjective well-being among the elderly: Evidence from 2017 China General Social Survey. Front Public Health **9**, 677643 (2021)


DRAFT

# Social capital endowment's role in the intergenerational transmission of education

Alessandra Trimarchi[a], Maria Gabriella Campolo[b], and Antonino Di Pino Incognito[b]

[a] Universität Wien; alessandra.trimarchi@univie.ac.at
[b] Università di Messina; mgcampolo@unime.it; dipino@unime.it

**Abstract**

This paper aims at disentangling the effect of parents' fertility and education on children's educational outcomes by means of social capital endowment. We use parents' social capital endowment as exogenous factor predicting parents' education and fertility, and children's social capital values. We apply an ordered Probit to estimate children's probability to acquire a certain educational level, using the predicted instrumental variables. We use a sample of respondents aged 18-60 years old from the cross-sectional Bank of Italy's Survey on Household's Income and Wealth (SHIW) 2010. We found that parents' social capital endowment is a good predictor of parents' education, fertility, and children's own social capital. All these aspects affect differently the probability of a child to acquire a tertiary degree.

**Keywords:** education, intergenerational transmission, social capital, fertility

DRAFT

## 1. Introduction

Children's school outcomes are often a consequence of enduring inequalities of opportunities [5]. Substantial social stratification research converges on the fact that parental educational background is an important determinant of children educational outcomes [3,8]. Highly educated parents have more economic and cultural resources to sustain their children during education, and to motivate them to continue schooling.

Social stratification literature has emphasized the role of parental socio-economic background, and hence, children's social capital, for their educational outcomes [1,6]. Bourdieu [1] sees social capital as a tool of reproduction for the dominant class, whereas Coleman [6] emphasizes the social control aspect, where trust, information channels, and norms are characteristics of the community. In this latter framework, the family is responsible to adopt certain norms to advance children's life chances.

Recently, a strand within the social mobility and stratification literature paid attention to the role of demographic pathways in the transmission of education from one generation to the other. For instance, it has been found that for highly educated women there is no effect of (own) education on the probability to have a highly educated child, because of the negative effects of tertiary education on women's likelihood to marry and have children [5,13].

In this paper, we combine the two strands of research, aiming at disentangling the effect of parents' fertility and education by means of the social capital endowment. Previous research on the role of social capital on the level of education of the individual is vast. Differently from previous studies, we consider the complexities of the relationship between education and fertility in the parents' generation, using parents' social capital endowment as exogenous factor predicting parental background. In addition, we

also identify the effect of social capital endowment of children on their own education goal as a result of the transmission of social capital values by their parents. In doing this, we assume the components of parents' social capital endowment as exogenous predictors of children's social capital endowment. To evaluate the individual endowment of social capital, which can be defined as the values related to individuals' networks and communities, we use the categorization suggested by the Social Capital Benchmark Study (see, among others, Saguaro Seminar, 2001). This approach provides a broad conceptualization by examining indices that also relate to the pre-requisites necessary to build social capital, the aspects considered are: social trust, trust in legal system and public institutions, membership and group participation; altruism; informal interaction among individuals [12]. Furthermore, we have considered the aspect of "shared norms" ([7]), which lies outside the Saguaro Benchmark Study's approach. (

To this aim, we use the cross-sectional Bank of Italy's Survey on Household's Income and Wealth (SHIW) 2010, which retrospectively reports, in addition to sociodemographic information, a section regarding questions on the individual endowment of social capital. This information can also be traced back to the parents of the respondent. Using these data, it is possible to empirically derive associations between endowment of parents' social capital and the intergenerational transmission of education.

We expect that social capital endowment is an important component which predicts the level of education of the parents and their fertility behavior. We expect that a positive inclination of parents towards norms and attitudes about the relevance of the family and having children impacts positively parents' fertility and, indirectly, negatively affects children's probability to obtain a tertiary degree. This is in line with the so-called dilution theory and siblings' effects on individuals' human capital [9]. Next, we hypothesize that the relevance of family and having children is negatively associated with the educational level of the parents and, indirectly, reduces the positive impact of parents' education on the children's probability to obtain a tertiary degree.

## 2. Data and Methods

We used a sample of males and females aged between 18 and 60 years, drawn from the second "rotation" of the Italian cross-sectional SHIW survey of the Bank of Italy 2010. Dropping out the students in the sample, we finally analysed a sample of 2773 subjects. .. Information on the parents who do not live in the household of the respondents is collected with retrospective questions. At first, we correctly identify the associations between social capital endowment of the parents, their educational level and fertility, and children's social capital endowment. Predicted parents' education and fertility, as well as predicted social capital endowment of children, are then used as instrumental variables to predict children's education level.

We use answers to specific questions included in the questionnaire to obtain information on proxy variables that reasonably fall into the categories of social capital (following Putnam's Saguaro Seminar benchmark [11]). The SHIW 2010 survey contains information on various proxy variables of the individual endowment of social capital, in the questionnaire the respondent has to evaluate each social capital aspect with a score that ranges between 1 and 10: 1) tolerance, 2) compliance with authority (parents, educators), 3) respect of laws, 4) norms related to the relevance of the family and having children, 5) norms regarding being successful at work 6) being careful in trusting people. These questions are asked to the respondent in two different fashions. At first, it is asked, to what extent, in the education received by the respondent, the family of origin insisted on transmitting the values listed above. Then, it has been asked to the respondent how much himself/herself insisted (or thought it is reasonable to insist) on the importance of the same values in the upbringing of children. In the first case, it is possible to retrospectively survey a proxy-measure of the values transmitted from parents to children and which we considered as part of the social capital endowment of the parents. The proxy measures of the parents' social capital endowment are adopted as exogenous instruments of parents' education and fertility, as well of children's social capital endowment. The corresponding "instrumented" variables are then jointly used in the equation that predicts children's educational attainment.

We apply a two-stage approach: at first, we estimated simultaneously the education level of both parents, adopting a two-equation seemingly unrelated regression (SUR) model, in which the two dependent variables measure the years of schooling of the father and mother of the respondents. Regressors sets of both equations include social capital components as instruments. In this first stage, we also estimated the fertility of parents through a truncated Poisson regression (see, among others, [10]) in which

DRAFT

the dependent count variable, given by the number of children, is truncated to zero because we cannot include in our sample individuals who did not become a parent. Furthermore, also at this stage, we estimated the components of children's social capital endowment, performing a SUR regression model and using parents' components of social capital as instruments.

In the second stage, we apply an ordered Probit regression to estimate the probability of the respondents to obtain lower secondary educational level, a high school degree, or tertiary degree. In this regression, we include as explanatory (instrumental) variables the previously predicted parents' education and fertility, and the predicted components of children's social capital endowment.

## 3. Preliminary Results

The results obtained at the first stage by performing SUR and truncated Poisson regressions show that the components of social capital are significant predictors of both education and fertility of the parents, as well as of social capital endowment of children. In line with our expectations, the higher the score on attitudes and norms related to the relevance of the family and having children, the lower parents' level of education. The association is reversed when the outcome is parents' fertility. Performing appropriate overidentification and weak-instruments tests, we verify the validity and utility of parents' components of social capital as instruments in a two-stage estimation procedure (testing results are omitted here for the sake of brevity but remain available upon request).

Predicted parents' education and fertility, as well as predicted respondents' social capital endowment, strongly impact children's probability to obtain a certain educational degree. In Tab. 1 we report the ordered probit estimation results, measuring the impact of some relevant explanatory variables and the predicted parents' education and fertility by means. In line with expectations, and previous literature, parents' education positively affects children's education, while parents' fertility negatively. In addition, respect to previous studies, marginal analysis results show how the (negative) percentage effect of parents' fertility on children's educational career is stronger than the positive effect of parents' education (Table 2).

Finally, marginal effects of social capital components of children (reported in Table 2) show how predicted attitude to obedience to authorities negatively impacts the probability to be graduated, while predicted respect of laws and attitude to be successful at work have a positive impact. The analysis of profiles reported in Figure 1 show different results for women and men with regards to the effect of some components of parents' social capital endowment. However, the interpretation of these results requires further deepening.

This study presents also some limitations. It should be mentioned that due to data availability, the used measures of social capital endowment transmitted from parents to children are inevitably subjective. To clean the measurement from this subjective bias, we would need parents' responses, which, unfortunately, are not available. Moreover, the proxy variables used should be interpreted as micro-level representations of the endowment of social capital, mainly constituted by the importance of intra-community links. At a macro level, a social capital endowment could be thought of as states' responsiveness to society or the institutional capacity of a community ([7]).

DRAFT

Table 1: Second stage – Ordered Probit Regression

| Dependent variable: | Children's education level (secondary=1; high school=2; degree=3) | | |
|---|---|---|---|
| | Coef. | *Std. Err.* | p-value |
| Woman: 1=yes; 0=otherwise | -0.44 | *0.07* | *** |
| Age | -0.01 | *0.00* | * |
| Residence area of the descendants: 1= South and Islands; 0=North-Centre | -0.17 | *0.08* | * |
| Woman had a child before twentieth year: 1=yes; 0=otherwise | -1.01 | *0.18* | *** |
| Man worked before twentieth year: 1=yes; 0=otherwise | -1.34 | *0.07* | *** |
| Woman worked before twentieth year: 1=yes; 0=otherwise | -0.66 | *0.07* | *** |

*Table 1 continued:*

| | | | |
|---|---|---|---|
| Predicted parent education (IV) | 0.12 | *0.03* | *** |
| Predicted parent fertility (IV) | -0.41 | *0.16* | ** |
| Predicted obedience to parents and educators (IV) | -0.42 | *0.09* | *** |
| Predicted respect of law (IV) | 0.37 | *0.07* | *** |
| Being successful at work (Predicted IV) | 0.06 | *0.04* | |
| Risk preference (result of PCA score) | 0.13 | *0.02* | *** |
| Constant cut 1 | -1.03 | | ** |
| Constant cut 2 | 0.44 | | |
| Pseudo $R^2$ | 0.13 | | |

*Note:* *** p <0.001; ** p <0.01; * p <0.05

Table 2: Second stage – Ordered Probit Regression Marginal Analysis – Average effects (in percentage) of instrumental variables on offspring's education goals (semi-elasticity)

| | % Effect on tertiary education | *95% CI* | | % Effect on secondary education | *95% CI* | |
|---|---|---|---|---|---|---|
| Predicted parent education (IV) | 21.02 | *9.30* | *32.74* | 5.39 | *2.36* | *8.42* |
| Predicted parent fertility (IV) | -72.05 | *-126.17* | *-17.92* | -18.47 | *-32.44* | *-4.51* |
| Predicted obedience to parents and educators (IV) | -74.94 | *-105.39* | *-44.50* | -19.22 | *-27.16* | *-11.27* |
| Predicted respect of law (IV) | 65.81 | *40.28* | *91.34* | 16.88 | *10.19* | *23.56* |
| Being successful at work (Predicted IV) | 10.99 | *-1.33* | *23.32* | 2.82 | *-0.35* | *5.99* |

DRAFT

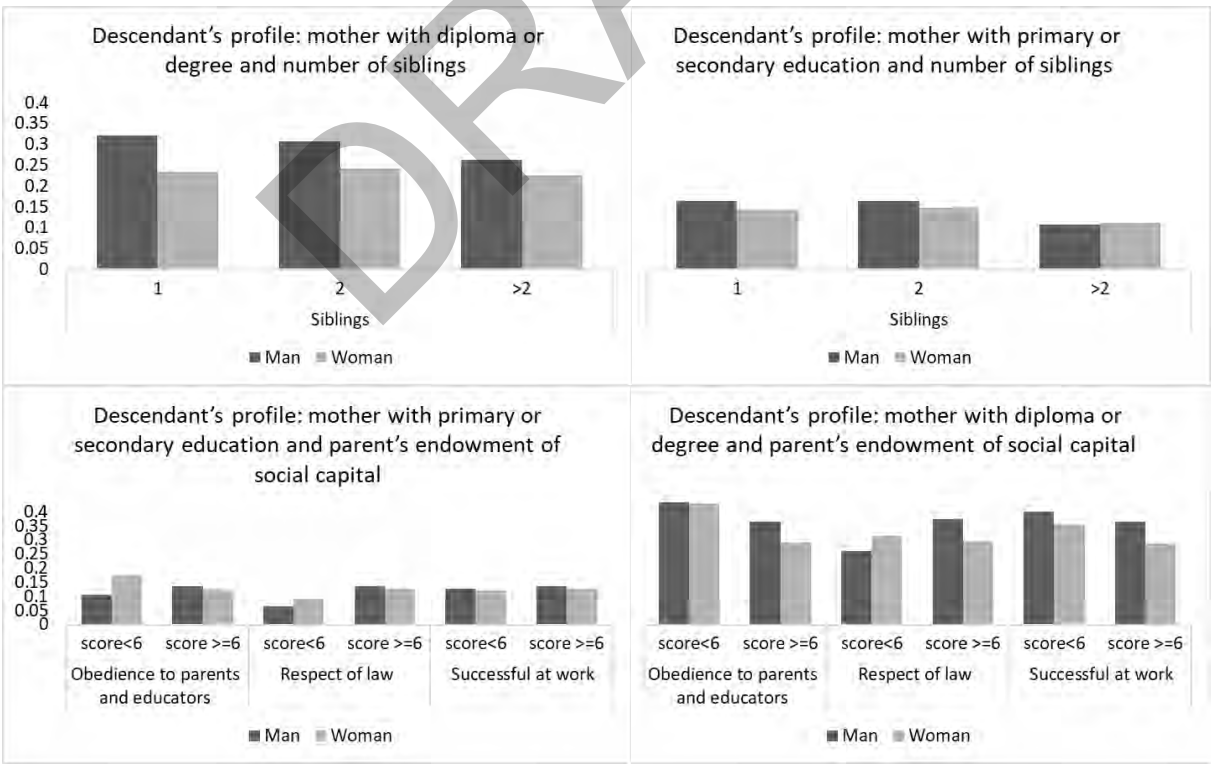



Figure 1: Probability of a descendant to be graduated.

## References


[1] Bourdieu, P.: The Forms of Capital. In: Richardson, J. (ed.) Handbook of theory and research for the sociology of education, pp. 241–258. Greenwood , New York (1986)

[2] Breen, R.: Educational Expansion and Social Mobility in the 20 th Century. Soc. Forces 89(2), 365--388 (2010)

[3] Breen, R., Luijkx, R., Miiller, W., Pollak, R.: Nonpersistent inequality in educational attainment: Evidence from eight European countries. Am. J. Sociol. 114(5), 1475--1521 (2009)

[4] Breen, R., Jonsson, J.O.: Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility. Annu. Rev. Sociol. 31(1), 223--243 (2005)

[5] Breen, R., Ermisch, J., Helske, S. Educational reproduction in Europe: A descriptive account. Demogr. Res. 41, 1373--1400 (2019)

[6] Coleman, J.S.: Social capital in the creation of human capital. Am. J. Sociol. 94(1) Supplement: Organizations and institutions: Sociological and economic approaches to the analysis of social structure, 95--120 (1988)

[7] Engbers, T.A., Thompson, M.F., Slaper, T.F.: Theory and measurement in social capital research. Soc. Indic. Res. 132(2), 537--558 (2017)

[8] Engzell, P., Mood, C., Jonsson, J.: It's All about the Parents: Inequality Transmission across Three Generations in Sweden. Sociol. Sci. 7, 242--267 (2020)

[9] Ferrari, G., Dalla Zuanna, G.: Siblings and human capital: A comparison between Italy and France. Demogr. Res. 23, 587--614 (2010)

[10] Hardin, J.W., Hilbe, J.M.: Regression models for count data from truncated distributions. The Stata J. 15(1), 226--246 (2015)

[11] Putnam, R.D.: Better Together: Report of the Saguaro Seminar on Civic Engagement in America (2000). Available at: http://robertdputnam.com/better-together/the-report/. Cited 24 Jan 2023

[12] Saguaro Seminar. (2001). The social capital community benchmark survey. Cambridge: John F. Kennedy School of Government, Harvard University.

[13] Skopek, J., Leopold, T.: Educational Reproduction in Germany: A Prospective Study Based on Retrospective Data. Demography, 57(4), 1241--1270 (2020)


DRAFT

# Streaming Data from Social Networks to Track Political Trends

Emiliano del Gobbo[a] and Barbara Cafarelli[a]

[a]Department of Economics, Management and Territory, University of Foggia;
emiliano.delgobbo@unifg.it, barbara.cafarelli@unifg.it

**Abstract**

One of the most interesting areas of research that has emerged since the introduction of these platforms is the analysis of material from social networks. Over the years, several studies have tried to exploit social media data to acquire information that could be useful for predicting or understanding people's opinion on political and current issues, but also for brand and product marketing purposes. Our research focuses on the analysis of Twitter data collected during the 2022 Italian general elections. A statistical learning approach has been exploited to analyse the flow of user-submitted comments that covered the political discourse. It is possible to understand public opinion trends in an electoral context using existing tools.

***Keywords:*** statistical learning, text mining, streaming data, opinion mining

DRAFT

## 1. Introduction

Individuals often express their views on facts, current events, politics, and sports on internet platforms such as forums and web news comments. In recent years, the widespread use of social networks has attracted more people to the internet and made it easier for them to share their opinions. Opinions expressed in social media text messages are a valuable resource for extracting knowledge. Text data mining techniques are used to extract knowledge from this type of data, as text messages are unstructured data, and specific methods have been developed.

One of the most common types of text analysis is sentiment analysis, which is used to determine the sentiment of people on various topics in order to predict future events or identify emerging trends. Several studies have shown that opinions expressed on social networks are a good indicator of overall public opinion [see 8; 4; 10; 5, among others].

Our research focuses on the specific case of the 2022 Italian general election, that involves the full engagement of the entire population, and a large coverage on media. This election has been the first one occurring in late summer and it came after the resignation of Prime Minister in July. This event took the parties unprepared and forced them, not only to rashly organize a electoral campaign, but also to build alliances plan, that changed up to few weeks from the election day.

For this study we used data collected from Twitter, through the Twitter API and Socialgrabber, a custom software that collect tweets in archives. The data collected and examinated in this research span from 22 july 2022 (the day after the Prime Minister resignation) to 27 Sempteber 2022 (the day after the general election).

## 2. Background

Data are the base resource, but their exploitation in any application is due mainly to Data Mining techniques. *Data Mining* is the process that allows finding hidden patterns using automated or semi-automatic systems [12]. Data Mining techniques are widely used in marketing and sales, as they often can elaborate large amounts of information, providing worthy insights and a relavant competitive advantage with a direct return on profit. The evolution of methodologies for Sentiment Analysis, is particularly entangled with progresses in the Natural Language Processing (NLP) field. Wang et al. [11] present one of the earliest research about streaming sentiment analysis. The authors implemented a system with a visual dashboard for real-time analysis of public sentiment expressed on Twitter toward presidential candidates in the 2012 U.S. election. To measure the sentiment the authors trained a Naive Bayes model, considering as features the TF-IDF matrix generated from tweets. This approach required the preparation of a big corpus manually classified. The final purpose was the implementation of a visualization tool to monitor the evolution of sentiment towards specific subjects. Rahnama et al. [9] proposes Vertical Hoeffding Tree, a parallel decision tree classifier to classify tweets. Their parallelization-based solution reduces both the computing effort and the necessity to store the data stream. A study about the UK's general elections in 2015 [2] has again sought to predict electoral results through data collected on Twitter. Although it failed to properly predict the party that would have won more seats, a factor that also depends on how the votes are geographically distributed. However, the research was able to predict the order of Parties in terms of the number of votes received, even succeeding in predict the spread out of a not-important party in the previous elections, UKIP's Nigel Farage. The study in this case made use of a lexical-based sentiment software that assigns a score between +5 and -5 at each word based on whether it is considered positive or negative. A study on the referendum for the Brexit [3], the referendum launched in the United Kingdom to decide the exit from the European Union, showed how NLP (Natural Language Processing) techniques applied to a dataset of data collected on the web, social networks and forum comments and newspapers have allowed a fairly accurate prediction of election results.
Finally, text mining techniques have been shown to be relevant also to track the evolution of long-term debates on the social media platform, as in [6] where the debate of the last year before Brexit in the UK was analyzed, detecting the most relevant topics and their time collocation through Latent Dirichlet Allocation and Dynamic Topic Modelling. The analysis of big social data can be generalized to the rest of the offline world and it's useful not only for trying to predict the election results (which is still an insidious operation when the candidates are close in elector approval), but above all to interpret the feelings, opinions, and issues affecting people, allowing to conduct more effective political campaigns, as well as implementing more appropriate business and marketing strategies.

DRAFT

## 3. Methodology

The first step of our research is related to the collection of data from Twitter. This process required a selection of the relevant keywords for the electoral debate. The keyword selection involved the candidates' parties and leaders and the most relevant politician and public personalities connected to them. The most critical factor was that two month before the elections parties were not yet organized for the electoral campaign and alliances was not too, due to the unexpected election. Therefore the number of parties, candidates, and alliances changed over time, and this required updating the keywords according to the changes. Over 4 millions of tweets have been collected in the time frame of two months. The data have been processed with Python scripts. To compute the sentiment, the Python Library *sent-it* was used [1]. The library contains a model based on the current state of the art for NLP, BERT [7] and fine-tuned on the Italian corpus of tweets to classify positive and negative attitude contents. The processing is run simulating a data streaming from the full collection of data. The sentiment model is trained with older tweets than the dataset used, and the data are binned by time, as this would occur in a real data stream. In this way, the implemented approach does not benefit from owning the full dataset beforehand

## 4. Results and Limitations

Figure 1 provides some preliminary results. Each graph shows the rate of positive tweets over the total number of tweets by day assigned to that party. Each tweet is assigned to a party based on the keywords in the text. Tweets with keywords belonging to more than a party have been removed from the dataset. Only parties with at least 1000 tweets by day have been selected to avoid unreliable time series due to the insufficent data. The overall results show how the general debates on tweets seem to feature a negative attitude towards parties: the positive rate is almost always under 30%. While for some parties, such as "Fratelli d'Italia" the tweet time trend seems to follow the one predicted by polls, others, such as "Lega" and "Movimento 5 Stelle" seem to be opposite to the trends tracked by poll. Some of the unexpected results may be explained by the limitations of the methodology adopted. First of all, is the keywords selection: that has a strong impact on the gathered tweets, for example some of the politician names used as keywords may lead to the selection of tweets with a specific attitude different from the general attitude towards the party. The second limitation is intrinsic in the model used to evaluate the attitude of tweets: the model was trained to predict attitude on a dataset of tweets not updated with the current discussion themes. This aspect is particularly relevant for textual data, as the way the attitude is expressed is particularly related to the domain of discussion. The last one is related to party positioning: Twitter is not a perfect representation of the population, and some groups of people may be more represented on social media, or more active than others, influencing social media analysis results. While the last point represents one of the main limitations of social network political analyses, in the future the second point may be addressed by building a model more tailored and able to grow and adapt together with the debate, reducing the human efforts in the construction of training datasets.

DRAFT

## References


[1] Federico Bianchi, Debora Nozza, and Dirk Hovy. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021.

[2] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using Twitter to predict the {UK} 2015 General Election. *Electoral Studies*, 41:230 – 233, 2016.

[3] Fabio Celli, Evgeny A. Stepanov, Massimo Poesio, and Giuseppe Riccardi. Proceedings of the Predicting Brexit: Classifying Agreement is Better than Sentiment and Pollsters. In *Proceedings PEOPLE Workshop*, pages 110–118, Osaka, Japan, 12 2016.

[4] Andrea Ceron, Luigi Curini, Stefano M. Iacus, and Giuseppe Porro. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, 16(2):340–358, 2014.

[5] Emiliano del Gobbo, Lara Fontanella, Sara Fontanella, and Annalina Sarra. Geographies of twitter debates. *Journal of Computational Social Science*, 5(1):647–663, May 2022.

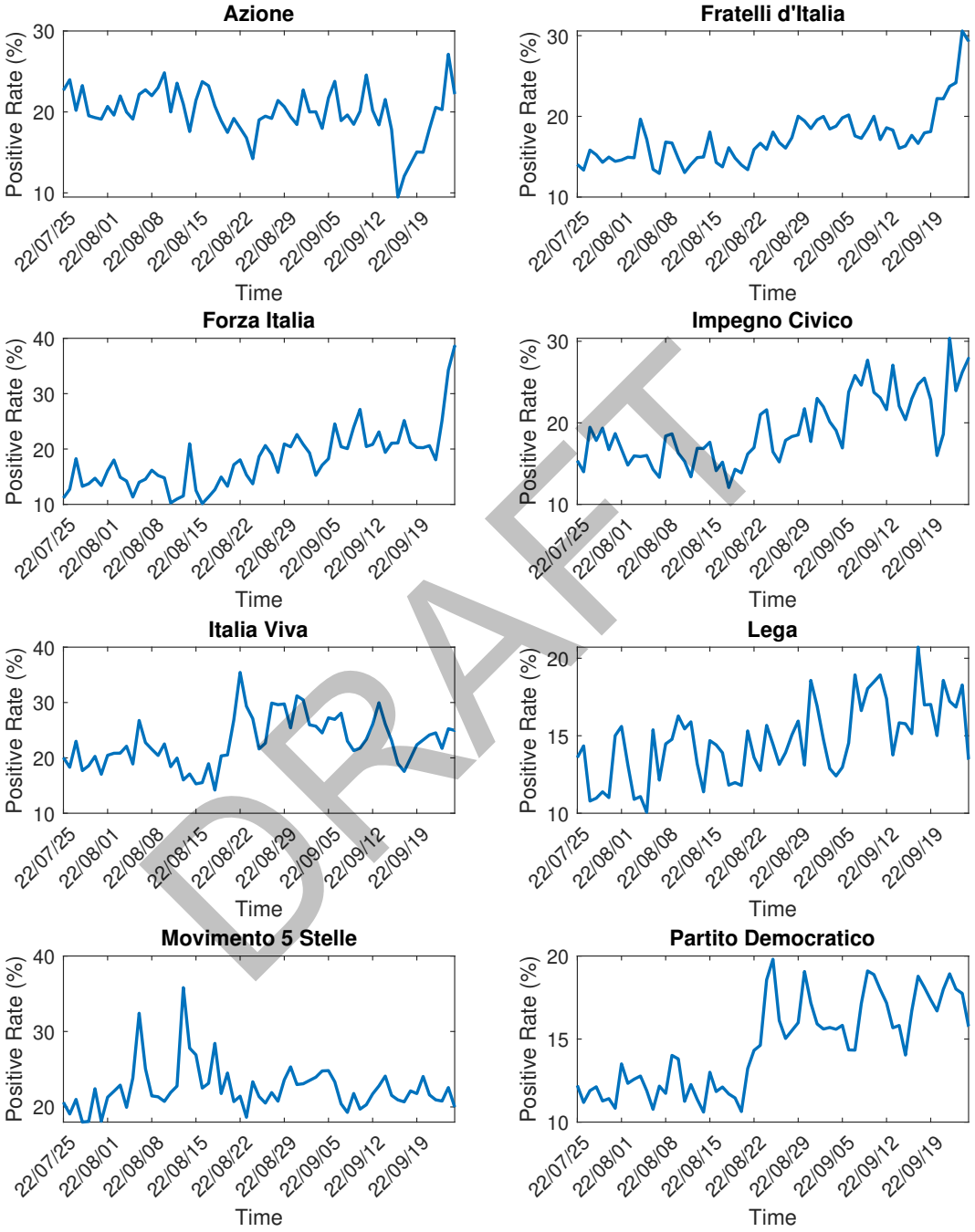
[6] Emiliano del Gobbo, Sara Fontanella, Annalina Sarra, and Lara Fontanella. Emerging topics in brexit debate on twitter around the deadlines. *Social Indicators Research*, 156(2):669–688, Aug 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[8] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[9] Amir Hossein Akhavan Rahnama. Distributed real-time sentiment analysis for big data social streams. In *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 789–794, 2014.

[10] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting




DRAFT

Figure 1: Rate of positive tweets by day and party

Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Leopoldstraße 139, 80804 Munich, Germany, 2010.

[11] Hao Wang, Doğan Can, Abe Kazemzadeh, FranÃ§ois Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. pages 115–120. Association for Computational Linguistics, 2012.

[12] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data mining : practical machine learning tools and techniques*. Elsevier, 2011.

DRAFT

# The scientific production on gender dysphoria: a bibliometric analysis

Maria Gabriella Grassia[a], Marina Marino[a], Massimo Aria[b], Rocco Mazza[c], Luca D'Aniello[a], Agostino Stavolo[a]

[a] Department of Social Sciences, University of Naples Federico II;
mariagabriella.grassia@unina.it, marina.marino@unina.it,
luca.daniello@unina.it, agostino.stavolo@unina.it
[b]Department of Economics and Statistics, University of Naples Federico II;
massimo.aria@unina.it
[c] Department of Political Science, University of Bari "Aldo Moro", Bari, Italy;
rocco.mazza@uniba.it

DRAFT

**Abstract**

Defined by Diagnostic and Statistical Manual of Mental Disorders (DSM-5), gender dysphoria is a psychological condition characterized by marked incongruence between the gender expressed by an individual and the gender assigned at birth. The scientific community has long debated the issue of gender dysphoria, first defining it as a gender identity disorder until now when it has been de-pathologized from mental disorders. The aim of the work, therefore, is to map the scientific production of gender dysphoria topics through bibliometric techniques in order to analyse scientific productivity and study their contents.

**Keywords:** gender dysphoria, bibliometric analysis, science mapping

## 1. Introduction

Gender dysphoria is a psychological condition in which a person subjectively feels their identity and gender does not conform to their birth-attributed sex, often causing clinically significant impairment insocial areas, affective-relational areas, and other areas of life (Frew et al. 2021). People who experience gender dysphoria may describe themselves as transgender, nonbinary, or trans+, due to the perception that their gender identity does not match their assigned sex at birth (American Psychiatric Association,2013).

Over the years, the clinical definition of gender dysphoria is changed, moving from "gender identity disorder" to "gender dysphoria" in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). According to the DSM-5, dysphoria is manifested by the following criteria: (a) the desire to be treated as the opposite gender; (b) the belief that one has the typical feelings and reactions of the opposite gender; (c) marked incongruence between an individual's experienced/expressed gender and primary and/or secondary sexual characteristics; (d) a strong desire to shed one's primary and/or secondary sexual characteristics due to a marked incongruence with one's experienced/expressed gender; (e) a condition clinically significant that provide distress or impairment in social, occupational or other functioning; (f) desire for the primary and/or secondary sexual characteristics of the opposite gender.

In 2018, the World Health Organization released a new version of the International Classification of Diseases (ICD-11) that state gender dysphoria as "gender incongruence", removing it from mental disorders.

To better understand the theme in various aspects, it is necessary to analyse and map the literature about it. Mapping the scientific knowledge of a topic means performing a comprehensive review of its relevant scientific literature. This involves synthesizing previous research findings to effectively lever-age the existing knowledge base and identify potential avenues for future research. Bibliometrics allowsimplifying

the review process by offering a systematic, transparent, and repeatable methodology based on statistical measurements of scientific production, scientists, and scientific activities (Cuccurullo et al., 2016). This methodology involves the application of quantitative methods for evaluating, tracking, and quantifying published research within one or multiple fields over time.

In several research domains, bibliometric methods are implemented to assess the impact of the field, specific researchers, individual papers, and reference journals, as well as to identify knowledge inputs, research gaps, trends, and future opportunities (Zhao, 2010). Two primary bibliometric approaches used for investigating a research domain are performance analysis and science mapping (Noyons et al., 1999). In this study, we concentrate on science mapping, as it enables the identification and visualization of themes and trends, both synchronically (Callon et al., 1983) and diachronically (Cobo et al., 2011). This work aims to present a systematic literature review about gender dysphoria to analyse its scientific productivity, in terms of the number of documents published about this topic, and to study their contents through quantitative analyses.

## 2. Methodology

To investigate the content of scientific documents, we map their conceptual structure by conducting two related analyses: a term co-occurrence network analysis and a strategic or thematic map. The use of these techniques allows (i) measuring the relationship among terms, (ii) detecting the main research themes, and (iii) highlighting how they are developed.

The term co-occurrence network analysis (Wang et al., 2019) identify of one or more topics represented by a set of terms (e.g., keywords, terms extracted from titles, or abstracts). To understand the themes covered by a research field, a network representation is used in order to find which are the most important and the most recent research frontiers. By considering the network approach, we perform a term co-occurrence matrix. Each cell outside the principal diagonal contains values that measure the number of times two terms appear together in the collection of documents (co-occur). Then, the association index, proposed by Van Eck and Waltman (2009), is used to normalize the co-occurrences among terms. The index, which can assume values between 0-1, reflects the strength of the association amongterms.

The co-occurrence matrix can be represented as an undirected weighted network in which terms are nodes and the associations between linked terms are reflected by edges, showing both single terms and subsets of terms frequently co-occurring together. To identify subgroups of strongly linked terms, where each subgroup corresponds to a center of interest or a theme of the concept structure of collected documents, we perform community detection (Fortunato, 2010) using the Walktrap algorithm (Pons and Latapy, 2006).

Topics identified through the community detection algorithm can be plotted on a Strategic or Thematic Map (Cobo et al., 2011). The Thematic Map is a bi-dimensional map where the axes **x** and **y** are built using Callon centrality and density measures respectively (Callon et al., 1983). Centrality quantifies the level of significance that a theme holds within a research field, whereas density serves as an indicator of the degree of the theme's development. These two measures allow identifying four typologies of topics (Cahlik, 2000) according to the quadrants of the map in which they are positioned.

The first quadrant, on the upper-right of the map, groups motor topics, characterized by both high centralityand density. This means that they are both developed and important.

The second quadrant, on the upper left, identify isolated topics or niche ones. These topics are distinguished by well-developed internal links, as evidenced by their high density. However, due to their lack of significant external links, they possess only limited importance for the broader research field, as indicated by their low centrality.

In the third quadrant, on the lower left, there are emerging or declining themes, characterized by both lowcentrality and density. These topics are weakly developed or marginal. In the fourth quadrant, on the lower-right of the map, basic and transversal topics are located. They are characterized by high centrality and low density identifying important topics transversal to the different research fields.

## 3. Data retrieval and main findings

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA- Liberati et al., 2009) model was used to retrieve the scientific production. We queried the Web of Science (WoS) indexing database – launched by the Institute for Scientific Information (ISI) and now maintained by Clarivate

DRAFT

Analytics – using the following query: (TS= ("gender dysphoria") OR TS= ("gender identity disorder*")). We refine our search by selecting only documents classified as Articles, Proceedings Papers, Review Articles, and Book Chapters in English published from 1985 – the starting year of indexing documents available on WoS - to 2022. The collection, consisting of 2362 records, was exported into PlainText format.

To analyze the whole collection, we used the *bibliometrix* R package, an open-source tool for quantitative research in scientometrics and bibliometrics that includes all the main statistical and visualization methods for performance analysis and science mapping (Aria and Cuccurullo, 2017). After converting the collection into a data frame, we reported some descriptive information:

- The 2362 documents were published in 815 different journals;
- The total number of references is 50191;
- The average citation per document is equal to 27;
- The number of co-authors per document is 4.

Focusing on the annual scientific production (Figure 1), the first phase - from 1985 to 2000 - included transsexualism as the first gender-related diagnostic label to be included in DSM-3 within the macro-category of psychosexual disorders (Zucker, 2015).

With the advent of DSM-4 in the early 2000s, the terminology changed to "gender identity disorder", falling under the category of sexual and gender identity disorders, developing new thinking and insights from the scientific community. The release of DSM-5 in 2013, which included a section on gender dysphoria, increased the literature on the topic. The shift towards highlighting clinical distress as opposed to treating transgender identity as inherently problematic seems to be a less stigmatizing approach, particularly given that the absence of the term "disorder" has been noted to be more inclusive (Galupo et al., 2021).

Following this, the new diagnosis introduced in 2018 has resulted in a rise in scientific output. It is defined as "incongruence" between the experienced gender and the assigned gender, which includes non-binary individuals belonging to the transgender population. Non-binary individuals do not identifywithin the traditional male-female gender binary, but rather identify outside of it (Reisner and Hughto, 2019).

DRAFT

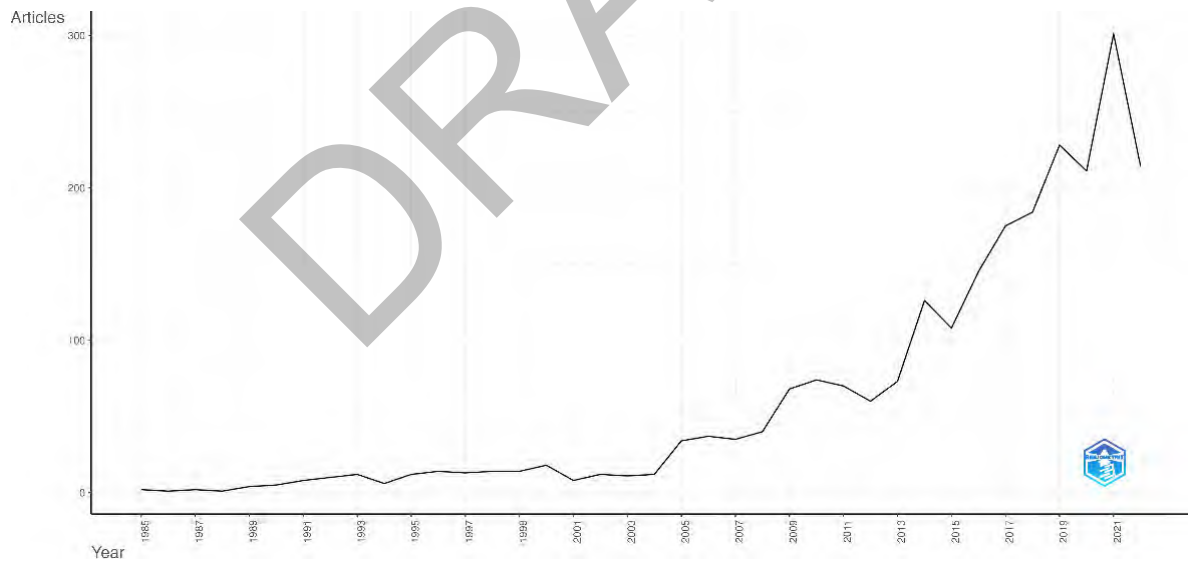



Figure 1: Annual Scientific Production of documents published from 1985 to 2022

To extract and analyse the concepts of the scientific documents collected, we perform a thematic map (Figure 2). Keywords Plus (IDs) were used as the unit of analysis to build the co-occurrence matrix and, consequently, the thematic map. IDs are words or phrases that frequently appear in the titles of anarticle's references, generated by an algorithm (Garfield, 1990) of the WoS database.

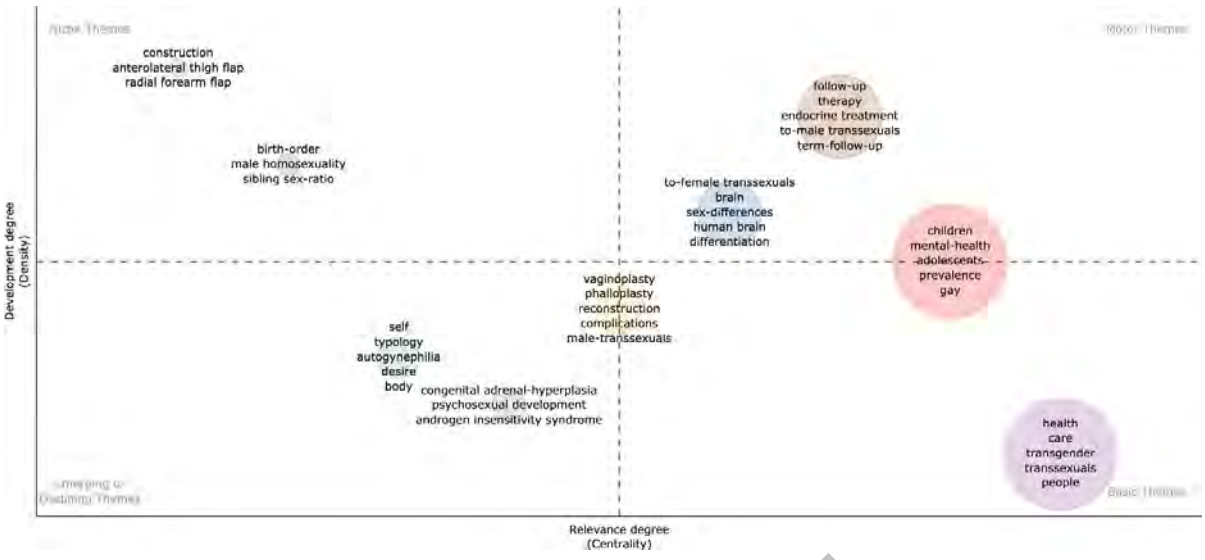



Figure 2: Thematic map of topics about gender dysphoria

Two relevant topics are highlighted as motor themes. The first one concerns the hormone therapiesthat transgender individuals receive during their transition, while the second one focuses on the biological and hormonal distinctions between male and female brains. Various studies, including Kurth et al. (2022) and Kurth et al. (2020), have suggested that the brain of a transgender person more closely resembles their gender identity rather than the gender assigned to them at birth. Straddling the motor and basic themes is a particular focus on gender dysphoria in children.

The basic theme, on the other hand, refers to health in transgender people, whose sex assigned at birth does not match their gender identity. Indeed, the scientific literature on the topic has always focused on how transgender people may experience gender dysphoria.

The emerging themes predominantly focus on medical concepts related to hormonal aspects. However, the other theme located in this quadrant pertains to the topic of transvestism, with autogynephilia being identified as the tendency of a male to experience sexual arousal from the idea of presenting as afemale and donning women's clothing (Lawrence, 2011).

Finally, the quadrant of niche themes contains topics inherent to the medicalization of sex reassignment surgeries.

## 4. Conclusions and future developments

The scientific production available on gender dysphoria appears to be strongly focused on medicalized aspects of gender reassignment. The literature encompasses not only the surgical procedures used to facilitate gender transition, particularly in transgender patients but also the hormonal treatments they undergo. Additionally, the study of gender dysphoria in children is an area of great significance within the field. Several studies have reported that dysphoria related to gender identity can manifest as early as two years of age. However, there is significant controversy over whether to support the social and/or medical transition of prepubertal children with gender dysphoria, as well as when to initiate such interventions. (Chen et al., 2018, Tompkins, 2019).

Future developments will be devoted to providing a bibliometric longitudinal analysis to study the evolution of gender dysphoria topics over time, dividing the reference timespan into different time slices. In addition, a specific emphasis will be placed on examining scientific literature within specific research domains (such as medicine, psychology, and social sciences) to gain insight into how the topic of gender dysphoria is approached and studied.

DRAFT

## References


[1] American Psychiatric Association: Diagnostic and statistical manual of mental disorders (5th ed.). Au-thor. (2013).

[2] Aria, M.; Cuccurullo, C.: bibliometrix: An R-tool for comprehensive science mapping analysis. J. In-formetr. 11, 959–975 (2017)

[3] Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E.: Fast unfolding of communities in largenetworks. J. Stat. Mech Theory Exp. (2008)

[4] Cahlik, T.: Comparison of the maps of science. Scientometrics, 49(3), pp. 373-387 (2000).

[5] Callon, M., Courtial, J. P., Turner, W. A., Bauin, S.: From translations to problematic networks: An introduction to co-word analysis. Social Science Information, 22 (2), pp. 191-235 (1983).

[6] Chen, D., Edwards-Leeper, L., Stancin, T., Tishelman, A.: Advancing the practice of pediatric psychology with transgender youth: State of the science, ongoing controversies, and future directions. *Clinical Practice in Pediatric Psychology*, 6(1), 73. (2018).

[7] Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F.: Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology, 62(7), 1382–1402 (2011)

[8] Cuccurullo, C., Aria, M., Sarto, F.: Foundations and trends in performance management. A twenty-five-year bibliometric analysis in business and public administration domains. Scientometrics, 108, 595-611.(2016).

[9] Fortunato, S.: Community detection in graphs. Phys. Rep., 486, 75–174 (2010)

[10] Frew, T., Watsford, C., Walker, I.: Gender dysphoria and psychiatric comorbidities in childhood: a sys- tematic review. Australian Journal of Psychology, 73(3), 255-271 (2021)

[11] Garfield, E.: Keywords Plus®: ISI's breakthrough retrieval method. Part 1. *Expanding your searching power on Current Contents on Diskette*, Current Contents, 32, pp. 5-9. (1990)

[12] Lawrence, A. A.: Autogynephilia: An underappreciated paraphilia. Sexual Dysfunction: Beyond the Brain-Body Connection, 31, 135-148 (2011).

[13] Liberati A., Altman D. G., Tetzlaff J., Mulrow C., Gøtzsche P. C., Ioannidis J. P. A., Clarke M., Devereaux P. J., Kleijnen J., Moher D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10), pp. e1-e34. (2009).

[14] Noyons, E., Moed, H., Van Raan, A.: Integrating research performance analysis and science mapping. Scientometrics, 46(3), 591-604. (1999)

[15] Pons, P., Latapy, M.: Computing communities in large networks using random walks. *J. Graph Algo-rithms Appl.*, *10*(2), 191-218. (2006).

[16] Reisner, S. L., Hughto, J. M.: Comparing the health of non-binary and binary transgender adults in astatewide non-probability sample. PLoS one, 14(8), (2019).

[17] Tompkins, A.: The Trans Generation: How Trans Kids (and Their Parents) Are Creating a Gender Rev-olution. (2019).

[18] Van Eck, N.; Waltman, L.: How to normalise co-occurrence data? An analysis of some well-known sim-ilarity measures. J. Am. Soc. Inf. Sci. Technol. 60, 1635–1651 (2009)

[19] Wang, H., Zhao, Y., Dang, B., Han, P., Shi, X.: Network centrality and innovation performance: the roleof formal and informal institutions in emerging economies, *Journal of Business & Industrial Marketing*,34(6), pp. 1388-1400 (2019).

[20] Zhao, D: Characteristics and impact of grant-funded research: a case study of the library and informationscience field. Scientometrics, 84(2), pp. 293-306 (2010).

[21] Zucker, K. J: The DSM-5 diagnostic criteria for gender dysphoria. Management of gender dysphoria: amultidisciplinary approach, 33-37. (2015).

[22] Kurth, F., Gaser, C., Sánchez, F. J., Luders, E.: Brain sex in transgender women is shifted towards genderidentity. Journal of Clinical Medicine, 11(6), 1582. (2022).

[23] Kurth, F., Gaser, C., Luders, E.: Development of sex differences in the human brain. *Cognitive neurosci-ence*, 12(3-4), 155-162. (2021).


DRAFT

# A latent variable approach to Millennials' knowledge of green finance

Maria Iannario[a], Alessandra Tanda[b], and Claudia Tarantola[b]

[a]Via L. Rodinò, 22, Naples, University of Naples Federico II; maria.iannario@unina.it
[b]Via San Felice, 5, Pavia, University of Pavia,; alessandra.tanda@unipv.it, claudia.tarantola@unipv.it

**Abstract**

In this paper, we consider a latent variable model for Millennials' knowledge of green finance. We draw inspiration from a survey conducted last year among students at Italian universities, as part of the COST project 'Fintech and Artificial Intelligence in Finance'. In the analysis, we consider a cumulative link model that accounts for heterogeneity in the response process. The model presented is discussed in a Bayesian framework.

***Keywords:*** Heterogeneity of variance, MCMC, *numeracy*, ordinal responses, scale effects

DRAFT

## 1. Introduction

In the latest years, the attention devoted to climate change consequences has grown considerably together with the awareness on the need to pursue more sustainable growth and perform a transition to greener economies (16). The interest in this topic comes from many directions. Companies adopt greener practices to reduce their environmental impact to attract investors and reduce their cost of funding, also for reputational concerns, see (23) and (30) among others. Investors place a lot of emphasis on ESG performance in their investment decisions in order to drive economic growth according to their values and to achieve long-term sustainable performance; see e.g. (8), (10) and (29). Policymakers developed guidelines and action plans to promote sustainable growth and manage the risks of climate change, accompanying companies to the transition (15); (13); (27); and, in addition, civil society has become more aware and attentive to sustainability and to the impact of their choices as citizens.

Among consumers and potential investors, Millennials find themselves in the position to effectively contribute to greener growth, through their purchase and investment choices. Millennials are individuals born between 1980 and 2000 (26); they generally show a strong sensitivity towards sustainability (4) and different risk attitudes and investment habits than the older generations (3). However, as the literature shows, this focus on sustainability issues does not always translate into more sustainable purchasing (4; 25). The authors of (25) claim that the sustainable behaviour of Millennials depends on their specific characteristics, including their ecological knowledge or *eco-literacy* (20).

Starting from these findings rooted in the managerial and consumers' behaviour literature, we extend the idea to the financial sphere of decisions. As the availability of funds is crucial to the transition and to the development of a greener economic system (15), it is important that savings are redirected towards those activities that are able to pursue a more balanced and inclusive growth, also contributing to the achievement of the Sustainable Development Goals (SDGs), issued by the UN (see https://sdgs.un.org). Globally, Millennials are estimated to have contributed to a surge in the amount invested in ESG funds (24).

Nevertheless, their contribution may currently be hampered by a low level of knowledge in finance and, specifically, 'green finance'. The level of financial knowledge and education among households remains very low, especially among the young investors. Bank of Italy reports that in 2020 the average level of financial literacy in Italy was 11.2 (out of 21); see (11) and (14). Additionally, young people (less than 35 years old in the survey) show a lower level of financial literacy than their older peers. Young people scored less than the average, especially for the component 'financial attitude' that investigates the ability to plan financial decisions from a long-term perspective. In an international comparison, Italy scores poorly compared to other countries and this further underlines the critical situation of the young Italian generation in terms of financial literacy. Furthermore, the Edufin Committee provides evidence on the lack of knowledge of sustainability in finance and 'green finance' in Italy (14), reporting that around 40% of the surveyed population has never heard the word 'ESG' and 1/5 never heard about 'sustainable finance'. Similarly, Millennials perform slightly better for ESG, with about 35% having never heard of 'ESG'; 26% have never heard of 'sustainable finance'. To foster Millennials' participation as investors in the transition to sustainable growth, it is necessary to understand the factors that determine their knowledge and behaviour.

Based on the previous considerations, this paper addresses the assessment of the degree of knowledge of 'green finance' by Millennials and the factors that determine this outcome. We examine a set of data collected via a survey distributed among students at Italy's largest university. The aim of the survey was to evaluate students' financial literacy. To assess their knowledge, respondents reported scores to items on ordinal rating scales. The observed score can be considered as a discretization of an underlying unobserved (latent) continuous variable, with every possible score for the ordinal response corresponding to an interval of the latent variable. The analysis presented here has been performed by means of a latent trait model accounting for heterogeneity in the response process (22). We work in a Bayesian framework, gaining flexibility in specifying the model and enhancing richness and accuracy in providing parameter estimates, see (17) for an application in a student evaluation context. A further advantage of the Bayesian setup is the possibility to use the same framework and approach even if the sample size is small (as in the examined case). For a general discussion of the advantages of the use of a Bayesian approach for ordinal data see, e.g, (21) and references therein. For some milestones regarding Bayesian approach in the context of ordinal data see (1), (2), (12), (18) and (19).

To the best of our knowledge, no previous study investigates the degree of awareness of the existence of 'green finance' among Millennials and whether this knowledge depends on their specific characteristics and aptitudes, such as numeracy.

The paper is organized as follows. The next section provides a brief description of the considered latent trait model in its location-scale version. Section 3. introduces the examined survey data and presents the Bayesian estimates of the examined model. Section 4. presents some preliminary findings deriving from the Bayesian analysis of the examined model.

DRAFT

## 2. Latent trait model description

Let $(x_i, y_i), i = 1, \ldots, n$; be a data set of size $n$, where the covariates $x_i$'s are assumed to be non-stochastic in $R^p$. The $i$-th measurement $y_i$ is the realization of a random variable $Y \sim G(y)$ conditioned on $x_i$. Variable $Y$ takes values in the finite set $\chi = \{1, \ldots, k\}$. Following (22), we assume that there exists an unobserved latent random variable $Y_i^*$ such that when $\alpha_{j-1} < Y_i^* \leq \alpha_j$, then $Y_i = j\,, j \in \chi$. Here $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_k = +\infty$ are the thresholds in the continuous support of the latent variable $Y^*$.

The $i$-th rating of $Y^*$ linearly depends on $p \geq 1$ covariate(s) through $\boldsymbol{x}_i$ as $Y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \sigma_i\epsilon_i$, $i = 1, 2, \ldots, n$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the covariates coefficients. In the latent regression $\sigma_i$ is the standard deviation of the noise variable $\epsilon \sim F_\epsilon(.)$, which may depend on covariates yielding $\sigma_i = \exp(\boldsymbol{z}_i\boldsymbol{\gamma})$. Here $\boldsymbol{z}_i$ is a row vector of the matrix $\boldsymbol{Z}$ which includes all the $q \geq 1$ relevant covariates and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$ are the related covariates coefficients.

The probability mass function of $Y_i$, for $j = 1, 2, \ldots, k$, can be expressed as

$$\begin{aligned} Pr\left(Y_i = j \mid \boldsymbol{\theta}, \boldsymbol{x}\right) &= Pr\left(\alpha_{j-1} < Y_i^* \leq \alpha_j\right) \\ &= F_\epsilon\Big[(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma_i\Big] - F_\epsilon\Big[(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma_i\Big]. \end{aligned}$$

Among the alternative choices for $F_\epsilon(.)$ we focus on the logit link function for robustness properties.

Since we do not have relevant prior information, following the approach prosed by (6) and (7), we use non-informative priors on all parameters of interest, letting the data guide the behavior of the posterior distributions. More precisely on the covariate coefficients we assign improper uniform priors, $unif(-\infty, +\infty)$, while on the thresholds we consider Student-$t$ priors with 3 degrees of freedom. This ensures that the tails are wide, while the distribution is still proper with finite mean and variance (here set –without loss of generality– equal to 0 and 2.5 respectively). In order to obtain posterior samples we rely on Markov Chain Monte Carlo (MCMC) method. In particular, we use `R` package `brms` (6), which implements a Hamiltonian MCMC using Stan (5). The ordering of the intercepts is ensured via the `order class` in Stan. More precisely the joint prior distribution is truncated to support over points satisfying the ordering constraints.

## 3. Data

The data come from the survey on 'Knowledge and Use of Fintech Products' administered as part of the European project CA19130 Fintech and Artificial intelligence in Finance. The sample consists of 385 Italian Millennials, of whom 0.511 are women, 0.563 are from Southern Italy, 0.210 are studying economics at university. The distribution of the dependent variable on the level of knowledge of green finance (ranging from 0 =I don't know to 5 =I know it perfectly) is in Figure 1.

Among the *numeracy* variables considered are the self-assessment of how good the respondent perceives himself/herself to be with fractions (0.388 Millennials give a good rating) and the correct answers to the Berlin numerical test (multiple-choice format) (9) in which the respondents who correctly formulated all answers to the four questions are indicated by *top* (0.035 respondents). Difficulties with financial language and stress in relation to financial decisions are expressed by 0.703 and 0.696 of the respondents, respectively.

The Bayesian estimates of the location and scale parameters are reported in Table 1 (posterior mean, MCMC Standard Error and 95% credible intervals). Standard convergence diagnostics have been considered. Possible interactions among covariates were tested, but they were related to not statistically significant parameters.

We run in parallel 4 chains of 2000 iteration with a burnin period of 1000 iteration each. The Bayesian estimate of the standard deviation is obtained from the posterior samples of log-disc (log-discrimination) with disc corresponding to the inverse of the standard deviation. More precisely, for every iteration $t$, $t = 1, 2, \ldots, T$, we transformed $\log(disc)^t$ to $sd^t$ with $sd^t = 1/exp(log(disc))$; the Bayesian estimates of $sd$ is then obtained as the average of all $sd^t$.

## 4. Preliminary results

Preliminary results show that the knowledge of 'sustainable finance' or 'green finance' is associated with a set of financial education items and *numeracy*. We also observe that knowledge of Fintech innovations, such as Crowdfunding, Roboadvisor and advanced technologies (AI) is associated with knowledge of 'green finance'. These two elements, financial technology, and sustainability are, in fact, often coupled together by policymakers to foster more inclusive growth. It is therefore not surprising that Millennials who are more attentive to Fintech innovations are also more aware of green finance issues.

In addition, Millennials attending university courses in the field of economics have a greater knowledge of green finance. This could be due to the inclusion of ESG and sustainability topics in courses

DRAFT

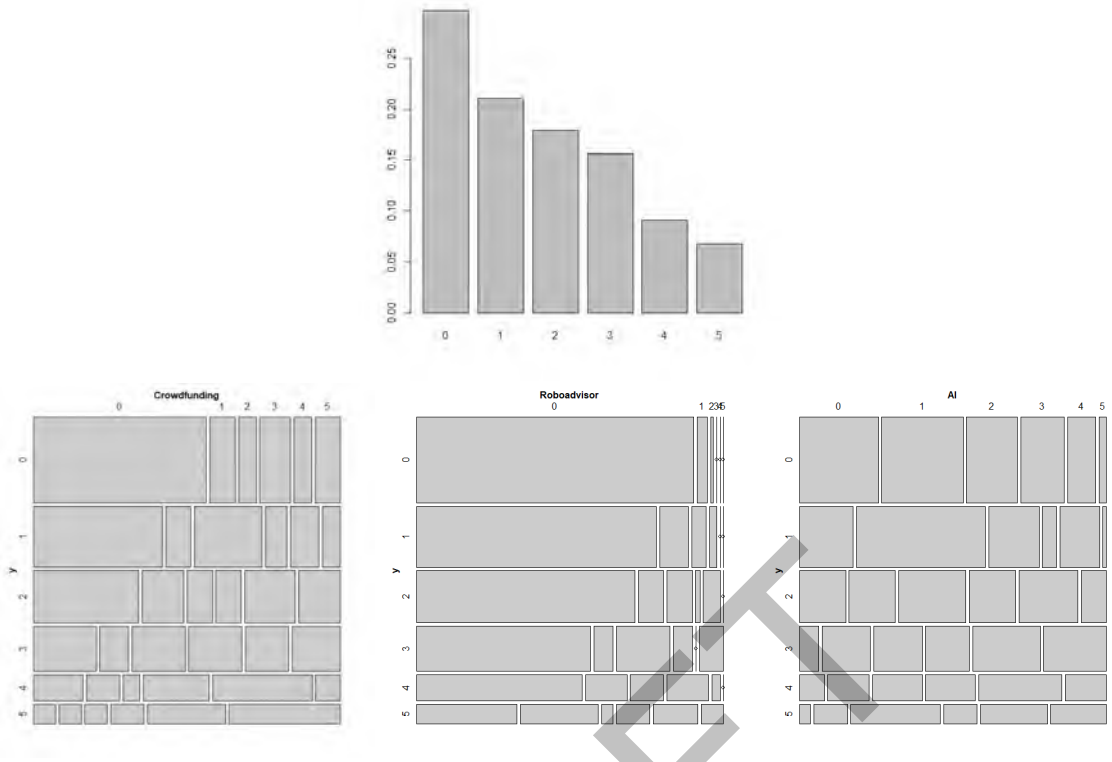



Figure 1: First row: Frequency distribution of the ordinal variable. Second row: Mosaic plot for the dependent variable vs Crowdfunding (first), Roboadvisor (second), and Artificial Intelligence - AI (third).

DRAFT

taken by university students in the field of finance. Numerical skills provide two seemingly counterintuitive results. Those who have a better ability to use fractions have a greater knowledge of 'green finance'. But the better performers in the Berlin numeracy test have less knowledge of our variable of interest $Y$.

*Numeracy* is a strong component of financial education and a basic skill for taking good financial decisions (28). However, in the field of 'green finance', this does not seem to be enough. In our analysis, the Millennials who score highest on *numeracy* tests have the least knowledge of green finance. This result may be determined by those Millennials studying in a STEM degree programme, who have excellent numerical skills, but have not yet encountered the topic of 'green finance'. We then evaluate the effect of financial knowledge in terms of familiarity with financial language and decisions.

Respondents who find financial language more difficult are also less knowledgeable about 'green finance'. At the same time, those who are more stressed when making financial decisions are more familiar with green finance. This could be due to the fact that stressed people might learn more before making decisions and, by doing so, become more aware of the different instruments available in the financial markets. This aspect probably deserves further investigation.

Our results are relevant for policy makers and for the design of future financial education initiatives. Further efforts should be devoted to increasing awareness of these key sustainability issues, along with general financial knowledge and *numeracy*.

**Acknowledgments** This work acknowledges research support by COST Action CA19130 'Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry' (FinAI), supported by COST (European Cooperation in Science and Technology).

Table 1: Bayesian estimation for the location-scale model; i.e., posterior mean estimates, standard deviations and $95\%$ Credible Intervals (CI) for its parameters

| | Estimate(Sd) | l-95% CI | u-95% CI |
|---|---|---|---|
| $\hat{\alpha}_1$ | 0.80(0.97) | -0.97 | 2.92 |
| $\hat{\alpha}_2$ | 5.72(1.62) | 3.10 | 9.56 |
| $\hat{\alpha}_3$ | 10.20(2.47) | 6.20 | 15.73 |
| $\hat{\alpha}_4$ | 15.64(3.61) | 9.83 | 23.70 |
| $\hat{\alpha}_5$ | 21.86(5.10) | 13.78 | 33.36 |
| *Crowdfunding* | 1.00(0.32) | 0.47 | 1.68 |
| *Roboadvisor* | 1.54(0.52) | 0.62 | 2.69 |
| *AI* | 0.92(0.35) | 0.33 | 1.72 |
| *economics* | 2.82(1.19) | 0.72 | 5.33 |
| *fractions* | 0.99(0.34) | 0.44 | 1.74 |
| *top* | -6.26(2.90) | -12.62 | -1.36 |
| *financial-language* | -1.47(0.47) | -2.51 | -0.69 |
| *stress-fin* | 0.90(0.42) | 0.19 | 1.79 |
| *log_disc_info* | 1.23(0.06) | 1.11 | 1.35 |
| *log_disc_franctions* | 1.13(0.04) | 1.05 | 1.22 |
| *log_disc_sud* | 1.26(0.15) | 0.99 | 1.57 |

DRAFT

## References


[1] Albert, J. H. and Chib, S.: Bayesian Analysis of Binary and Polychotomous Response Data. *JASA*, **88**, 422, 669–679 (1993)

[2] Albert, J. and Chib, S.: Bayesian methods for cumulative, sequential and two step ordinal data regression models. Technical report (1997)

[3] Beck, J. J., Garris III, R. O.: Managing personal finance literacy in the United States: A case study. Educ. Sci., **9**, 129 (2019)

[4] Bernardes, J. P., Ferreira, F., Marques, A. D., Nogueira, M.: Millennials: is 'green' your colour?. In IOP Conference Series: Materials Science and Engineering (Vol. 459, No. 1, p. 012090). IOP Publishing (2018)

[5] Betancourt, M., and Girolami, M.: Hamiltonian Monte Carlo for Hierarchical Models. arXiv 1312.0906. http://arxiv.org/abs/1312.0906 (2013)

[6] Bürkner, P.: brms: An R Package for Bayesian Multilevel Models Using Stan.J. Stat. Softw., **80**, 1–28 (2017)

[7] Bürkner, P., Vuorre M.: Ordinal Regression Models in Psychology: A Tutorial. AMPPS, **2**, 77-101 (2019)

[8] Chauhan, Y., Kumar, S. B.: Do investors value the nonfinancial disclosure in emerging markets?. Emerg. Mark. Rev., **37**, 32-46 (2018)

[9] Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., Garcia-Retamero, R.: Measuring risk literacy: The Berlin Numeracy Test. JDM, **7**, 25-47 (2012).

[10] Cornell, B. ESG preferences, risk and return. Eur. Financial Manag., **27**, 12-19 (2021)

[11] D'Alessio, G., De Bonis, R., Neri, A., Rampazzi, C.: Financial literacy in Italy: The results of the Bank of Italy's 2020 survey. Politica economica, **37**, 215-252 (2021)

[12] Dellaportas, P. Smith, A.F.M.: Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. JRSS C, 42, 443-459 (1993)

[13] ECB: Climate-related risk and financial stability. July 2021. Available at https://www.ecb.europa.eu/pub/pdf/other/ecb.climateriskfinancialstability202107 87822fae81.en.pdf

[14] Edufin: Rapporto Edufin 2021. Available at https://www.quellocheconta.gov.it/it/news-eventi/rassegna/Rassegna-Stampa/news_138.html

[15] European Commission: The European Green Deal, COM(2019) 640 final, 11 December. $https : //ec.europa.eu/info/sites/info/files/european - green - deal - communication_en.pdf.$

[16] Hafner, S., Jones, A., Anger-Kraavi, A., Pohl, J.: Closing the green finance gap. A systems perspective.Environ. Innov. Soc. Transit., **34**, 26-60 (2020)
[17] Iannario, M., Kateri, M., Tarantola, C.: Modelling scale effects in rating data: a Bayesian approach. Manuscript (2023)
[18] Johnson, V. E.: On Bayesian analysis of multirater ordinal data: An application to automated essay grading. JASA, 91, 42-51 (1996)
[19] Johnson V.E., Albert J.H.: Ordinal Data Modeling. Springer-Verlag, New York (1999)
[20] Kanchanapibul, M., Lacka, E., Wang, X., Chan, H. K.: An empirical investigation of green purchase behaviour among the young generation. J. Clean. Prod., **66**, 528-536 (2014)
[21] Liddell, T. M., Kruschke, J.K.: Analyzing ordinal data with metric models: What could possibly go wrong? J. Exp. Soc. Psychol., **79**, 328–348 (2018)
[22] McCullagh, P.: Regression models for ordinal data (with discussion). *J. R. Stat. Soc. B*, **42**, 109–142 (1980)
[23] Miroshnychenko, I., Barontini, R., Testa, F.: Green practices and financial performance: A global outlook. J. Clean. Prod., bf147, 340-351 (2017)
[24] MSCI: Swipe to invest: the story behind millennials and ESG investing. (2020) Available at https://www.msci.com/documents/10199/07e7a7d3-59c3-4d0b-b0b5-029e8fd3974b
[25] Naderi, I., Van Steenburg, E.: Me first, then the environment: Young Millennials as green consumers. Young Consumers (2018)
[26] Ng, E. S., Schweitzer, L., and Lyons, S. T.: New generation, great expectations: A field study of the millennial generation. JBP, **25**, 281-292 (2010)
[27] OECD: Financial Markets and Climate Transition: Opportunities, Challenges and Policy Implications, OECD Paris, (2021) https://www.oecd.org/finance/Financial-Markets-and-ClimateTransition-Opportunities-challenges-and-policy-implications.htm
[28] Sunderaraman, P., Barker, M., Chapman, S., Cosentino, S.: Assessing numerical reasoning provides insight into financial literacy. Appl. Neuropsychol. Adult, **29**, 710-717 (2022)
[29] Van Duuren, E., Plantinga, A., Scholtens, B.: ESG integration and the investment management process: Fundamental investing reinvented. J. Bus. Ethics, **138**, 525-533 (2016)
[30] Yu, E. P. Y., Tanda, A., Luu, B. V., Chai, D. H.: Environmental transparency and investors' risk perception: Cross-country evidence on multinational corporations' sustainability practices and cost of equity. BSE, **30**, 3975-4000 (2021)

DRAFT

# Archetypal analysis and latent Markov models: A step-wise approach

Lucio Palazzo[a], Rosa Fabbricatore[b], and Francesco Palumbo[a]

[a]Department of Political Sciences, University of Naples Federico II; lucio.palazzo@unina.it, fpalumbo@unina.it
[b]Department of Social Sciences, University of Naples Federico II; rosa.fabbricatore@unina.it

**Abstract**

In this contribution we exploit probabilistic archetypal analysis in a three-step approach involving latent Markov models to analyze discrete latent variables with a discrete-time follow-up scheme. Archetypal analysis provides class assignments that are subsequently used as single indicators in a latent Markov model to estimate the structural part. We apply the proposed strategy to a dataset concerning responses to a statistical anxiety questionnaire administered to university students attending an introductory statistical course.

***Keywords:*** latent Markov models, probabilistic archetypal analysis, statistical anxiety

DRAFT

## 1. Introduction

The analysis of longitudinal data (1) plays a relevant role in many fields concerning the human sciences, where many research questions concern the study across time. In general, a longitudinal data set consists of $S$ individuals (panel), each of which a set of $M$ items is measured at equally or unequally spaced time intervals. Various statistical models exist to analyze longitudinal data (2) and the specific application context leads to choosing the most appropriate. Among them, latent Markov models (LM, 8) play a relevant role and are considered in the present contribution. LM models basic assumptions are that the latent process follows a Markov chain process with a certain number of latent states and that the response variables, in the multivariate case, are conditionally independent (local independence). LM models are a statistical reference approach to study the state changes in individual or group characteristics over time among those considering discrete latent variables measured at different time points through a set of observed variables, while other external variables (covariates) can act as predictors of manifest or latent variables. However, this approach usually returns homogeneous groups and therefore it identifies profiles characterized by specific class response probabilities, rarely defining peculiar cases.

The goal of the present contribution is to allow for the study of an LM model that accounts for the switching between latent classes over time. According to (12), latent class (LC) analysis aims to assign individuals to a set of few discrete latent classes. The novelty of the proposal consists in the use of the archetypal analysis (AA) framework (6; 3; 11) to address the heterogeneity of the latent classes. Let $Y$ be a $N \times M$ pooled responses matrix of $S$ distinct subjects to $M$ items collected to the same panel over $T$ time intervals, where its generic element $y_{nm} \in \{1, \ldots, I\}$ is the response of the $n$-th individual to the $m$-th item. It is worth noticing that $N \leq S \cdot T$ and the equality holds in case of zero dropouts. Throughout the whole paper, the superscript $(t)$ is used whenever it is necessary to specify the time $t$.

The proposed procedure consists of three steps: Step 1 exploits the AA on the pooled data, without covariates, to find homogeneous groups defining profiles characterized by specific class response probabilities; Step 2 calculates the state membership for each individual at any considered time point; Step 3

estimates an LM model to investigate individual transitions among the considered groups over time and the effect of a set of covariates on the initial classification and transitions. Because of their high feasibility, the step-wise estimation approach is considered in this work. We apply the proposed method to a dataset concerning responses to a statistical anxiety psychometric questionnaire that was administered to a convenience sample of undergraduate students attending an introductory statistics course.

The present contribution is organized as follows: in Section 2. the proposed approach is presented along with the model formulations, while in Section 3. the strategy is applied in the considered context and the main results are described. In conclusion, Section 4. outlines some final remarks.

## 2. A three-step PAA-LM

The vast specific literature on data clustering offers several alternative clustering approaches; in this context, we propose the Archetypal Analysis (AA), initially introduced in (6), as an alternative to the most known and used data partitioning methods. Differently from other clustering methods, AA focuses on identifying extreme points, namely archetypes, preferring the maximum heterogeneity among groups to the maximum homogeneity within groups. In its first presentation, AA deals with continuous data; more recently, different variations and algorithms have been presented for a more robust analysis (see, e.g., 7). Among them, (11) proposed a probabilistic AA model (PAA) to extend its applicability to manifest variables according to several discrete distribution families. It allows identifying a few extreme profiles corresponding to the archetypes and the related groups according to the membership degrees. In the considered application, the resulting latent classes refer to more or less severe anxiety levels and, more in particular, to statistical anxiety (SA), which is inferred through a set of items measured according to an ordinal scale. In the following, the overall method is summarized.

*1. Probabilistic Archetypal Analysis.* When the dataset is composed of ordinal variables, as in our case, it is possible to consider them as originating from a multinomial model. In such a case, a multinomial PAA is defined by assuming that the observed values can be reconstructed by a linear combination of the underlying multinomial probabilities. Here, the PAA is applied to the pooled dataset $Y$.

The goal is to find, given $K$ the number of archetypes, two matrices $W$ (with dimension $K \times N$) and $H$ (with dimension $N \times K$) that solve the following optimization problem

$$\arg \min_{W,H \geq 0} -\sum_n \sum_m y_{nm} \log\left(\tilde{y}_{nm}\right), \qquad \text{where } \tilde{Y} = HWP, \tag{1}$$

subject to $\sum_j h_{nj} = 1$ and $\sum_i w_{ki} = 1$. The $\tilde{Y}$ is the reconstructed matrix of the observed values and the vector $p_{n\cdot}$ is the maximum likelihood point estimate from observation $y_{n\cdot}$. After obtaining the estimated matrices $W$ and $H$, it is possible to compute the archetypal profiles $Z = WP$ and the simplex plot through the use of the membership matrix $H$. Moreover, it is worth remembering that even if the archetypal profiles lie in the observation space and define "extreme" profiles of responses, they can not correspond to real observations of the sample.

*2. Subjects' membership assignment.* Given the properties of the coefficient in $H$, there are different approaches allowing to cluster of the data around the archetypes, see (10). The usual method, namely the crisp allocation rule, allocates responses in a cluster based on the nearest archetype. In addition, if we consider the scores to be values of a fuzzy membership function, then it is possible to assign the data to an additional central group. The fuzzy allocation rule (FAR) is summarized as follows

$$\mathcal{C}_k = \{y_{n\cdot} \,|\, h_{ik} > \tau\} \qquad\qquad k = 1, \ldots, K, \quad 0 < \tau < 1. \tag{2}$$

The threshold $\tau$ acts as a "purity" parameter: some respondents are assigned to the "closest" archetype, while the respondents whose membership scores do not reach the threshold are assigned to the so-called central class. In particular, if $\tau \geq 0.5$ then each data point is classified as belonging to only one group, but at the same time not all the data points will be classified. The "discarded" observations will lie in the so-called central group $\mathcal{C}_+ = \{y_{n\cdot} | y_{n\cdot} \notin \mathcal{C}_k, k = 1, \ldots, K\}$ that, along with the other groups, will induce an augmented partition of the data consisting of $K^* = K + 1$ classes.

DRAFT

*3. Structural model estimation.* At this point, the considered LM model is applied given the latent class memberships obtained through the FAR method. LM modeling framework is used to model the change in latent class membership over time concurrently accounting for the effect of covariates on initial and transition probabilities. The latent process is assumed to follow a first-order Markov chain, where changes only depend on the previous class. According to the three-step approach, the model is specified to have the modal class assignment obtained at Step 2 as a single fixed indicator.

Denote with $g_s^{(t)}$ the vector of the considered covariates for individual $s$ at time $t$. Let $X_s^{(t)}$ be the categorical latent variable for the $t$-th time point, taking values: $1, \ldots, K^*$. Logistic models can be used to model the effect of the covariates on initial and transition probabilities. In particular, taking the class $\mathcal{C}_+$ (corresponding to the $K^*$-th class) as the reference category, we have

$$\log \frac{P\big(X_s^{(1)} = k | g_s^{(1)}\big)}{P\big(X_s^{(1)} = K^* | g_s^{(1)}\big)} = \beta_{0k} + \boldsymbol{\beta}_k' g_s^{(1)}, \qquad k = 1, \ldots, K,$$

for initial probability, where $\beta_{0k}$ and $\boldsymbol{\beta}_k$ are respectively the intercept and the vector of coefficients for the covariate effects, and

$$\log \frac{P\big(X_s^{(t)} = k | X_s^{(t-1)} = l, g_s^{(t)}\big)}{P\big(X_s^{(t)} = K^* | X_s^{(t-1)} = l, g_s^{(t)}\big)} = \gamma_{0k} + \gamma_{0lk} + \boldsymbol{\gamma}_k' g_s^{(t)}, \quad k, = 1, \ldots, K, \quad l = 1, \ldots, K^*,$$

for transition probabilities, where $\gamma_{0k}$ is the intercept for the $k$-th latent class, $\gamma_{0lk}$ the intercept for the considered transition, and $\boldsymbol{\gamma}_k$ the vector of covariate coefficients. Coefficients related to the last category are set to zero because the last class is taken as the reference category. Moreover, transition probabilities are assumed to be time-homogeneous. Parameter estimation for the third step of the proposed strategy is performed employing the Maximum Likelihood (MML) approach using Latent GOLD 6.0 (13).

DRAFT

## 3. Application

The proposed approach was implemented to investigate students' SA during an introductory Statistics course. SA can be defined as "an anxiety that comes to the fore when a student encounters Statistics in any form and at any level" (9). SA can cause worry and discomfort when students are exposed to evaluative contexts, instructional situations, or just Statistics content or problems. The *Statistical Anxiety Scale* (SAS, 5) aims to assess these different aspects of anxiety by means of 24 items with a 5-point Likert scale ranging from 1 (no anxiety) to 5 (very much anxiety). The instrument includes 3 subscales: *(i) Examination anxiety* (8 items), referring to the anxiety encountered during a Statistics class or a Statistics test; *(ii) Fear for asking for help* (8 items), relating to the anxiety experienced when asking a peer, a tutor, or a professor for help in understanding statistical contents; *(iii) Interpretation anxiety* (8 items), occurring when students have to interpret or make a decision about statistical data. In the literature several antecedents of SA have been proposed; herein, we consider sex, math knowledge, academic motivation, self-efficacy, test anxiety, cognitive strategy, self-regulation, students' attitude toward Statistics (dimensions: affect, cognitive, value, difficulty), engagement (dimensions: affect, cognitive, behavioral).

The study involves 196 students enrolled in the psychology degree course at the University of Naples Federico II in Italy, attending an introductory Statistics course. Participants were predominantly females (75%), with a mean age of 19.86 years (sd = 2.88). At the beginning of the course, students were invited to fill out a questionnaire including socio-demographic questions and psychometric scales assessing psychological variables related to learning Statistics and SA. In addition, SA, attitudes toward Statistics, and students' engagement were measured three times during the course, at the end of each main statistical learning module. Indeed, the course may impact students' anxiety and attitude toward Statistics and also the level of engagement can change over time. Data collection was carried out through the MOODLE platform.

## 3.1 Main findings

Firstly, let us look at the measurement part of the model estimated in Step 1. In Figure 1 the profiles obtained with our approach are compared with those resulting from the classical LC analysis. Dotted vertical lines divide the items according to the SA subscales that are, in order, examination, asking for help, and interpretation. The colors are intended only to distinguish the profiles within each procedure. In the top panel the resulting archetypes are depicted, basing on the optimal choice of $k = 4$ and the addition of the central class, for a total of 5 distinct profiles. At first glance, the leftmost portion of the plot shows that all the archetypes have high tendency scores on examination anxiety, evidencing that students of the considered sample generally exhibit high anxiety levels during the examination. Conversely, the profiles cover a wider range of scores for what concerns the remaining anxiety dimensions.

Figure 1: Archetypal (top) versus latent (bottom) class profiles.

In more detail, the archetypes A1 and A2 (colored in black and red, respectively) present a similar anxiety level for examination and asking for help dimensions, while they present different patterns for the interpretation anxiety. Specifically, the group defined by A1 encompasses the respondents with low levels of asking-for-help and interpretation anxiety, whereas A2 is representative of the group with students with the highest level of examination and interpretation anxiety. The remaining A3 (green line) and A4 (blue line) present some peculiarities. The archetype A3 describes a profile with middle scores in all the anxiety components, representing students who tend to select the intermediate category of the response scale (i.e., $i = 3$), regardless of the encountered items. This result leads us to suppose that A3 caught a specific response style. On the other hand, A4 is representative of students who experience high levels of anxiety in both examination and asking-for-help situations. Focusing on the second subscale, it can be noticed that, despite the overall high levels of anxiety, A4 is characterized by low scores related to SAS5 ("Asking a private teacher to explain a topic that I have not understood at all") and SAS24 ("Asking a private teacher to tell me how to do an exercise") items. Thus, students belonging to the group defined by A4 present a high level of asking-for-help anxiety but don't feel anxious when they request a private teacher to explain a topic. Finally, A5 (light blue line) represents the central class and is defined by averaging the scores of the respondents belonging to the central group $\mathcal{C}_+$. About the weights of the classes defined by the obtained archetypes, results show that the largest class is $\mathcal{C}_+$ (34%), followed by $\mathcal{C}_3$ (20%), $\mathcal{C}_1$ (18%), $\mathcal{C}_4$ (17%), and $\mathcal{C}_2$ (11%).

Regarding the structural part of the model, Table 1 displays the matrix of transition probabilities. Generally, students tend to remain in the same group over time, showing stability in their SA profile. For

DRAFT

what concern the transitions, most of them occur toward the central class, indicating gradual changes in students' anxiety over time. On the other hand, students belonging to $\mathcal{C}_+$ transit mainly towards classes defined by archetypes A1 and A3.

Table 1: Estimated transition probabilities.

| Class at time $t-1$ | Class at time $t$ | | | | |
|---|---|---|---|---|---|
| | $\mathcal{C}_1$ | $\mathcal{C}_2$ | $\mathcal{C}_3$ | $\mathcal{C}_4$ | $\mathcal{C}_+$ |
| $\mathcal{C}_1$ | 0.490 | 0.066 | 0.087 | 0.055 | 0.302 |
| $\mathcal{C}_2$ | 0.117 | 0.499 | 0.105 | 0.037 | 0.242 |
| $\mathcal{C}_3$ | 0.096 | 0.026 | 0.532 | 0.126 | 0.220 |
| $\mathcal{C}_4$ | 0.043 | 0.047 | 0.147 | 0.492 | 0.271 |
| $\mathcal{C}_+$ | 0.112 | 0.036 | 0.233 | 0.094 | 0.525 |

In Table 2 are summarized the significant effects of the of individual covariates on initial classification and transitions, considering the central class $\mathcal{C}_+$ as reference. Statistical significance is evaluated through the Wald test. Results show that autonomous regulation in academic motivation (measured by the relative autonomy index) increases the probability of belonging to $\mathcal{C}_1$ at the beginning of the course, indicating a lower level of asking-for-help and interpretation anxiety. Math ability, instead, reduces the probability of having a high level of examination and interpretation anxiety (typical of students in $\mathcal{C}_2$).

Table 2: Regression parameters' estimates and standard errors (in parentheses)

| Latent class | Initial probability | | | Transition probability | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sex | Math ability | RAI | Sex | SATS Affect | SATS Cognitive | SATS Value | ENG Cognitive | ENG Behavioral |
| $\mathcal{C}_1$ (vs $\mathcal{C}_+$) | -0.34 (0.50) | 0.07 (0.06) | **0.04 (0.02)** | 0.27 (0.38) | **0.10 (0.03)** | -0.02 (0.04) | 0.002 (0.02) | -0.48 (0.34) | **0.58 (0.29)** |
| $\mathcal{C}_2$ (vs $\mathcal{C}_+$) | 0.03 (0.66) | **-0.11 (0.05)** | 0.01 (0.02) | 0.93 (0.71) | -0.01 (0.04) | -0.04 (0.04) | 0.01 (0.03) | **-0.74 (0.42)** | **1.23 (0.43)** |
| $\mathcal{C}_3$ (vs $\mathcal{C}_+$) | -0.89 (0.54) | -0.07 (0.05) | 0.005 (0.02) | -0.02 (0.33) | **0.12 (0.03)** | **-0.12 (0.03)** | **-0.05 (0.02)** | 0.08 (0.26) | -0.08 (0.24) |
| $\mathcal{C}_4$ (vs $\mathcal{C}_+$) | -0.01 (0.52) | 0.04 (0.05) | -0.01 (0.02) | 0.13 (0.42) | **-0.06 (0.03)** | **0.07 (0.04)** | 0.02 (0.02) | **-0.87 (0.35)** | 0.23 (0.27) |

Note: Significant effects are reported in bold; RAI = Relative Autonomy Index, SATS = Survey of Attitudes toward Statistics, ENG = Student Engagement in Statistics.

DRAFT

A positive affect attitude towards Statistics (students' positive feelings concerning Statistics) increases the probability of being in $\mathcal{C}_1$ (low level of asking-for-help and interpretation anxiety) and $\mathcal{C}_3$ (students selecting the intermediate category of the response scale), whereas it reduces the probability for $\mathcal{C}_4$ (high level of asking-for-help anxiety). Conversely, the cognitive component of attitude towards Statistics (students' belief about their ability in Statistics) affects the probability of being in $\mathcal{C}_4$ positively and the probability of being in $\mathcal{C}_3$ negatively. Moreover, perceiving the relevance of Statistics in the personal and professional life (value component of attitude) reduces the probability of providing a middle-category response pattern (namely, to be in $\mathcal{C}_3$). Finally, about the effect of students' engagement in Statistics, the cognitive component (cognitive strategies to monitor their own learning of Statistics) negatively affects the probability of being in $\mathcal{C}_2$ and $\mathcal{C}_4$, both with low levels of interpretation anxiety, whereas the behavioral component (e.g., regular study, participation during lessons) influences the students' membership in $\mathcal{C}_1$ and $\mathcal{C}_2$, increasing the probability of having a low level of asking-for-help anxiety. Sex appears to be a non-significant factor for both initial and transition probabilities.

## 4. Concluding remarks

In many applications involving longitudinal data, the interest is focused on the evolution of a latent individual characteristic that is measured by one or more occasion-specific response variables. In such a context, LM may be usefully applied given their nature as discrete latent variable models with a discrete-time follow-up scheme. In this contribution, we exploit probabilistic archetypal analysis in a three-step approach to obtain class assignments that are subsequently used as single indicators in an LM model to estimate the structural part.

This approach could represent a helpful tool, especially in the educational context, where distinguishing peculiar profiles according to students' performance and/or psychological characteristics can drive the development of more tailored formative and motivational feedback (4). PAA also allows accounting for the ordinal nature of the observed variables, usually following a Likert-type scale when assessing motivational and emotional factors. Moreover, the use of archetypes gives better interpretability of the corresponding profiles and of the respondents belonging to their archetypal classes, thus making it easier to generate *ad hoc* recommendations for each archetypal class.

## References


[1] Zeger, S.L. & Liang, KY, (1986). Longitudinal data analysis for discrete and continuous outcomes, Biometrics, **42**(1), 121–130.

[2] Bartolucci, F., Farcomeni, A. & Pennoni, F. (2012). Latent Markov models for longitudinal data, CRC Press, Boca Raton.

[3] Bauckhage, C., & Thurau, C. (2009). Making Archetypal Analysis Practical. In DAGM-Symposium (pp. 272-281).

[4] Adabbo, B., Fabbricatore, R., Iodice D'Enza, A. & Palumbo, F. (2021). Statistics Knowledge assessment: an archetypal analysis approach. In C. Perna, N. Salvati, and F. Schirripa Spagnolo (Eds.), BoSP SIS2021 (pp. 1388-1393). Pearson, Milano.

[5] Chiesi, F., Primi, C., & Carmona, J. (2011). Measuring statistics anxiety: Cross-country validity of the Statistical Anxiety Scale (SAS). Journal of psychoeducational assessment, **29**(6), 559-569.

[6] Cutler, A., & Breiman, L. (1994). Archetypal analysis. Technometrics, **36**(4), 338-347.

[7] Eugster, M., & Leisch, F. (2009). From spider-man to hero-archetypal analysis in R.

[8] Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. Developmental psychology, **44**(2), 457-467.

[9] Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments–a comprehensive review of the literature. Teaching in higher education, **8**(2), 195-209.

[10] Ragozini, G., Palumbo, F., & D'Esposito, M. R. (2017). Archetypal analysis for data-driven prototype identification. Statistical Analysis and Data Mining: The ASA Data Science Journal, **10**(1), 6-20.

[11] Seth, S., & Eugster, M. J. (2016). Probabilistic archetypal analysis. Machine learning, **102**, 85-113.

[12] Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. The sage encyclopedia of social sciences research methods, **2**, 549-553.

[13] Vermunt, J. K. & Magidson, J. (2020). Upgrade manual for Latent GOLD 6.0. Statistical Innovations Inc., Belmont.


DRAFT

# From high school to university: academic intentions and enrolment of foreign students in Italy

Francesca Di Patrizio[a], Eleonora Trappolini[b], and Cristina Giudici[b]

[a] Istat; dipatriz@istat.it
[b] Sapienza University of Rome; eleonora.trappolini@uniroma1.it
[b] Sapienza University of Rome; cristina.giudici@uniroma1.it

**Abstract**

This work adopts a longitudinal perspective in analysing the academic intentions and enrolment of foreign students in Italy, by citizenship. Using logistic regression models, we investigate how citizenship is associated with the academic intentions and the transition to university of a national representative cohort of foreign students and their classmates, enrolled in the last year of Italian high school in 2015. We use a unique dataset linking survey data of the "Integration of the second generation", carried out by Istat in 2015, with administrative data on students' university enrolment, from the Ministry of University registers. We found higher academic intentions among foreign students, compared to their Italian classmates, with Ecuador and Peru, China and the Philippines showing a higher probability of academic intentions and enrolment than both Italians and the other groups.

**Keywords:** Foreign students, Citizenship, Integration of the second generation survey

DRAFT

## 1. Background

A vast literature shows that students' educational intentions are key predictors of their future educational attainments. However, concerning foreign students, academic pathways may differ widely, depending on structural factors and individual characteristics.

Educational pathways of foreign students have been widely analysed by international literature (Zapfe & Gross 2021). Looking at the European context, several comparative studies show that the educational opportunities and school outcomes of second-generation students from the same group of origin vary considerably across countries (Crul et al. 2012, Griga & Adjar 2014, Gabrielli et al. 2022). Despite lower school results, children of immigrants generally tend to exhibit higher aspirations than ethnic majority peers. A great body of literature refer to this finding as "immigrant optimism" (Kao and Tienda 1995), or "bold choices" (Jackson 2012). Concerning school-university transition, a large sociological literature exists, showing mixed results, whereas there is a lack of statistical and demographic studies. A general finding in this field is that a stratified secondary school system reduces the educational chances of students with migrant background. Conversely,

a low-stratified secondary school system improves the probability of people with a migrant background attaining a higher education degree (Griga & Hadjar 2014, Zapfe & Gross 2021).

In this context, existing research in Italy is characterised by two major weaknesses: studies generally focus on small samples, that are not representative of the national situation and do not allow for a detailed analysis by citizenship. Moreover, existing literature in Italy rarely addresses this subject using a longitudinal perspective.

The aim of this work is to empirically contribute to this literature by investigating the academic intentions and the transition to university of a cohort of students enrolled in the last year of the Italian High school in 2015. We use a unique dataset that links survey data (the 2015 Integration of the second generation, carried out by Istat) with administrative records on university enrolment in a.y. 2015/16, 2016/17 and 2017/18.

To frame our research, we formulated three research hypotheses: (1) academic intentions of foreign students are higher than intentions of their Italian classmates; (2) despite their higher academic intentions, foreign students are less likely to enrol in university; (3) differences by citizenship exist, reflecting social and cultural specificities.

## 2. Data and methods

We created a pooled longitudinal dataset linking survey data of the 'Integration of the second generation' (hereafter, ISG) carried out by Istat in 2015, and administrative data on university enrolment from the Student register of the Ministry of University ("Anagrafe dello studente" hereafter, ANS-MUR).

The ISG survey collects information on a sample of 1,400 school of secondary education with at least 5 foreign students (more than 68,000 students). ISG analyses socio-demographic characteristics, behaviours and attitudes of students with a migrant background, namely those students without the Italian citizenship (Conti & Prati, 2020). In addition, this survey allows to distinguish the different models of integration of the first ten nationalities (Romania, Albania, Morocco, China, Philippines, Moldova, Peru, Ukraine, Ecuador and India) and the age at arrival, considering the age of enrolment at different levels of education in Italy (born in Italy, arrived before 6 years old, between 6 and 10 years old, and 11+)[1]. Since our work studies differences in the academic aspirations and enrolment of Italian and foreign students, from this survey we selected only students enrolled in the last year of the Italian high school in 2015.

The ANS-MUR archive contains information on students enrolled and those who obtained a degree in an Italian University.

The ISG and ANS-MUR record linkage, performed using an individual anonymised code subsequently removed, allows to follow and distinguish, among the students enrolled in the last year of the Italian high school in 2015 (previously selected from the ISG survey), those that enrolled at university within three years from 2014/2015 and those who decided not to enrol at university.

Therefore, our final sample is composed of 3,940 students (46.6% foreign students).

We used two dependent variables. The first one is the 'academic intention' of students, which is derived from the ISG survey with different possible answers including 'go to University', 'go to work', 'follow a vocationally-oriented course', 'stay at home', and 'I do not know[2]'; and the 'academic enrolment' of students derived from the ANS-MUR archive. For analysis purposes, the 'academic intention' variable takes value 1 when the student's answer is 'go to University', 0 otherwise. The 'academic enrolment' variable takes value 1 if the student enrolled within three years from 2014/2015 (i.e., within 2015/16, 2016/17 and 2017/18)[3], 0 otherwise.

[1] For further details, see Istat (2017), Conti, C., Quattrociocchi L. (a cura di), L'indagine sull'integrazione delle seconde generazioni: obiettivi, metodologia e organizzazione, 2017.

[2] The cases that answered 'I do not know' were excluded from the sample.

[3] We chose to follow students within three years from 2014/2015 because according to ad-hoc elaborations on the ANS-MUR archive, we observed that in the a.y. 2018/19, 75% of the second-generation enrolled had graduated from high

DRAFT

To test our research hypotheses, we applied two logistic regression models and conducted two separate analyses. In the first analysis, foreign status (foreigners vs Italians) was the main independent variable. In the second analysis, to overcome the Italians-foreigners dichotomy, we analysed any differences between Italians and foreign subgroups where we distinguished 9 communities (Romania, Albania, China, Philippines, Morocco, Moldova and Ukraine, Ecuador and Peru, Other HDC[4], Other HMPC[5]).

In all analyses, we used a set of control variables which include socio-demographic characteristics, the student's school career, the attitude towards studying of both students and their families and the relational aspects of young people with their peers.

We applied population weights provided in the dataset. First, we estimated the odds ratios. Then, to improve the readability of results we computed predicted probabilities of the outcomes with 95% confidence intervals for pairwise comparisons. In addition, confidence intervals were centred on the predictions and had lengths equal to 2 × 1.39 × standard errors. This was necessary to obtain an average level of 5% for Type I errors in pairwise comparisons of a group of means (Goldstein & Healy, 1995).

## 3. Preliminary results

Figure 1 shows the results of the adjusted predicted probabilities for our first dependent variable, 'academic intentions', by foreign status (Figure 1a) and citizenship (Figure 1b).

Looking at the results considering migrants as a heterogeneous group, we observed that foreign students show higher academic intentions than their Italian classmates (55.2% and 50.2%, respectively) (Figure 1a). However, the analysis by citizenship highlights differences between and among the different foreign subgroups examined. Specifically, we found that students from Ecuador and Peru, the Philippines and China register a higher probability of academic intentions than both Italians and the other groups (75.4%, 70.7% and 65.4%, respectively). Conversely, Romanian students show the lowest probability of academic intentions than all the other subgroups (43.4%). Students from Albania, Morocco, Ukraine and Moldova, and other HDC and HMPC do not show significant differences in the academic intentions with respect to Italian classmates (Figure 1b).

Figure 2 shows the adjusted predicted probabilities for our second dependent variable, 'university enrolment', by foreign status (Figure 2a) and citizenship (Figure 2b). By comparing Italians and foreigners (considering them as a single group) results show no differences between the two groups (Figure 2a), while the analysis by citizenship reveals that the subgroups that show higher academic intentions (Ecuador and Peru, China and Philippines) are the same ones that also show a higher probability of academic enrolment (60.8%, 60.2% and 57.1%, respectively). We detected that students from Romania (35.7%) and other HDC (36.0%) show the lowest probability compared to Italian classmates and the other subgroups. Finally, students from Albania, Morocco, Ukraine and Moldova, and other HMPC do not show significant differences in the university enrolment with respect to Italian classmates (Figure 2b).

DRAFT

school in the same year, 14% one year earlier, and 4% two years earlier, compared to values registered by Italian students of 83%, 8% and 2%, respectively.

[4] Foreigners coming from Highly Developed Countries.

[5] Foreigners coming from High Migratory Pressure Countries.

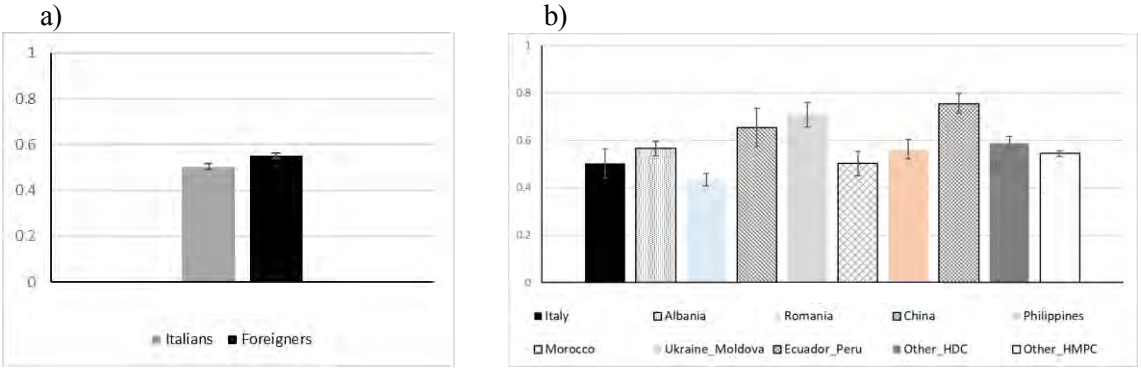



Figure 1: Adjusted predicted probabilities of the 'academic intention' by foreign status (panel a) and by foreign subgroups (panel b)

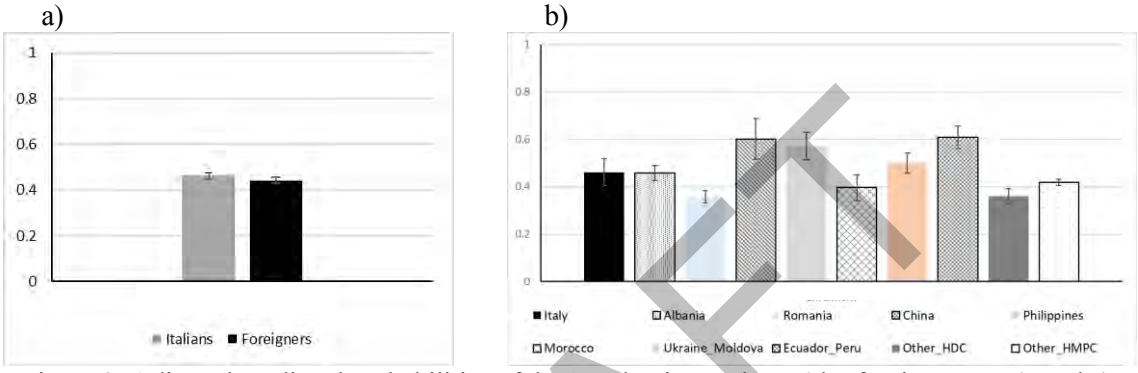



Figure 2: Adjusted predicted probabilities of the 'academic enrolment' by foreign status (panel a) and by foreign subgroups (panel b)

DRAFT

## 4. Discussion and further development

This work wants to contribute to the contemporary discussion on ethnic diversity as a major challenge for policy-makers, given the pivotal role that is played by education in the successful economic and social integration of children of immigrants.

Using a unique dataset which links the ISG survey data conducted by Istat in 2015 and the ANS-MUR archive, which contains administrative records, this is the first study which analyses academic intentions and the transition to university of foreign students in Italy. Results confirm our research hypotheses: overall, we found that academic intentions of foreign students are higher than intentions of their Italian classmates. Even though results show no differences between the two groups in terms of university enrolment, a more detailed analysis allows us to highlight differences by citizenship. Although the analysis of the factors underlying these patterns go beyond the scope of this work, we could speculate that differences by citizenship reflect social and cultural specificities of the highly heterogeneous foreign population residing in Italy. In order to deeply understand the mechanisms that explain educational choices of foreign students enrolled in Italian High School, further development of this study envisages to take into account the age at arrival, as a proxy of the social and educational integration level.

## References

[1] Conti, C., Prati, S. (2020). Identità e Percorsi di Integrazione delle Seconde Generazioni in Italia. Vita e Percorsi di Integrazione degli Immigrati in Italia [Identity and Integration Paths of the Second Generations in Italy. Life and Integration Paths of Immigrants in Italy].

[2] Crul, M., Schnell, P., Herzog-Punzenberger, B., Wilmes, M., Slootman, M., & Gómez, R. A.: School careers of second-generation youth in Europe. The European second generation compared, Amsterdam University Press (2012).
[3] Gabrielli, G., Longobardi, S., Strozza, S.: The academic resilience of native and immigrant-origin students in selected European countries. Journal of Ethnic and Migration Studies, 48(10), 2347-2368 (2022).
[4] Griga, D., Hadjar, A.: Migrant Background and Higher Education Participation in Europe: The Effect of the Educational Systems, European Sociological Review, 30-3, 275--286 (2014).
[5] Goldstein, H., Healy, M. J.: The graphical presentation of a collection of means. Journal of the Royal Statistical Society: Series A (Statistics in Society), 158(1), 175-177 (1995).
[6] Jackson, M., Jonsson, O. J., & Rudolphi, F.: Ethnic Inequality in Choice-driven Education Systems. A Longitudinal Study of Performance and Choice in England and Sweden, Sociology of Education, 85(2), 158-178 (2012).
[7] Kao, G., Tienda, M.: Optimism and achievement: the educational performance of immigrant youth, Social Science Quarterly 76, 1: 1-19 (1995).
[8] Zapfe, L., & Gross, C. (2021). How do characteristics of educational systems shape educational inequalities? Results from a systematic review. International Journal of Educational Research, 109, 101837.

DRAFT

# Growth models for the Progress Test in Italian dentistry degree programs

Giulio Biscardi[a], Leonardo Grilli[a], Carla Rampichini[a], Laura Antonucci[b], and Corrado Crocetta[c]

[a]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy; giulio.biscardi@unifi.it, leonardo.grilli@unifi.it carla.rampichini@unifi.it

[b]Department of Clinical and Experimental Medicine, University of Foggia, Foggia, Italy ; laura.antonucci@unifg.it

[c]Department of Political Science, University of Bari Albo Moro, Bari, Italy corrado.crocetta@uniba.it

DRAFT

**Abstract**

In 2017 and 2018, for the first time in Italy, the students enrolled in Dental Schools performed a Progress test. The data available for these two years allows us to analyse the variability of students' knowledge during the course years to check the occurrence of a peak of ability. Furthermore, under certain conditions of the teaching provided by the universities, we can evaluate the coherence of the teaching results with the core curriculum. We propose an analysis of such data based on a mixed effects growth curve binomial model for the proportion of correct answers, considering fixed effects for the disciplines and random effects for the universities. We represent the trajectory of the Progress Test for each university's different fields using a polynomial time function. Using the Empirical Bayes prediction of level 2 residuals, the approach allows us to compare the trajectories of Italian universities, highlighting a substantial heterogeneity in the starting levels and growth rates.

***Keywords:*** Binomial Growth Curves, Empirical Bayes prediction, Mixed effect model, Progress Test

## 1. Introduction

In 2017 a Progress Test was administered for the first time to students of 31 out of 36 Italian Universities with a degree program in Dentistry.

The tests were administered throughout Italy on the same day and consisted of 300 multiple-choice questions divided into different subject areas. This instrument gives an account of the level of knowledge achieved by students in the disciplines covered during the degree program. All students received the same test regardless of the enrollment year (from first to sixth). This feature allows us to estimate the learning curve across the 6 year of course, even in absence of longitudinal data.

The Progress Test was repeated in subsequent years. We have access to data from editions 2017 and 2018, thus we can repeat the analysis to evaluate the persistence of the findings.

We specify a growth model [3] where the response variable is the proportion of correct answers to each test item at the university level. The data have a hierarchical structure with items nested into universities, thus the growth model is multilevel with random effects for the universities. The items are also nested into disciplines, for which we use fixed effects.

The growth of knowledge over time is typically nonlinear, increasing until it reaches a peak, after which it remains constant or decreases [4]. The literature on Progress Tests reveal that base and clinical disciplines show different patterns [1], hence we will allow for distinct parameters for the two types of disciplines.

The same data have been considered in [1], who give details on the test and show the results from a linear growth model. In this contribution, we adopt a similar approach, except that we overcome the limitations of the linear model by a binomial-logit specification of the proportion of correct answers.

## 2. Data structure

The available dataset on the Progress Test has 11,160 records. There are 31 universities with questions covering 30 disciplines. A key feature is that all students have to answer the same questions regardless of the enrollment year (first to sixth). Therefore, there are $31 \times 30 \times 6 = 5,580$ records for each edition (2017 and 2018). We dropped 450 records with missing test results, thus the number of records per university ranges from 180 to 360. In addition, we removed universities in which the number of test takers is low [2] (less than 50% in case of more than 150 expected participants, or less than 60% otherwise). As a result, the dataset for the analysis has 8,670 records, with 24 universities in 2017 and 25 in 2018.

The percentage of correct answers tends to increase from the first to the sixth year of enrollment, showing a growth like in Progress Tests administered in the faculties of Medicine and Surgery [6; 7].

## 3. A growth model for the learning curve

Growth models are widely used to analyse the trajectories of individuals over time [3]. In our case, they can be used to model the learning curve across the six years of enrollment. As we do not have access to individual data, we specify a model for the proportions of correct responses for the universities, estimating the pattern at the university level. The heterogeneity among universities is accounted by random effects.

The growth model for the Progress Test is specified as follows:

$$y_{djt} \sim Bin\left(p_{djt}, n_{djt}\right) \tag{1}$$

$$logit\left[p_{djt}\right] = (\alpha_{0d} + u_{0j}) + (\alpha_{1d} + u_{1j}) \times t + (\gamma_1 + \gamma_2 \times c_d) \times t^2 \tag{2}$$

with

- $d = 1 \ldots 30$ disciplines
- $j = 1 \ldots 25$ universities
- $t = 0 \ldots 5$ enrolment years (with $t = 0$ for the first year etc.)
- $c_d$ dummy variable for the type of discipline (1: clinic, 0: base)
- $\gamma_1, \gamma_2$ coefficients of the quadratic term
- $\alpha_{0d}$, $\alpha_{1d}$ discipline fixed effects
- $[u_{0j}, u_{1j}]^T \sim N\left(\mathbf{0}, \mathbf{\Sigma}\right)$ random effects for the universities

Note that the intercepts and linear terms of the curves vary across disciplines due to the fixed effects $\alpha_{0d}$, $\alpha_{1d}$ and vary across universities due to the random effects $u_{0j}$, $u_{1j}$. On the other hand, the quadratic term only vary across the type of discipline (base vs clinic).

The marginal likelihood of the model (1)-(2) involves an integral with respect to the random effects that is not in closed form. For this reason, we use Laplace's method to approximate the integrand function with a Normal distribution centered on the value of the random effect that maximizes the integrand. We choose to use the Laplace approximation, instead of Gauss-Hermite quadrature, because it is computationally faster while yielding similar estimates. We estimated the model using the `meqrlogit` command of `Stata`.

DRAFT

Once the parameters of model (1)-(2) have been estimated, we assign values to the random effects for each university by means of the Empirical Bayes method, which provides the Best Linear Unbiased Prediction (BLUP). The shrinkage property of the Empirical Bayes residuals pulls towards zero the values for the universities with a low number of respondents.

# 4. Results

The model's fixed and random effects in both years will be presented and discussed in this section. We will pay attention to the two types of shapes obtained from the curves of the proportions of correct answers. Since we separately estimated a model for 2017 and one for 2018, the results obtained in both years will be compared and discussed.

## *4.1 Fixed effects*

In both editions of the Progress Test, the base disciplines have starting values higher than clinic disciplines, while the fixed effects on growth for these disciplines are among the lowest.

In contrast, the disciplines with a lower starting value are clinical. Looking at the fixed effects on growth, we see that the values of the clinical disciplines are higher than those of the basic disciplines.

In both editions of the test, the coefficient of the quadratic term is significantly different between base and clinic disciplines, leading to a different shape of trajectories. Most of the trajectories for the base disciplines are a parabola with downward concavity, which means that students' knowledge in these subjects increases until they reach a peak beyond which knowledge decreases. This pattern is expected since base disciplines are treated only early in the degree program. On the other hand, the trajectories of the clinical disciplines have a shape much closer to a straight line, pointing out that knowledge is still increasing without reaching a peak. An example of this difference can be seen in Figure 1, in which the trajectories of the predicted probabilities of correct answers for three basic disciplines (behavioral sciences, chemistry and physics) and three clinical disciplines (endodontics, plastic surgery and dentistry) are shown for an average university, i.e., with random effects equal to zero.

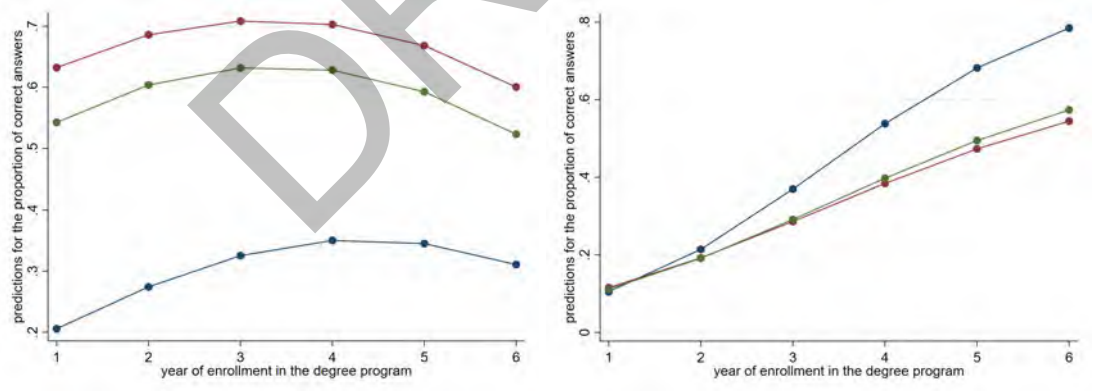



Figure 1: Predicted probabilities of correct answers for three basic disciplines (left panel) and three clinical disciplines (right panel) for an average university, 2017

## *4.2 Random effects*

The random intercept and the random slope have statistically significant variances. Their correlation is negative, namely universities whose students have a higher knowledge at the beginning tend to grow at a lower rate. The initial gap among universities is due to the self-selection of students, who choose the universities with major reputation. In most cases, the initial gap in the knowledge is closed within the end of the degree program due to a faster learning process.

DRAFT

The Empirical Bayes predictions of the random effects with confidence intervals are arranged in the caterpillar plot [5] shown in Figure 2. The intervals are wider for universities with fewer respondents. In most cases the intervals do not intersect the zero, so that the corresponding random effects are significantly different from zero.

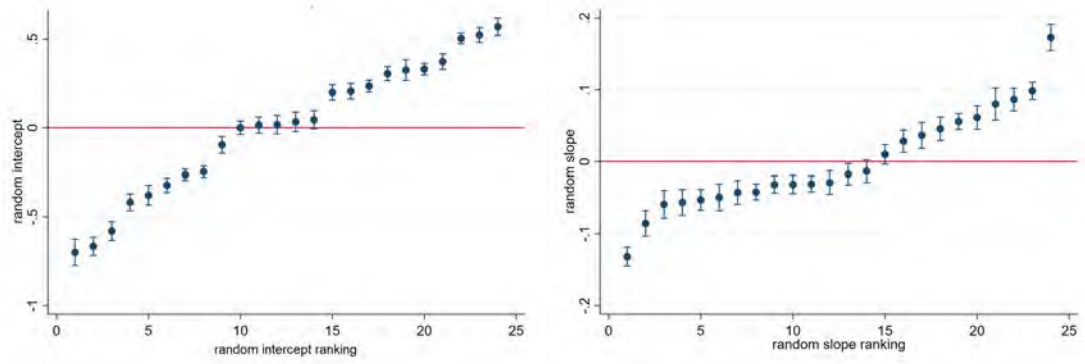



Figure 2: Caterpillar plot of Empirical Bayes predictions of random intercepts (left panel) and random slopes (rigth panel). All universities, 2017

DRAFT

## 5. Final remarks

The analysis has shown that the disciplines of the Progress Test have quite different trajectories, with the performance in the base disciplines reaching a plateau or even decreasing in the last years of the degree program. The Empirical Bayes predictions allowed us to build rankings of the universities with confidence intervals accounting for uncertainty. The results are suitable to support policy interventions.

## References

[1] Antonucci, L., Biscardi, G., Firza, N., Grilli, L.: Un modello ad effetti misti con dati longitudinali per il progress test dei corsi di laurea italiani in odontoiatria. In: Ernesto Toma. Metodi e Analisi Statistiche, pp. 219-234 Università degli Studi di Bari Aldo Moro. (2022)

[2] Crocetta, C., Brindisi, M., Lo Muzio, L. : Analisi dei risultati del progress test 2017 dei Corsi di Laurea in Odontoiatria e Protesi Dentaria. In: Journal of Italian Medical Education 2.78, pp. 3487-3493 (2018)

[3] Hedeker, D.: An introduction to growth modeling. In: The Sage handbook of quantitative methodology for the social sciences, pp. 215-234 (2004)

[4] McNeish, D., Dumas, D., Torre, D., Rice, N.: Modelling time to maximum competency in medical student progress tests. In: Journal of the Royal Statistical Society. Series A: Statistics in Society. (2022)

[5] Rabe-Hesketh, S., Skrondal, A.: Multilevel and Longitudinal Modeling Using Stata. Forth edition. College Station, TX: Stata Press (2022)

[6] Tenore, A., Basili, S., Lenzi, A., Marangon, M., Proietti, M.: Il Progress Test 2011. In: Journal of Italian Medical Education 56, pp. 2487-2509 (2012)

[7] Tenore, A., Basili, S., Proietti, M.: Il Progress Test 2012. In: Journal of Italian Medical Education 60, pp. 2699-2704 (2013)

# The COVID-19 pandemic and academic E-learning: Italian students and instructors' perceptions

Francesco Santelli[a], Teresa Gentile[b], Davide Bizjak[c] and Lorenzo Fattori[d]

[a] University of Trieste; fsantelli@units.it
[b] University of Salerno; [c] University of Naples Federico II
[d] Aosta Valley University

**Abstract**

The term e-learning refers to a learning method involving new multimedia and internet technologies capable of totally replacing face to face meetings in academic contexts. With the advent of the COVID-19 pandemic, the closure of universities and higher education institutes forced a sudden and unexpected transition from face-to-face teaching to distance learning to mitigate the spread of SARS-CoV-2. This shift to online teaching and learning is a game changer for higher education.

The goal of this work is to identify the factors that impacted the adoption of e-learning systems within Italian universities during the period of the COVID-19 pandemic situation. Specifically, through a quantitative empirical survey and a statistical analysis with EFA method, it was possible to understand how the perceived usefulness and ease of use of e-learning platforms have influenced the online learning and teaching experience of Italian students and professors.

**Keywords:** e-learning, covid pandemic,exploratory factor analysis, exploratory graph analysis, digital skills

DRAFT

## 1. Introduction

The term e-learning system refers to a learning method that involves the use of new multimedia and internet technologies capable of totally replacing face to face meetings in academic contexts (Azeiteiro et al., 2015; European Commission, 2001; Guri-Rosenblit, 2005; Holmes & Gardner, 2006; Khan, 2005; Qwaider, 2011). In this sense, the use of specific technologies such as tutorials, audio, video, web pages, or games or other, provided synchronously and / or videotaped (asynchronous) also through the support of mobile devices could facilitate and improve access to e-learning systems able to completely substitute the didactic activities carried out in university classrooms (Rice & Gregor, 2016).

The forced and unprecedented shift to online teaching is seen as a game changer for education change (notably higher education) and fostering innovation (Flores et al., 2022) and has generated a different approach in students and teachers in carrying out teaching activity.

## 2. Design of the survey

The questionnaire was administered online. The questionnaire was disclosed in mid-March 2020, during the first lockdown period resulting from the health emergency for COVID-19, when Italy had issued a ministerial decree on guidelines and tools for implementing e-learning in universities. For the formulation of the content of the various items of the questionnaire, reference was made to the existing literature on the adoption of e-learning systems in universities (Flores et al., 2021). The questionnaire was disseminated through certain social media and online channels with some specific groups of possible participants (Kayam & Hirsch, 2012). Completion time was estimated at just under 10 minutes. Furthermore, in the initial part, based on current privacy legislation, it was indicated that the answers would be processed and managed anonymously and that participation in the questionnaire was voluntary (Kayam, & Hirsch, 2012).

The survey was aimed at professors (professors, researchers, and adjunct professors) and students (enrolled in a first- and second-level degree course or a research doctorate) from Italian universities.

A total of 473 respondents were involved in the survey. The questionnaire is divided into the 9 sections below to allow the acquisition of different information from generic to the more specific

## 3. Methodology and data preparation

According to the research questions of the present work, statistical methodology follows the framework of the Exploratory Factor Analysis (EFA) (Costello and Osborne, 2005; Fabrigar and Wegener, 2011). It is a statistical method used to reveal the relationships among a set of variables, i.e., to highlight the presence of a consistent structure of factors. EFA is a technique belonging to the large family of factor analysis, whose predominant aim is to identify the underlying and not necessarily explicit associations between measured variables. It is commonly conceived as a technique that identifies a set of latent constructs that are not observed, underlying a battery of measured variables that is instead observed.

When dealing with unknown factors and manifest variables with no a priori fixed constructs, this approach can assess if a few emergent latent factors are present in the data. Thus, it is handy when dealing with data about a (relative) recent topic yet to explore.

In this case, given that the sudden epidemic emergency led to a quick switch in learning strategies and the whole situation can be seen as the first time that the modern academic world has to face this kind of challenge, an exploratory approach to figure out recurrent patterns in questionnaire items can be considered a safe first multivariate picture of the COVID effect in the Italian context.

EFA is based on the principle of the common factor model (Velicer and Jackson, 1990). Within the framework of this model, a combination of common factors, unique factors, and measurement errors leads to the definition of the manifest variables (Norris and Lecavalier, 2010). Each unique factor impacts only one manifest variable and does not influence the correlation among manifest variables.

Factor loadings are measures of the weight of a given unique factor on the manifest variable.

In this case, the underlying assumption is that each item can be related only to one factor, while with the adequate rotation (not orthogonal), axes can correlate each other's. Still, the overall correlation/association structure drives the final result about the constructs.

In our case, it is interesting that it is possible to split the data in two samples: "teachers" and "students." We expect most relationships will not diverge too much from one data set to another, while a few minor differences will highlight peculiar behaviors. Overall, learning experiences can not be considered as a whole, and the two groups' expectations and challenges can differ.

## 3. Data description

From these data, we will focus on the questions related to the respondent's personal characteristics and items strictly associated to the e-learning process during the pandemic time, such as the type of platform, role in the university, type of degree/Ph.D., items related to behaviors and perceptions about e-learning and specific platforms, and overall effects of the pandemic context and online-distance learning.

Among the 473 respondents, students are prevalent, 351 of them. Among them, 235, the majority, is in the course of obtaining a Bachelor Degree. Among the students also, there are also 23 who are in the course of obtaining a PhD: this is because, for Italian law, PhDs are students, and also because they are more involved in learning than in teaching and researching.

The majority of respondents among professors is composed by full or associate professors (67 respondents), but there are also 55 researchers of three different types: Post-Doc, Scolarship Researchers and Researchers.

Nearly everyone among the respondents has been involved in e-learning: only two respondents still needed to have e-learning experience before responding to the survey.

To test if items are suited for Factorial Analysis, we propose results of Kaiser-Meyer-Olkin factor adequacy. Value of this KMO between 0.8 and 1 indicates a very high adequacy for factor analysis. For our data such measures is equal to 0.95.

Results indicate that data are suited for factor analysis. Similar results (0.94; 0.88) are obtained for the students data and professors data, respectively.

DRAFT

In the following, given the nature of the data, we will use Polychoric correlation matrix suited for ordinal data, such the Likert scale that we are using for the items involved in the analysis. The two different Polychoric correlation matrix (students and professors data) show a similar pattern, even with slight differences.

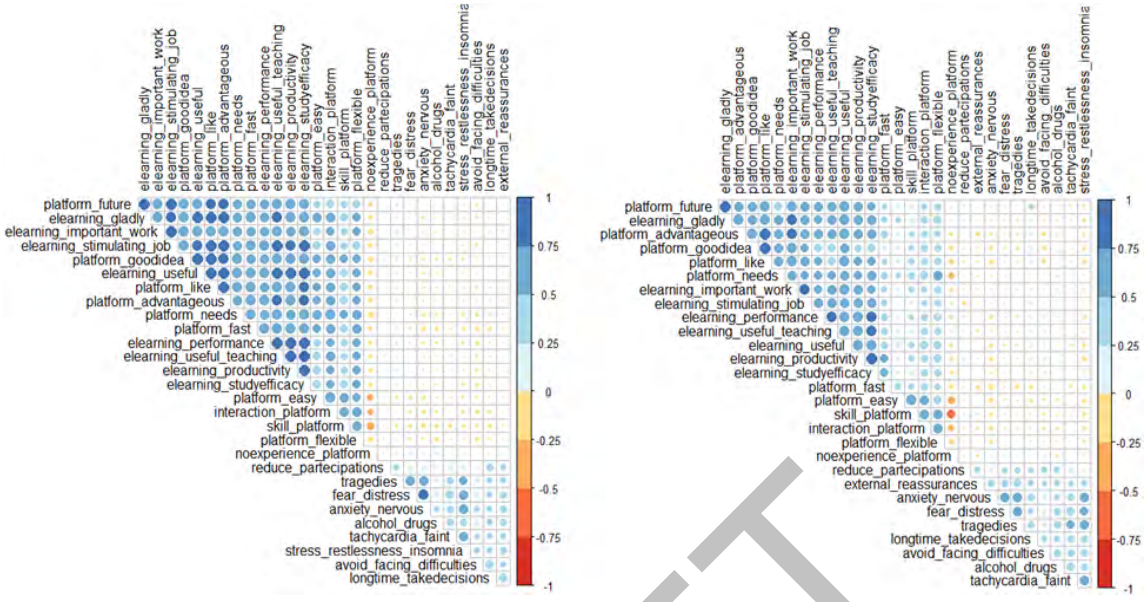
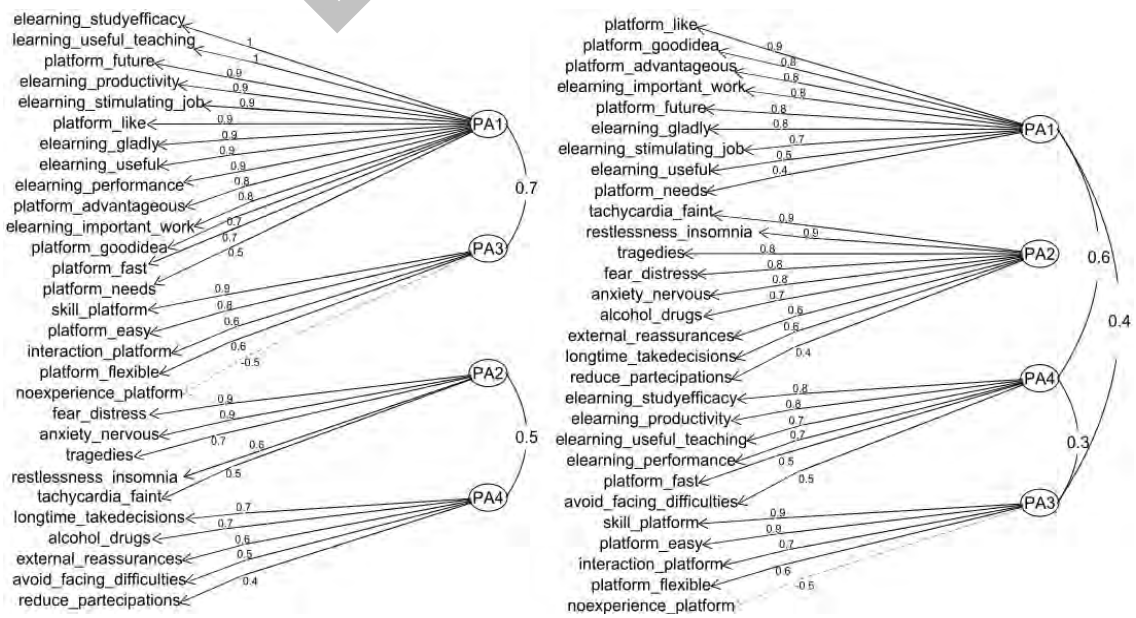



Figure 1: polychoric correlation matrix of the 29 items. Left: Students Data. Right: Professors Data.

## 4. Main EFA findings

Parallel Analysis (PA) (Braeken, 2017) and Exploratory Graph Analysis (EGA) (Christensen, 2020) estimate the number of dimensions in factor analysis. Results differ slightly between the two approaches (students: 3 with PA and 4 with EGA; professors, 4 with PA and 6 with EGA but with two isolated items). Thus, we propose a compromise solution with 4 factors for each data-set. As a control over such identified factors, Cronbach alpha is computed. Given the exploratory approach of the analysis, we try to understand if the observed factors can be reliable and to what extent. Most results are promising, with factor 3 in students data (0.58) and factor 3 in professors data (0.57) being the most problematic.




DRAFT

Figure 2: EFA factors identification. Both case 4 factors as best solution. Left: Students Data. Right: Professors Data.

Table 1: GLM regressions using covariates to explain individual scores on the different factors.

| | | | | | |
|---|---|---|---|---|---|
| ***Students*** | **Type of Degree** | PA1 | PA3 | PA2 | PA4 |
| | 5 Years - PhD | *Ref.* | | | |
| | Bachelor | 0.719*** | −0.481* | *No effect* | |
| | Master | 0.830** | *No effect* | | |
| ***Students*** | **Geographical Area** | PA1 | PA3 | PA2 | PA4 |
| | Center | *Ref.* | | | |
| | North | *No effect* | | | *No effect* |
| | South | | | | −0.626* |
| -- | -- | -- | -- | -- | -- |
| ***Professors*** | **N of Courses** | PA1 | PA3 | PA2 | PA4 |
| | 0-1 | *Ref.* | | | |
| | 2-3 | 0.480+ | *No effect* | *No effect* | *No effect* |
| | 4+ | *No effect* | −0.785* | | 0.692+ |
| ***Professors*** | **Field** | PA1 | PA3 | PA2 | PA4 |
| | STEM | *Ref.* | | | |
| | Med-Vet-Bio | 0.958+ | −0.762+ | | |
| | Humanities | 0.735+ | −0.785* | *No effect* | |
| | Economics-Statistics | 0.751+ | *No effect* | | |
| ***Professors*** | **Method online** | PA1 | PA3 | PA2 | PA4 |
| | Still don't know | *Ref.* | | | |
| | Asynchronous | 1.735** | *No effect* | | |
| | Syncronous | *No effect* | | | |
| ***Professors*** | **Platform type** | PA1 | PA3 | PA2 | PA4 |
| | University + External | *Ref.* | | | |
| | Only University | *No effect* | *No effect* | | |
| | Only External | −0.684* | | | |
| ***Professors*** | **Geographical Area** | *No effect* | | | |
| ***Professors*** | **Role** | *No effect* | | | |
| ***Professors*** | **Frontal lesson** | *No effect* | | | |

DRAFT

## 5. Discussion and Conclusion

**Students:** The first factor (PA1) identifies mainly an underlying dimension tied to the efficacy and usefulness of e-learning in general, while the second (PA3) locates the relation between the individual and the e-learning platform. For factor PA1, the most influential items are "elearning_studyefficacy" and "learning_useful_teaching", both with a loading close to 1. For factor PA3, the most influential item is "skill_platform" with a loading of 0.9, representing the answers to "I believe I can soon become proficient in using the e-learning platform". The factor PA2 seems to individuate a dimension of anxiety and fear, while the last factor PA4 relates to the items pertaining to difficulties of interaction, participation, and the necessity of reassurances. For factor PA2, the most influential items are "fear_distress" and "anxiety_nervous", both with a loading of 0.9. For factor PA4, the most influential items are "longtime_takedecisions" and "alcohol_drugs", both with a loading of 0.7.

Professors: The first (PA1), third (PA4), and fourth (PA3) all correlate among them and seem to identify latent dimensions pertaining to the use of e-learning and platforms. More specifically, for factor PA1, the most influential item is "platform_like", with a loading of 0.9. For factor PA4, the most influential items are "elearning_studyefficacy" and "elearning_productivity", both with a loading of 0.8. For factor PA3, the most influential items are "skill_platform"

and "platform_easy", both with a loading of 0.9". The other factor (PA2), however, gathers together all the items related to fear and anxiety. The most influential items for this factor are "tachycardia_faint" and "restlessnees_insomnia", both with a loading of 0.9. For professors, the ideal number of courses for the best teaching effort and understanding e-learning and platform is between two and three. The anxiety level tends to increase with the number of courses assigned to an individual professor: in particular, with more than four courses to dispense, the professor's anxiety level tends to raise very significantly, and this seems the only occasion in which an influence of e-learning on anxiety level is clearly traceable.

Moreover, STEM professors seem to be less optimistic about e-learning, while medical sciences and Humanities professors have more difficulties with platform use. Professors who had their lessons in an asynchronous way (lessons were first recorded and subsequently delivered to students) have higher e-learning appreciation scores. Finally, professors who only used external platforms and not platforms provided by their universities have the worst idea of e-learning in general.

In conclusion, the results seem consistent with the Technology Acceptance Model (TAM): from the analysis, it is possible to note that the degree of satisfaction with e-learning is higher where users indicate a greater perception of reliability and ease of use of the platform. This happens for both students and professors. Nevertheless, the most evident difference between students and professors, when it comes to factors, is the difference in anxiety. This is not surprising, given the higher age and responsibilities of professors than that of the students: short-term and long-term anxiety collapse to just one dimension, while fear may be less about one's career development and more about the health of oneself and its loved ones.

Finally, the general lack of influence of e-learning on anxiety levels should be regarded as the clearest result of this study. As shown, this lack of influence emerges from both datasets analyzed, for students and professors.

DRAFT

## References


[1] Almaiah, M. A., Al-Khasawneh, A., & Althunibat, A. (2020). Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic. Education and Information Technologies, 1.

[2] Azeiteiro U. M., Bacelar-Nicolau P., Caetano F. J., & Caeiro S. (2015), Education for sustainable development through e-learning in higher education: experiences from Portugal. Journal of Cleaner Production, 106, 308-319.

[3] Braeken, J., & van Assen, M. A. (2017). An empirical Kaiser criterion. Psychological Methods, 22, 450 – 466. http://dx.doi.org/10.1037/ met0000074

[4] Christensen, A. P., & Golino, H. (2021). Estimating the stability of the number of factors via Bootstrap Exploratory Graph Analysis: A tutorial. Psych, 3(3), 479-500.

[5] Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical assessment, research, and evaluation, 10(1), 7.

[6] European Commission (2001), Communication from the Commission to the council and the European parliament The eLearning Action Plan Designing tomorrow's education, Bruxelles, COM, 172.

[7] Fabrigar, L. R., & Wegener, D. T. (2011). Exploratory factor analysis. Oxford University Press.

[8] Favale, T., Soro, F., Trevisan, M., Drago, I., & Mellia, M. (2020). Campus traffic and e-Learning during COVID-19 pandemic. Computer Networks, 107290.

[9] Flores, M. A., Barros, A., Simão, A. M. V., Pereira, D., Flores, P., Fernandes, E., ... & Ferreira, P. C. (2022). Portuguese higher education students' adaptation to online teaching and learning in times of the COVID-19 pandemic: personal and contextual factors. Higher Education, 83(6), 1389-1408.

[10] Guri-Rosenblit, S. (2005). 'Distance education'and 'e-learning': Not the same thing, Higher education, 49(4), 467-493.

[11] Holmes, B., Gardner, J. (2006), E-learning: Concepts and practice. Sage.

[12] Khan, B. H. (2005), Managing e-learning: Design, delivery, implementation, and evaluation, IGI Global.

[13] Qwaider, W. Q. (2011). Integrated of knowledge management and E-learning system, International Journal of Hybrid Information Technology. 4(4), 59-70.

# Working Students and job market outcomes: Insights from the University of Florence

Gabriele Lombardi[a], Valentina Tocchioni[a], and Alessandra Petrucci[a]

[a]Department of Statistics, Computer Science, Applications "Giuseppe Parenti" - DiSIA, University of Florence; gabriele.lombardi@unifi.it, valentina.tocchioni@unifi.it, alessandra.petrucci@unifi.it

**Abstract**

Does having been a working student during the higher education studies matter for future job market outcomes? This article addresses this question by analysing administrative data, matching individual-level data on the career of students at the University of Florence, Italy, with records on employment contracts for the same students from the Italian Ministry of Labor. A survival analysis for ordered events is employed in order to estimate the risk of stipulating hierarchical-ranked types of employment contracts. Coherently with expectations, it turns out that former working students graduates exhibit a positive risk of being employed in a shorter time. Moreover, students from lyceums are more likely to obtain a higher-skilled contract.

***Keywords:*** Secondary Education, Job Market Outcomes, Working students, Higher education.

DRAFT

## 1. Introduction

The present article tries to investigate the association between the fact of having been or not a working student during the tertiary education career and the labor markets outcomes for those graduates from the University of Florence. In particular, we explore the idea that the decision of being a working student may condition at a very early stage the future chances of students' social mobility. Indeed, Italy could exhibit a standstill of compulsory education in reducing inequalities (7), and even if it is included among the countries that were able to reduce inequalities for the transition to the upper secondary education, it remained behind concerning the stratification at the tertiary education level (2). Even though universal education should have helped in overcoming the classism in educational attainment, and in fostering social mobility, Italy is still characterized by mostly upper-class students that enrol in secondary school's academic tracks, such as lyceums. On the other side, students from the working class are more likely to choose technical and vocational schools; their families' investment on education turns out to be burdensome, and they need strongly to "bet" on their childrens' ability in order to allow them enrolling an academic track (14; 5). In a nutshell, Italy exhibits social origins as the main driver for educational choices after compulsory schooling, fostering a school track implicitly based on classes, which tends to mantain the status quo (8).

In this framework, focusing on the decision of working during the tertiary education path could be a useful aspect for exploiting the issue of social stratification between education and employment. Indeed, this choice should be taken by those students who need to support themselves or have at least some monetary needs to satisfy. Nonetheless, economic needs are not the only reason for getting a job in the meanwhile of a degree course; also, another motivation can be the will of anticipating the entry in the labor market, so to signal a greater experience gathered in advance with regard to not working colleagues

(12). Thus, it is still uncertain if working during studies reflects more the presence of economic difficulties, the will of signalling resourcefulness to the future recruiters, or just that tertiary education is seen as an accessory activity, as in the "parking lot" hypothesis (4).

# 2. Data and Model

## 2.1 Data

The following analysis is based on data obtained thanks to an agreement between the Italian Ministry of Labor and the University of Florence. Regarding the available data, administrative information about the student population of the University of Florence are merged with the compulsory notifications which have to be provided by every employer (public or private) who stipulates, extends, transforms or terminates any employment relationship. In particular, these data refer to 28073 graduates at the University of Florence between 2008 and 2013, who stipulated an official job contract between the 1st of January 2008 and the 30th of September 2015. For each graduate, careers were built starting from the first compulsory notification available. In this regard, it is important to notice the absence of an information about the exact graduation day, only the year being available. Thus, evaluating if certain jobs were obtained before or after the achievement of graduation is impossible: in those cases, a job is considered obtained after graduation if it is stipulated in the same calendar year of students' graduation. Coherently, the employment before graduation identifies working students. Indeed, recruiters could have some priors for evaluating positively or negatively this particular category of students.

Also, information about the Secondary Education is available: namely, the year of diploma, the final mark obtained, and the different typologies of high schools, coded according to four categories: *Lyceum*; Technical Schools; Vocational Schools; other schools. A proxy for the possibility that a student had to repeat one or more years of school is obtained through a dummy that indicates if the student graduates when s/he was older than 19 years old, which is the prescribed age at which a high school degree in Italy is supposed to be obtained.

DRAFT

Table 1: Descriptive Statistics: Means (for continuous variables) and proportions (for discrete) for the entire sample and sampled by job type.

| | At least a Low Job | At least a Medium Job | At least a High Job | Full Sample |
|---|---|---|---|---|
| **Secondary Ed.** | | | | |
| Final Grade | 82.55 | 81.98 | 84.45 | 82.56 |
| Fail | 0.14 | 0.13 | 0.11 | 0.13 |
| *Type of Secondary School:* | | | | |
| Lyceums | 0.66 | 0.65 | 0.70 | 0.66 |
| Technical | 0.25 | 0.28 | 0.20 | 0.24 |
| Vocational | 0.02 | 0.02 | 0.02 | 0.02 |
| Other | 0.07 | 0.06 | 0.08 | 0.07 |
| **Job Career** | | | | |
| Working Student | 0.32 | 0.25 | 0.26 | 0.33 |
| Prev.Job Low | 0.50 | 0.33 | 0.27 | 0.55 |
| Prev.Job Medium | 0.17 | 0.31 | 0.15 | 0.30 |
| Prev.Job High | 0.10 | 0.14 | 0.37 | 0.30 |
| No. Observations | 15066 | 9608 | 9962 | 28073 |

Information about university performance (e.g. average mark, time-to-degree) are not used because it represents an acknowledged source of endogeneity in the analysis of working students. Indeed, the decision of working and studying at the same time affects both the performance and the speed in achieving the degree.

Finally, job contract levels are classified thanks to a framework designed by the Italian National Institute of Statistics (ISTAT), where contracts are sorted on a scale 1-8 according to a hierarchical principle. Consequently, the outcomes are classified into three categories: *Low Job*, which includes classes from 4 to 8, namely elementary workers, craftsmen, specialized workers, farmers, specialized workers in tertiary sector, executive jobs in tertiary sector; *Medium Job*, which includes class 3, that are technical professions; *High-Job*, which includes classes 2 and 1, namely intellectual and scientific high-skill professions, and legislators, businessmen, managers and entrepreneurs.

Table 1 reports some descriptive statistics for the main covariates included in the analysis by type of contract. Among the main evidences, as two thirds of the whole sample are composed by lyceum students, one third of subjects are former working students. Graduates who have obtained at least a low job exhibit a higher share of students who failed a year during secondary education (14%) and of former working students (32%). 50% of them have stipulated a low job after another low job. On the other side, graduates who have obtained one or more medium contracts show the lower average final grade (81.98) and the higher share of former students coming from technical schools. Unsurprisingly, workers in high contracts include the highest average final grade (84.45), the lower share of students who failed one or more years (11%) and the highest share of lyceum students (70%).

## 2.2 Model

As in the data all the information obligations were available up to the $30^{th}$ of September, 2015, for each individual, then the information had to be conformed to survival structure. Consequently, starting from the first employment contract for each individual and focusing on all contracts which lasted at least 30 days, we categorized each of them as as a Low, Medium or High Job. Moreover, time spells were created for the periods in which each worker had not any contract in effect, ar a contract shorter than 30 days, since we are considering a monthly time unit. Those spells start from the end date of their own previous spell, and they finish on the start date of the following one. If the last contract available for a certain person results closing before $30^{th}$ of September, 2015, then a spell is created up to that date[1]. Thus, we are able to consider each working spell considering its job qualification.

Consequently, we build a *Cox proportional hazard model for subdistribution* (9; 15). Given $i = 1, ..., k$, the cause-specific hazard ratio is (6):

$$h_i(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta t, \text{failure from cause } i | T \geq t)}{\Delta t}$$

where, $t$ is a certain point in time, $T$ is the time to first failure, and the overall hazard rate, which is the probability that the failure occurs into the prescribed time interval, is $h(t) = \sum_i h_i(t)$, and the probability a failure occurs because of $i$ is $h_i(t)/h(t)$.

Thus, as we can observe for each individual multiple failures for competing risks, it is necessary to control for the possibility of tied events, i.e. events with the exactly same survival time. In order to deal with this problem, the method proposed by Efron (11) is preferred since, even if fit statistics are worse than those of other methods, efficiency and parameters estimates are basically equal with others and the bias is even lower in presence of a large number of ties (3).

As already stated, in the available data events can be recurrent, and the order in which they show up for each individual matters. Accordingly, starting from the partial likelihood function by (10), the estimation is corrected for multiple failure events (1), and stratified according with the order of events (15). For the $i$th failure type ($i$ = Low, Medium, High) the *hazard function* will assume the form:

[1]Unfortunately, after this reconstruction several contracts appeared overlapping their time spells for the same individual. Maybe arbitrarily, a choice was taken of maintaining case by case only the contract with the longest time length.

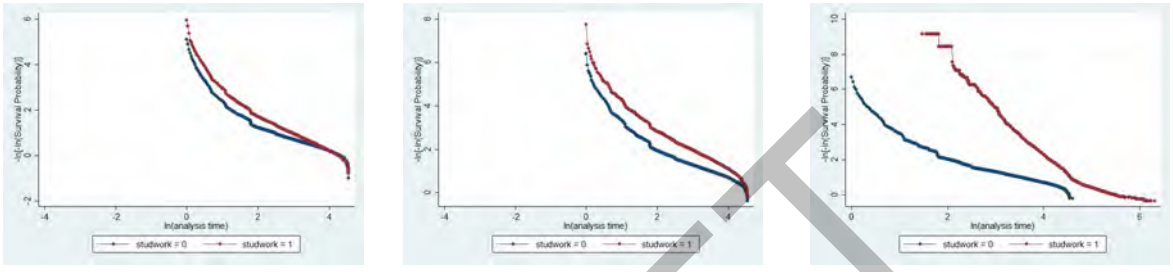
DRAFT

$$h_i(t; \beta_i, X(t)) = h_{i0}(t) exp[X(t)^T \beta_i],$$

where $h_{i0}$ is the latent baseline hazard function, $X(t)$ is the matrix of the covariates until time $t$, and $\beta_i$ is the vector of the coefficients.

## 3. Results and Discussion

Figure 1 allow us to check graphically the proportionality of hazards for having been or not a working student, which do not seem to present a picture so much far from proportionality, with the only exception of subfigure 1(c) for high jobs.



(a) Outcome: Low job (b) Outcome: Medium job (c) Outcome: High job

Figure 1: Proportional Hazard Plot for being a Working Student: (a) Low Job; (b) Medium Job; (c) High Job.

DRAFT

Hence, looking at Table 2 some straightforward conclusions can be taken. First of all, the fact of having been a working student is positively associated with every kind of qualification, even if the association is very low and weak with regard to medium jobs. Probably the two stronger relationship with low and high qualifications reflect two different frameworks: on the former side, those graduates experiencing the harshest economic needs, so having worked during tertiary education studies, remain stocked in lower level jobs. On the latter side, successful working students are positively rewarded by recruiters, in the long run. Nonetheless, this emerges as a puzzling result, deserving deeper exploration of the underlying mechanisms as a further development of this study.

Moreover, we can notice how the risk of obtaining a low job is negatively associated with the final grade and positively with fails during secondary education. Moreover, students from lyceums have a lower risk of ending up in such a job level than every other high school type. The risk of obtaining a low job is also negatively associated with age, probably because this kind of contracts are more likely to be gained at the very beginning of individuals' career. As it can be seen, the classification of the previous job is never significant across the three models, with a single exception. Indeed, having had a low job is positively associated with the risk of stipulating another low job, as these contracts may cause a sort of career stagnation.

Regarding the risk of stipulating a medium job, it is negatively associated with the final grade, but it does not seem to be related with having repeated one or more years during secondary education. Graduates coming from technical schools exhibit higher risk than those from lyceums. Also in this case, the risk is negatively associated with age.

A quite different picture emerges concerning high jobs. In this case, we find a higher risk for those with a high final mark and those who did not fail during secondary education. Students from lyceum are positively associated with jobs with a high qualification more than any others. Finally, this is also the only case in which age at event is positively significant, probably because it is necessary a longer time and experience for ending up in such contracts.

Nonetheless, secondary education emerges as a strong source of social stratification. Indeed, it seems sufficient to be a well-performing student coming from a lyceum to strongly increase the chances of obtaining a high job. But, as said before, these characteristics are still devoted to upper-classes graduates.

Concluding, this article has shown a preliminary analysis about the association between the fact of having been a working students and job market outcomes. Further developments should include other socio-demographic covariates that may play a role in shaping this association, add some interaction terms that may moderate the effect, and benefit of more recent graduates' cohorts.

Table 2: Conditional Risk Set Model for Ordered Failure Events, stratfied by occurence, and with tied events controlled by Efron Method.

| | Conditional Risk Set Model (Main) | | | | | |
|---|---|---|---|---|---|---|
| | Low | | Medium | | High | |
| | $\hat{\beta}$ | s.d. | $\hat{\beta}$ | s.d. | $\hat{\beta}$ | s.d. |
| **Secondary Ed.** | | | | | | |
| Final Grade | -0.004*** | (0.001) | -0.074*** | (0.024) | 0.009*** | (0.001) |
| Fail | 0.182*** | (0.025) | -0.050 | (0.031) | -0.159*** | (0.033) |
| *High School Type (Ref: Lyceums)* | | | | | | |
| Technical | 0.043** | (0.020) | 0.074*** | (0.024) | -0.286*** | (0.026) |
| Vocational | 0.167*** | (0.060) | -0.004 | (0.069) | -0.402*** | (0.078) |
| Other | 0.082** | (0.033) | -0.144*** | (0.045) | -0.190*** | (0.038) |
| **Job Career** | | | | | | |
| Working Student | 0.308*** | (0.037) | 0.058* | (0.034) | 0.205*** | (0.035) |
| *Previous Job (Ref: No Job)* | | | | | | |
| Low | 0.797** | (0.385) | -0.186 | (0.332) | -0.361 | (0.326) |
| Medium | -0.369 | (0.385) | -0.237 | (0.335) | -0.271 | (0.326) |
| High | -0.079 | (0.386) | -0.340 | (0.333) | 0.480 | (0.328) |
| Age at Event | -0.062*** | (0.003) | -0.027*** | (0.003) | 0.024*** | (0.002) |
| **Control Variables** | | | | | | |
| Faculties | Yes | | Yes | | Yes | |
| Year at Event | Yes | | Yes | | Yes | |
| No. of Subjects | 28073 | | 28073 | | 28073 | |
| No. of Failures | 15066 | | 9608 | | 9962 | |
| Time at Risk | 871173.5 | | 1059287.2 | | 1060461.2 | |
| Log pseudolike. | -121147.86 | | -76446.537 | | -79107.682 | |

DRAFT

## References


[1] Andersen, P. K., & Gill, R. D.: Cox's regression model for counting processes: a large sample study. The annals of statistics, 1100-1120. (1982)

[2] Blossfeld, P. N., Blossfeld, G. J., & Blossfeld, H. P.: Changes in educational inequality in cross-national perspective. In Handbook of the life course (pp. 223-247). Springer, Cham. (2016)

[3] Borucka, J.: Methods of handling tied events in the Cox proportional hazard model. Studia Oeconomica Posnaniensia, 2(2), 91-106. (2014)

[4] Bratti, M., Checchi, D., & De Blasio, G.: Does the expansion of higher education increase the equality of educational opportunities? Evidence from Italy. Labour, 22, 53-88. (2008)

[5] Checchi, D.: University education in Italy. International Journal of Manpower, Vol. 21 No. 3/4, 2000, pp. 177-205. (2000)
[6] Cleves, M., Gould, W., Gould, W. W., Gutierrez, R., & Marchenko, Y. : An introduction to survival analysis using Stata. Stata press. (2008)
[7] Cobalti, A.: Schooling inequalities in Italy: trends over time. European Sociological Review, 6(3), 199-214. (1990)
[8] Contini, D., & Triventi, M. : Between formal openness and stratification in secondary education: Implications for social inequalities in Italy. In Models of Secondary Education and Social Inequality. Edward Elgar Publishing. (2016)
[9] Cox, D. R. : Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202. (1972)
[10] Cox, D.R. : Partial likelihood. Biometrika, 62(2), 269-276. (2016)
[11] Efron, B. : The efficiency of Cox's likelihood function for censored data. Journal of the American statistical Association, 72(359), 557-565. (1977)
[12] Finocchietti, G. : Students and universities in Italy in an age of reform. European Journal of Education, 39(4), 459-469. (2004)
[13] Meggiolaro, S., Giraldo, A., & Clerici, R. : A multilevel competing risks model for analysis of university studentsâ careers in Italy. Studies in Higher Education, 42(7), 1259-1274. (2017)
[14] Panichella, N., & Triventi, M. : Social inequalities in the choice of secondary school: Long-term trends during educational expansion and reforms in Italy. European Societies, 16(5), 666-693. (2014)
[15] Prentice, R. L., Williams, B. J., & Peterson, A. V. : On the regression analysis of multivariate failure time data. Biometrika, 68(2), 373-379. (1981)

DRAFT

# Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation

Elena Sabbioni[a], Enrico Bibbona[a], Gianluca Mastrantonio[a], and Guido Sanguinetti[b]

[a]Politecnico di Torino; elena.sabbioni@polito.it, enrico.bibbona@polito.it, gianluca.mastrantonio@polito.it,
[b]Scuola Internazionale Superiore di Studi Avanzati (SISSA); gsanguin@sissa.it

**Abstract**

The analysis of RNA data plays a crucial role in understanding cellular differentiation. One widely-used methodology for analyzing RNA data is scVelo. However, in this paper, we show that, among other issues of scVelo, the current model formalization suffers from identifiability problems. We propose a Bayesian version of scVelo with modifications that address these issues.

***Keywords:*** scVelo, RNA, Bayesian, identifiability

DRAFT

## 1. Introduction

RNA velocity is a critical biological metric that facilitates the reconstruction of cellular differentiation at single-cell level. It provides insight into the future state of each cell, and it is closely associated with transcription from DNA to RNA, as well as the quantity of spliced mRNA in each cell. By analyzing RNA velocity, researchers can gain a deeper understanding of the underlying mechanisms driving cellular differentiation, which has important implications for fields such as developmental biology and disease research. Single-cell RNA sequencing (scRNA-seq) techniques are commonly used to measure the abundance of unspliced and spliced mRNA in each cell for each gene, which is essential for inferring RNA velocity. However, these techniques are destructive, as they permit only a single observation of gene expression for each cell before it is destroyed. In this sector, one of the most influential works is scVelo, presented in [1]. Despite its success in the scientific community, there are several criticisms when it is analyzed from a mathematical and statistical point of view.

As a primary contribution, we reframe the model under a Bayesian framework, which provides better insight into the parameters that can be estimated and identified. The use of Bayesian inference enables us to compare the posterior estimates of the parameters with their corresponding priors and compute credible intervals. In contrast, such comparisons and interval estimates are not possible with the point estimates provided by scVelo. Through simulated examples, we demonstrate that the "time" parameter is not identifiable, which, to the best of our knowledge, has not been previously identified in the literature, representing another contribution of this paper. Furthermore, we propose modifications to the model that addresses some of the criticisms of the original scVelo. Collectively, these contributions have the potential to enhance the accuracy and reliability of RNA velocity inference, as well as provide new insights into cellular differentiation.

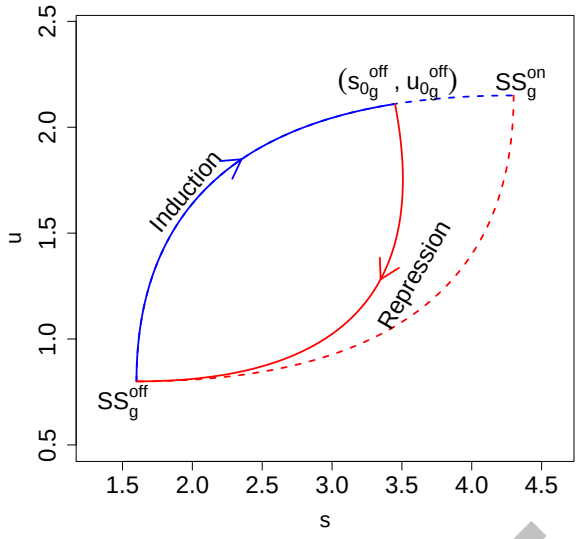



Figure 1: Solution of (2) in the space $(s, u)$ for gene $g$. The solid line represents the gene's behavior if an early switch at time $t_{0g}^{\text{off}}$ occurs, while the dashed line represents the potential behavior if the gene reaches the steady state $\text{SS}_g^{\text{on}}$ during the inductive phase. The blue upper curve corresponds to the gene's inductive phase, characterized by the rate $\alpha_g^{\text{on}}$, and the red lower curve represents the repressive phase, associated with $\alpha_g^{\text{off}}$.

## 2. Mathematical model

Let us consider a system with $n_g$ genes and $n_c$ cells. The model assumes a straightforward chemical reaction network (CRN) to represent the processes of transcription, splicing and degradation, with gene-specific rates, according to mass-action kinetics. The CRN representation of the process, for a given $g$, is

$$\text{DNA} \xrightarrow{\alpha_g^k} \text{unspliced mRNA} \xrightarrow{\beta_g} \text{spliced mRNA} \xrightarrow{\gamma_g} \emptyset \tag{1}$$

that is associated with the following ordinary differential equation (ODE) system:

$$\begin{cases} \frac{du_g(t)}{dt} = \alpha_g^k - \beta_g u_g(t), \\ \frac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t), \\ u_g\left(t_{0g}^k\right) = u_{0g}^k, \\ s_g\left(t_{0g}^k\right) = s_{0g}^k, \end{cases} \tag{2}$$

where $u_g(t)$ and $s_g(t)$ are the reads of unspliced and spliced mRNA at time $t$ in a cell. The solution to the ODEs is

$$\begin{cases} u_g(t) = u_{0g}^k e^{-\beta_g \tau_g^k} + \frac{\alpha_g^k}{\beta_g}(1 - e^{-\beta_g \tau_g^k}) \\ s_g(t) = s_{0g}^k e^{-\gamma_g \tau_g^k} + \frac{\alpha_g^k}{\gamma_g}(1 - e^{-\gamma_g \tau_g^k}) + \frac{\alpha_g^k - \beta_g u_{0g}^k}{\gamma_g - \beta_g}(e^{-\gamma_g \tau_g^k} - e^{-\beta_g \tau_g^k}) \quad \text{with } \tau_g^k = t - t_{0g}^k. \end{cases} \tag{3}$$

It should be noted that the variable $t$ is not the real time, but a representation of the cell position in the ODE dynamic, which is often called *pseudotime* in the literature, see, for example, [3].

RNA velocity is defined as

$$v_g(t) := \frac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t).$$

Accurately estimating the model parameters is crucial for obtaining a reliable estimator of this biological quantity.

DRAFT

**Parameters description** For each gene, there exist two transcription rates, indicated as $\alpha_g^{\text{on}}$ and $\alpha_g^{\text{off}}$ with $\alpha_g^{\text{on}} > \alpha_g^{\text{off}}$, represented in (2) as $\alpha_g^k$, with $k \in \{\text{on}, \text{off}\}$. The rates regulate the conversion of DNA into unspliced mRNA, as depicted in (1) (first and second components). This implies that a gene can exist in two distinct states: an inductive phase, regulated by the transcription rate $\alpha_g^{\text{on}}$, and a repressive phase, where transcription either occurs at a lower rate or is absent altogether, dictated by $\alpha_g^{\text{off}}$. It is assumed that each gene can be activated and then repressed only once, which is justified by the assumption that the total time length of the biological processes is sufficiently small. The rates $\beta_g$ and $\gamma_g$, illustrated in (1) (from the second to the fourth component), are responsible for the splicing and degradation mechanisms of mRNA. The gene time dynamic is depicted in Figure 1.

The ODE system has two theoretical steady states, that are only gene-dependent and are identified by the coordinates

$$\text{SS}_g^{\text{off}} = \left( \frac{\alpha_g^{\text{off}}}{\beta_g}, \frac{\alpha_g^{\text{off}}}{\gamma_g} \right), \qquad \text{SS}_g^{\text{on}} = \left( \frac{\alpha_g^{\text{on}}}{\beta_g}, \frac{\alpha_g^{\text{on}}}{\gamma_g} \right),$$

in the space $(s, u)$, see Figure 1. After the cell is created, each gene remains at $\text{SS}_g^{\text{off}}$ for a time period of $t_{0g}^{\text{on}}$ before being activated and entering the inductive phase, represented by the upper blue arc in Figure 1. Time $t_{0g}^{\text{on}}$ is not identifiable (see Section 3) since there is no information in the data regarding the real time point at which the cells are observed, and hence, without loss of generality, we assume $t_{0g}^{\text{on}} = 0$. This means that $s_{0g}^{\text{on}} = \alpha_g^{\text{off}}/\beta_g$ and $u_{0g}^{\text{on}} = \alpha_g^{\text{off}}/\gamma_g$. However, before reaching the second steady state $\text{SS}_g^{\text{on}}$, the repressive phase is triggered at time $t_{0g}^{\text{off}} + t_{0g}^{\text{on}}$ and the dynamic follows the evolution depicted by the solid line in Figure 1.

In scVelo the pre-processed data $(Y_{u,cg}, Y_{s,cg})'$ are assumed to be normally distributed and the unspliced and spliced components to be independent, i.e.

$$Y_{u,cg} \sim \mathcal{N}(u_g(t_{cg}), \sigma^2) \qquad Y_{s,cg} \sim \mathcal{N}(s_g(t_{cg}), \sigma^2) \qquad Y_{u,cg} \perp\!\!\!\perp Y_{s,cg}. \tag{4}$$

where $t_{cg} = \tau_{cg} + t_{0g}^{k_{cg}}$ and $u_g(t_{cg})$ and $s_g(t_{cg})$ are evaluated in a time that is both cell- and gene-specific. The description of the data used to estimate the model is discussed in Section 3. The scVelo algorithm estimates the following parameters: $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g, t_{0g}^{\text{off}})$ for each gene, and $(\tau_{cg}, k_{cg})$ for each cell and gene.

There are several criticisms of this model, that, in our opinion, raise questions about the reliability of the results, which will be discussed in the next section.

DRAFT

## 3. Critical issues of scVelo

One of the main concerns is related to the estimation of the cell- and gene-specific $\tau_{cg}$. Since single-cell data only provides a single observation for each cell, inferring $\tau_{cg}$ is inherently difficult, if not impossible. Despite this, the authors did not acknowledge this issue. As a first contribution, we demonstrate in Section 4. that $\tau_{cg}$ is at best weakly identifiable by showing that, the posterior distribution of $\tau_{cg}$ closely resembles the one of $\tau_{c'g}$, for $c \neq c'$, and they are both very similar to the prior.

Single-cell RNA-sequencing dataset contains discrete counts, describing the number of measured RNA molecules in each cell. In scVelo, a series of pre-processing steps are applied to the raw data. This includes filtering out genes that are not expressed, normalizing the counts to account for differences in sequencing depth across cells, and smoothing the gene expression profiles among groups of cells with similar genetic expressions. Additionally, the logarithm of the pre-processed counts is taken. The variables $(Y_{u,cg}, Y_{s,cg})'$ in equation (4), are the results of this pre-processing. While these pre-processing steps are common in many biological pipelines, they significantly alter the nature of the data by transforming them from discrete counts to continuous values. This transformation can be problematic, especially in real data applications where the original counts are often very low (often in the range [0,10]). The pipeline introduces dependence across genes which are not taken into account in the model, that assumes independence, see (2) and (4). Additionally, the use of a logarithmic transformation is questionable since ODE equations and solutions are not invariant under a non-linear transformation.

Despite time-dynamic being dependent on four parameters $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g)'$, only three of them are identifiable due to the lack of information on $t_{cg}$ and its scale in the data. We can easily see that, if $r \in \mathbb{R}^+$, then the parameters $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g)'$ and $(\alpha_g^{\text{off}}/r, \alpha_g^{\text{on}}/r, \beta_g/r, \gamma_g/r)'$ produce the same likelihood if $\tau_{cg}$ is substituted with $r\tau_{cg}$. This is because under both sets of parameters, the same value $(u_g(t_{cg}), s_g(t_{cg}))$ is obtained. While some of the issues discussed here have been previously addressed in the literature (e.g., [2; 5]), the non-identifiability of $\tau_{cg}$ and its impact on other parameter estimates has not been adequately emphasized to the scientific community.

In conclusion, scVelo has a further drawback in that it only provides point estimates of the parameters and does not compute any measure of their precision. This absence reduces the reliability of the results as it is not possible to assess the statistical differences among the parameters accurately.

## 4. The Bayesian Implementation

In our model formulation, we choose to use the original data without the non-linear pre-processing steps applied in scVelo. The only step we keep is the filtration of the genes that are not sufficiently expressed. As a result, our data $(Y_{u,cg}, Y_{s,cg})'$ is discrete, and $(u_g(t_{cg}), s_g(t_{cg}))'$, obtained as solution of (2), represents the mean of the original count data. A natural choice for modeling $(Y_{u,cg}, Y_{s,cg})'$ is the Poisson, because this distribution arises from the chemical master equation (CME) associated with the CRN (1). Specifically, in the steady state, the Poisson distribution is the distribution of mRNA counts of a single gene in a single cell, and in the transient part, CME distribution can be expressed as the convolution of multinomial and product Poisson distributions [4]. On the other hand, to increase the model flexibility and to take into account extra sources of variability we use Negative Binomial distribution, which is an overdispersed version of the Poisson. Specifically, we assume that

$$Y_{u,cg} \sim \mathcal{NB}(u_g(t_{cg}), \eta_g) \qquad Y_{s,cg} \sim \mathcal{NB}(s_g(t_{cg}), \eta_g) \qquad Y_{u,cg} \perp\!\!\!\perp Y_{s,cg}$$

Here the Negative-Binomial is parameterized in terms of its mean $\mu$ and the overdispersion parameter $\eta$, such that if $X \sim \mathcal{NB}(\mu, \eta)$, then $\mathbb{V}(X) = \mu(1+\mu\eta)$. As prior distributions we define $\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \gamma_g, \eta_g \sim \mathcal{N}_{[0,+\infty)}(0, 10000)$, where $\mathcal{N}_{[a,b]}$ is a truncated Normal distribution with support in $[a, b]$, and $P(k_{cg} = \text{on}) = 0.5$. For $\tau_{cg}$ we define a mixed-type distribution with two masses of value $0.1$ on $\tau_{cg} = 0$ and $\tau_{cg} = \infty$, respectively, and with probability $0.8$ we have $\log \tau_{cg} \sim N(0, 100)$. The two masses are used to locate the cell in the steady states. The prior on $t_{0g}^{\text{off}}$ must depend on the set $\{\tau_{cg}, k_{cg}\}_{c=1}^{n_c}$ since

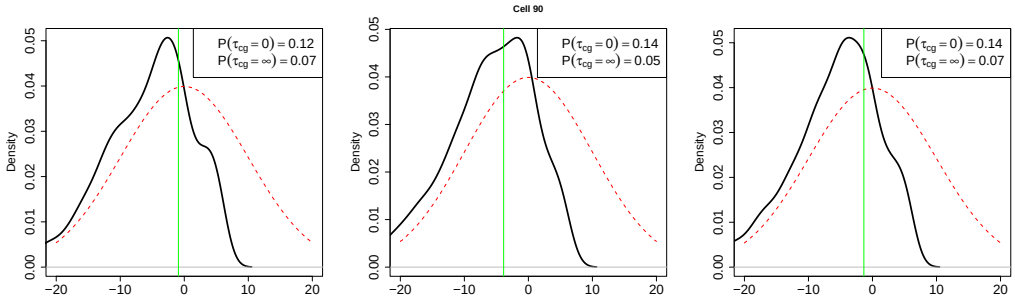
$$t_{0g}^{\text{off}} \geq \max\{\tau_{cg} | k_{cg} = \text{on}, c = 1, \ldots, n_c\} = \tau_{g,\max}^{\text{on}} \tag{5}$$

hence, we assume the following: $\log t_{0g}^{\text{off}} | \{\tau_{cg}, k_{cg}\}_{c=1}^{n_c} \sim N_{[\log \tau_{g,\max}^{\text{on}}, \infty)}(0, 100)$ To avoid identifiability issue, parameter $\beta_g$ is fixed to 1.

**Simulation setting** We simulate data with $n_g = 5$, $n_c = 3600$, and $\gamma_g$ randomly generated in $[0.5, 0.8]$, $\alpha_g^{\text{off}}$ in $[0.05, 1]$, $\alpha_g^{\text{on}}$ in $[2, 5]$, and $\eta_g$ in $[0.01, 0.1]$. These intervals have been chosen such that the empirical distribution of the raw data mimics the one of the real pancreatic dataset used in [1]. There are different issues when simulating parameters $\tau_{cg}$, $t_{0g}^{\text{off}}$, and $k_{cg}$. Indeed, we have to satisfy equation (5) and, to have realistic and diversified locations of $(s_g(t_{cg}), u_g(t_{cg}))$, as well as $(s_g(t_{0g}^{\text{off}}), u_g(t_{0g}^{\text{off}}))$ close to $\text{SS}_g^{\text{on}}$, $\text{SS}_g^{\text{off}}$ or in between. To achieve this, a sequence of if/else conditions were implemented, however for the sake of brevity these details are omitted. We run the model for $100000$ iterations, with thin $40$ and burning $10000$, having then $2250$ posterior samples.

**Discussion of the results** The posterior distributions of different $\tau_{cg}$, as shown in Figure 2, reveal that there are minimal differences between them. Additionally, the posterior distribution is similar to the prior distribution. Comparable results are obtained when changing the prior distribution, which are not presented for the sake of brevity. It should be pointed out that, under this setting, for $\log \tau_{cg} < -5$ and $\log \tau_{cg} > 5$ the coordinates $(s_g(t_{cg}), s_g(t_{cg}))$, for all $c = 1, \ldots, n_c$ and $g = 1, \ldots, n_g$, are approximately equal to the steady states with a difference of order $10^{-3}$.

DRAFT




(a) (b) (c)

Figure 2: Gene- and cell-specific $\tau_{cg}$ model. Prior (red dashed line) and posterior (black solid line) of the logarithm of $\tau_{cg}$ for three cells. The vertical line represents the true value.

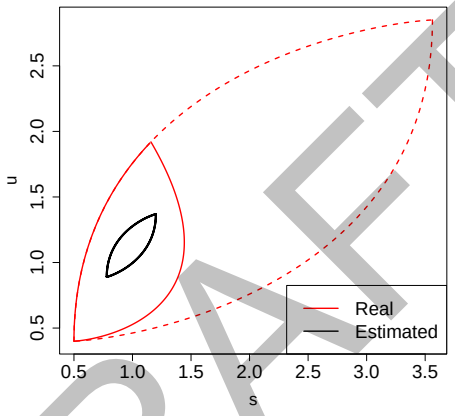



Figure 3: Phase plot in the space $(s, u)$ for a given gene. The solid lines show the solutions (3) obtained with the real parameters (red) used to simulate the data and the posterior means (black). The dashed lines correspond to the potential dynamic if the steady state $\mathrm{SS}_g^{\mathrm{on}}$ is reached. For the estimated solution, the dashed and the solid line coincide.

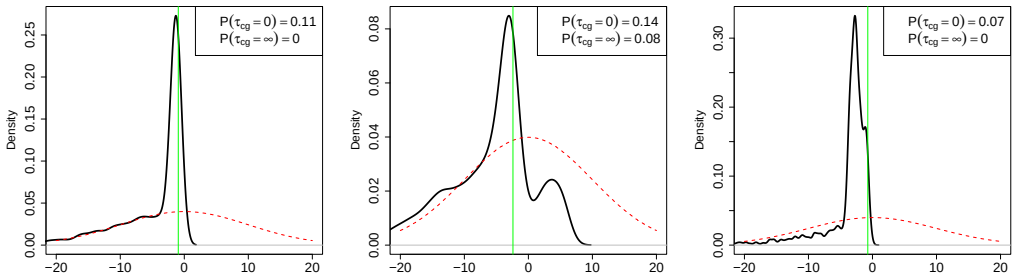
DRAFT

This illustrates the difficulty in estimating these parameters and emphasizes the importance of considering the full distribution rather than just point estimates. This issue cannot be detected with the original scVelo implementation, which only provides point estimates as output. As a consequence of this weak identifiability, all the other parameters are wrongly estimated, i.e., the associated 95% credible intervals do not contain the true value, and the entire structure in the space $(s, u)$ describing the time dynamic, is very different from the real one, as shown in Figure 3.

Simulated examples demonstrate that estimating the parameter $\tau_{cg}$ and other unknowns in the model is feasible when we have repeated measures for each $(c, g)$. The results obtained with $n_c = 8$, $n_g = 5$ and 450 repetitions are shown in Figure 4 as an example. However, in the case of single-cell data, it is not possible to have true repetitions since the variable $t_{cg}$ is unobservable/unknown. Instead of repetitions, a mixture model can be used where data share a common coordinate in the space $(s, u)$. These results suggest that this approach may be a viable direction.

## 5. Conclusions and further developments

This study highlights the issues present in the current formalization of the scVelo model. Specifically, we focus on the weak identifiability of the variable $\tau_{cg}$ and its impact on the estimation of other parame-




(a) (b) (c)

Figure 4: Repeated measurements model. Prior (red dashed line) and posterior (black solid line) of the logarithm of $\tau_{cg}$ for a three genes. The vertical lines represent the true values.

ters. To address this, we introduce a new Bayesian version of scVelo and evaluate its performance using synthetic data. Upon inspection of the posterior distribution of $\tau_{cg}$, we observe that, for a given gene, all distributions are comparable and closely resemble the prior distribution, indicating weak identifiability of the parameters.

In addition, we propose a potential solution to overcome the identifiability problem, which shows promising results in our initial investigations. We are pursuing this direction as a possible way forward in improving the performance of scVelo.

DRAFT

## References

[1] Bergen, V. and Lange, M. and Peidli, S. and Wolf, F. A. and Theis, F. J.: Generalizing RNA velocity to transient cell states through dynamical modeling. Nature biotechnology **38.12**, 1408–1414 (2020)

[2] Gorin, G. and Fang, M. and Chari, T. and Pachter, L.: RNA velocity unraveled. PLOS Computational Biology **18.9** (2022)

[3] Gupta, R., Cerletti, D.Gut, G., Oxenius, A., Claassen, M.: Simulation-based inference of differentiation trajectories from RNA velocity fields. Cell Reports Methods, **2** 1–15 (2022)

[4] Jahnke, T. and Huisinga, W.: Solving the chemical master equation for monomolecular reaction systems analytically. Journal of mathematical biology **54.1** 1–26 (2007)

[5] Marot-Lassauzaie, V. and Bouman, B. J. and Donaghy, F. D. and Demerdash, Y. and Essers, M. A. G. and Haghverdi, L.: Towards reliable quantification of cell state velocities. PLOS Computational Biology **18.9** (2022)

# Approximate Bayesian Computation for Probabilistic Damage Identification

Cecilia Viscardi[a], Silvia Monchetti[b], Luisa Collodi[a], Gianni Bartoli[b], Michele Betti[b], Michele Boreale[a], and Fabio Corradi[a]

[a]University of Florence - DiSIA, Viale Morgagni, 65, Florence; cecilia.viscardi@unifi.it, luisa.collodi@unifi.it, michele.boreale@unifi.it, fabio.corradi@unifi.it

[b]University of Florence - DICEA, via di S. Marta, 3, Florence; silvia.monchetti@unifi.it, gianni.bartoli@unifi.it, michele.betti@unifi.it

DRAFT

**Abstract**

Damage identification analyses are fundamental to guarantee the safety of civil structures. They are often formalised as inverse problems whose solution ignores any source of uncertainty that could be accounted for by using appropriate statistical models. Unfortunately, these models often exhibit an intractable likelihood function. We propose quantifying uncertainty through a fully Bayesian approach based on Approximate Bayesian Computation (ABC), a class of methods that overcome the evaluation of the likelihood and only require the ability to simulate from the model. Furthermore, we suggest a strategy to reduce ABC computational burden using Neural Networks. Finally, we test the method at work on a damaged beam to discuss its strengths and weaknesses.

***Keywords:*** Damage Identification, Uncertainty Quantification, Approximate Bayesian Computation, Neural Networks

## 1. Introduction

In the civil engineering field, structural monitoring and damage detection techniques have received growing attention since they are paramount for tasks of control and preservation. Damages modify the mechanical properties of a civil structure and can cause changes in the dynamic behaviour of the system described by natural frequencies and modal shapes. Thus, these quantities can be exploited to infer the existence of structural damages, their location and their entity. A common practice is addressing the issue as an inverse problem: optimal values of the parameters describing the properties of the system are found by minimising a distance measure between the experimental data (e.g., observed frequencies) and data produced by a Finite Element Model (FEM) - i.e. a numerical method for solving the differential equations that describe the dynamic behaviour of the system as a function of the mechanical parameters and the structural configuration. Once solved the inverse problem, one can evaluate whether there have been variations of the mechanical properties carrying pieces of information on the location and the entity of the damage. However, such a procedure ignores all the sources of uncertainty - e.g. unobserved characteristics of the system, variations of the material properties as well as measurement errors. It follows that predictions about future dynamics are taken as assured. A probabilistic damage assessment allows for taking into account different sources of uncertainty. More specifically, a Bayesian probabilistic damage identification procedure provides posterior distributions over the location and the entity of the

damage, thus avoiding a false sense of confidence. Moreover, posterior predictive distributions enable an evaluation of the uncertainty around the prediction of the future dynamic behaviour.

In the literature, there are few works addressing the problem of incorporating uncertainty in damage identification (7; 9, among others). They rely on strong assumptions and describe the relationship between observed data and mechanical properties through simple models that imply tractable likelihood functions. Implementing likelihood-free methods is a possible strategy to provide a finer description of reality. They allow a straightforward integration of the uncertainty induced by latent variables and variables having complex dependence structures. In (4; 3) the authors resort to Approximate Bayesian Computation (ABC), a likelihood-free approach, however, they do not adopt a fully Bayesian perspective and aim only at finding point estimates of the model parameters.

This paper is aimed at giving a formal statistical definition of the probabilistic model for damage's location identification building upon a proper framework for uncertainty quantification (5; 2). Furthermore, we describe a strategy to get fully Bayesian estimates using a suitable ABC algorithm. In particular, we propose a procedure based on a surrogate generative model derived via Neural Networks. We speculate that this approach will provide a flexible tool that allows straightforward integration into the model of many sources of uncertainty.

## 2. Bayesian Inference in models for uncertainty quantification

Let us denote by $\theta$ the variable object of our inference, that is the location of a damage in a structure, and by $y_0$ some observed characteristics related to its dynamic (e.g., the frequencies). Our aim is to derive the posterior distribution $\pi(\theta \mid y_0) \propto \pi(\theta)p(y_0 \mid \theta)$, given the prior distribution, $\pi(\cdot)$, and the likelihood function, $p(\cdot \mid \theta)$. [1] In this framework, the evaluation of posterior quantities requires a simulated inference approach because many unobserved variables interact with the damage's location and affects its relation with the frequencies. These variables must be included in the model as latent variables.

Let $x$ be the latent variables and $\xi = (\theta, x)$ the vector of all the unknown quantities. In principle, Monte Carlo (MC) or Markov Chain Monte Carlo (MCMC) methods allow us to get samples from a posterior distribution defined on an augmented space: $\pi(\xi \mid y_0) = \pi(\theta, x \mid y_0) \propto \pi(\theta, x)p(y_0 \mid \theta, x)$. However, here even the likelihood function $p(y_0 \mid \theta, x)$ is analytically intractable and its evaluation may be computationally demanding. To give insights into the reasons for this intractability, we provide a formal statistical definition of the model for uncertainty quantification (2). The key elements of the model are:

- $y^R(\theta)$: the vector of real values of the frequencies when the damage's location is $\theta$;
- $y^M(\xi)$: the output of a simulator that reproduces the real process. The simulator may be a numerical model for partial differential equations (e.g. the FEM) and takes both $\theta$ and $x$ as inputs;
- $b(\xi) = y^R(\theta) - y^M(\theta, x)$: the discrepancy between the model and the reality. It may come from incorrect or missing physical characteristics, as well as the simplification of the problem needed to put it in a digital framework (e.g. space discretisation in FEM).
- $y^E(\xi) = y^M(\xi) + \eta(\xi)$: the emulator. It is an approximation of the simulator and $\eta(\xi)$ is the discrepancy between the simulator and the emulator.
- $y_0(\theta) = y^M(\xi) + b(\xi) + e$ : the observed data. They typically differ from the real process for some measurement errors $e$.

In this scenario, the probability $p(y_0 \mid \xi)$ can be retrieved from $p(y_0, b, e \mid \xi)$ via marginalisation:

$$p(y_0 \mid \xi) = \int\int p(y_0 \mid e, b, \xi)p(e, b \mid \xi)de\,db = \int\int \delta_{y_0}(y^M + b + e)p(e)p(b \mid \xi)de\,db \qquad (1)$$

where $\delta_{y_0}(\cdot)$ is the Dirac measure. Note that Eq (1) comes from the assumption that measurement errors

[1] For the sake of simplicity, our notation does not discriminate between probability density functions and mass functions that can be distinguished from the context.

DRAFT

are independent of $\xi$ and $b$, and from the fact that the simulator is a deterministic numerical model that, once a vector $\xi$ is given as input, always returns the same output $y^M(\xi)$.

MC and MCMC algorithms for the computation of the $\pi(\xi \mid y_0)$ would involve multiple point-wise evaluations of $p(y_0 \mid \xi)$ and each of them requires the solution of the integrals in Eq (1). The computation of $y^M(\xi)$ makes exact calculations infeasible and numerical approximations are computationally demanding: a single evaluation of the integrals would require many runs of the FEM. This motivates the choice of simulation-based methods, such as Approximate Bayesian Computation (ABC).

## 3. ABC for probabilistic damage identification

The origin of ABC methods can be traced back to (11; 8) but, in the last twenty years, huge progress has been made in this field. For a comprehensive description of the method, we refer the reader to (10). The key idea of the basic ABC algorithm is to get samples from an approximate posterior distribution by converting samples from the prior into samples from the posterior in three steps: 1) generate $N$ parameter values from the prior distribution $\pi(\cdot)$; 2) generate simulated processes $y_i \sim p(\cdot \mid \theta_i)$ for $i \in \{1, ..., N\}$; 3) retain only parameter values $\theta_i$ such that $d(y_i; y_0) \leq \epsilon$, where $d(\cdot\,;\,\cdot)$ is a distance function and $\epsilon \geq 0$ is a tolerance threshold.

The algorithm avoids the evaluation of the likelihood function. It only requires the ability to produce samples from $p(\cdot \mid \theta)$ using a *generative model* that can be thought of as a computer code which takes parameters $\theta$ as inputs, performs stochastic calculations that involve latent variables $x$, and outputs simulated data $y$. However, this solution to the problem of the intractability of the likelihood comes at the cost of introducing at least one source of approximation in the estimate of the posterior distribution. In particular, the quality of the approximation depends on the tolerance threshold $\epsilon$: the approximate posterior distribution converges to the true posterior distribution as $\epsilon \to 0$.

Besides the basic ABC algorithm, many other sampling schemes have been proposed – see (10, Ch 4). Here, we resort to a sampling scheme inspired by the Population Monte Carlo ABC (PMC-ABC) presented in (1). It is displayed in Algorithm 1 where $K_j(\cdot \mid \theta_i^{j-1})$ is a Normal distribution with mean $\theta_i^{j-1}$ and variance equal to twice the weighted empirical variance of $(\theta_1^{j-1}, ..., \theta_N^{j-1})$ , and $\alpha$ is a tuning parameter between 0 and 1. The output of the algorithm is a sample from the following approximate posterior distribution

$$\pi_\epsilon(\theta \mid y_0) = \pi(\theta) \int \mathbb{1}\{d(y; y_0) \leq \epsilon\} p(y \mid \theta) dy$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and $\epsilon = \epsilon_M$ is the value of the threshold adaptively chosen.

Algorithm 2 describes the generative model used to get samples from $p(\cdot \mid \theta)$. Note that each run of the generative model involves a call to the FEM. Here we propose a strategy to reduce the computational cost of this procedure. In particular, we replace the simulator with a less expensive emulator based on Neural Networks.

DRAFT

**Algorithm 1** ABC-PMC

Sample $\theta_1^0, ..., \theta_N^0$ from $\pi(\cdot)$.
Sample $y_i$ using Alg. 2 giving $\theta_i^0$ as input for each $i \in \{1, ..., N\}$.
Let $d_i = d(y_i; y_0)$ for each $i \in \{1, ..., N\}$.
Put $\epsilon_1$ equal to the $\alpha$-quantile of the distribution of $(d_1, ..., d_N)$.
**for** $j = 1, 2, \ldots, M$ **do**
  Set $i = 0$
  **while** $i < N$ **do**
    Sample $\theta^*$ from $q_j(\cdot) = \frac{\sum_{i=1}^{N} w_i^{j-1} K_j(\cdot|\theta_i^{j-1})}{\sum_{i=1}^{N} w_i^{j-1}}$.
    Sample $y^*$ using Alg. 2 giving $\theta^*$ as input.
    Compute $w^* = \frac{\pi(\theta^*)}{q_j(\theta^*)} \mathbb{1}\{d(y^*; y_0) \leq \epsilon_j\}$.
    **if** $w^* > 0$ **then**
      Let $\theta_i^j = \theta^*$, $d_i^j = d(y^*; y_0)$, $w_i^j = w^*$ and $i = i + 1$.
    **end if**
  **end while**
  Put $\epsilon_{j+1}$ equal to the $\alpha$-quantile of the distribution of $(d_1^j, ..., d_N^j)$.
**end for**

**Algorithm 2** Generative model

Take $\theta$ as an input.
Sample $x$ from its prior distribution.
Compute $y^M(\theta, x)$ using the simulator (FEM) or the emulator (ANN).
Sample $b(\xi)$ and $e$ from their distributions.
Return $y = y^M(\xi) + b(\xi) + e$.

**Approximating the simulator using Neural Networks** Artificial Neural Networks (ANN) are computational models that use experience to learn functions. We resort to *feedforward neural networks* which process information from inputs to outputs through intermediate computations and without feedback connections. The basic processing unit of an ANN is the *neuron* that receives inputs from other neurons and computes its own output using a linear combination based on previously defined weights and bias, and an *activation function*. Neurons are then arranged in layers: the first layer is called the input layer, the last layer is the output layer and in between layers are the hidden layers that determine the depth of the network. The network learns by adjusting weights and biases to minimise the prediction error, which is the value of the loss function that quantifies the difference between the output of the network and the output taken from a training set.

In the case of the probabilistic damage identification, the relation that links the unknowns, $\xi$, to the simulated process, $y^M$, is specified by a deterministic function reproduced using the FEM. Thus, an ANN can be used as an emulator that replaces FEM in Algorithm 2. The ANN takes as input layer $\xi$ and gives as output layer an emulated process $y^E(\xi)$. To train the ANN we need a training set built by considering $S$ vectors, $\xi^1, ..., \xi^S$, giving them as input to the FEM, and taking $S$ simulated process as output. It follows that our approach gives a computational advantage as long as $S$ is smaller than the number of simulations from the generative model in Algorithm 1.

DRAFT

## 4. Simulation study

To illustrate the proposed method, we considered the example of a simply supported beam modelled by using the commercial code ANSYS®. The total length of the structure is 5000 mm and the cross-section is a IPE240 steel profile – see Figure 1.

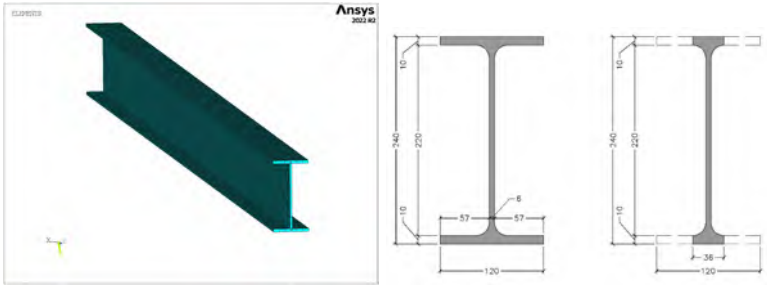



Figure 1: Discretisation in ANSYS®(left) and cross-section (right) of the beam without and with damage.

The induced localised damage is imposed by cutting the beam flanges. This damage is reproduced in ANSYS®by reducing the section shown in Figure 1. The effects, in terms of observable dynamic characteristics of the beam, are a reduction of the frequencies and changes in modal shapes. Here, we want to infer the location of the damage, $\theta \in (0, 5000)$, using three observed frequencies, $y_0 = (f_1, f_2, f_3)$ expressed in Hz. The latent variable $X$ represents the uncertainty on the restraint conditions of the beam, in particular on its location. We assumed $\theta \sim \text{Uniform}(0, 5000)$ and $\frac{X}{100} \sim \text{Exponential}(\lambda = 1.5)$. Measurement errors $e = (e_1, e_2, e_3)$ are distributed as a Multivariate Normal with mean $\mu_e = (0, 0, 0)$ and covariance matrix $\Sigma_e = 0.15^2 I_3$. We included in the model also a random discrepancy $b \sim \text{Uniform}(-0.2, 0.2)$.[2]

In our simulation study, the observed data $y_0 = (32.63, 96.73, 208.61)$ have been produced running the FEM assuming $\theta^{\text{true}} = 2423.8$ and $x^{\text{true}} = 6.18$ (values generated at random). We trained the ANN with 2 inputs ($\theta$ and $x$), three hidden layers with 50 neurons, and 3 outputs ($f_1$, $f_2$ and $f_3$). We used the Relu activation function for all the layers, the Mean Square Error (MSE) as loss function, and Adam (6) as the optimization algorithm. Our training set has size $S = 160\,000$. The performance of the trained network is good enough to consider negligible the discrepancy between the simulator and the emulator (MSE $= (1.09 \cdot 10^{-5}, 4.88 \cdot 10^{-6}, 9.48 \cdot 10^{-6})$ and $R^2 = (0.99, 0.99, 0.99)$ computed on a test set).

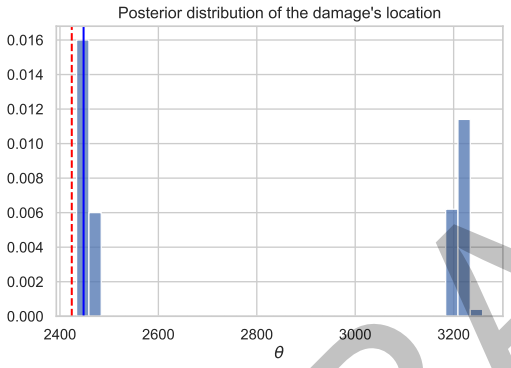



Figure 2: Posterior distribution of $\theta$ with $\theta^{\text{true}}$ (red line) and $\theta^{\text{MAP}}$ (blue line).

| | **Posterior predictive distributions** | | |
|---|---|---|---|
| | mean | s.d. | 95% CI |
| $f_1$ | 38.60 | 2.22 | 32.88; 41.03 |
| $f_2$ | 113.31 | 6.26 | 97.70; 120.85 |
| $f_3$ | 243.56 | 14.41 | 210.60; 263.30 |

Table 1: Mean, standard deviation and 95% credible intervals of the posterior predictive distributions.

We ran Algorithm 1 for 20 minutes with $N = 500$ and using the Euclidean distance. In the given budget of time, the final number of iterations is $M = 22$, and all of them required more than 100 000 calls to the generative model to accept 500 parameter proposals. The final threshold is $\epsilon_M = 0.07$.

Looking at Figure 2 we can see that the Maximum a Posteriori estimate of the damage's location, $\theta^{\text{MAP}} = 2448$, is very close to $\theta^{\text{true}}$. Note that the FEM discretises the beam using meshes of size 50 mm, meaning that the difference $\theta^{\text{MAP}} - \theta^{\text{true}} = 24.2$ mm can be ignored since it is too small to be detected by the model. The bimodality of the posterior distribution is due to the symmetry of the beam. However, in the reality this perfect symmetry does not occur, thus we speculate that posterior distributions based on real data will be unimodal. The uncertainty about the damage's location has been propagated to future frequencies by computing posterior predictive distributions described in Table 4. The variability of the distribution increases moving from the first to the third frequency. In a more realistic framework, the uncertainty would be even larger and ignoring posterior predictive distributions may lead to the observation of completely unexpected scenarios.

## 5. Discussion and future work

In this work, we investigate the use of a formal model for uncertainty quantification in the identification of damages in civil structures. A probabilistic approach is essential to be aware of the uncertainty

[2]Prior distributions have been set exploiting information coming from preliminary investigations as well as the experts' knowledge.

DRAFT

around estimates and predictions and to conduct a more conscious process of decision-making. However, including different sources of uncertainty often leads to complex probabilistic models with an intractable likelihood function. We propose a likelihood-free approach to provide fully Bayesian estimates. The presented ABC method overcomes problems related to the computational cost of the simulator resorting to an emulator. In particular, we exploit the deterministic nature of the function that links parameters and latent variables to the frequencies and propose an emulator based on ANNs.

Our exploratory analysis showed that the method is able to infer the damage's location and gave some insights into the uncertainty of future frequencies pointing out the importance of considering posterior predictive distributions. This aspect makes the proposed framework particularly relevant in the structural health monitoring field.

One of the main strengths of the proposed approach is its flexibility. In our example, we assumed simple Gaussian and Uniform distributions over the measurement errors and the bias of the model. However, the method allows one to straightforwardly replace them with more complex random variables – e.g. considering Gaussian processes (5). In fact, we need only the ability to produce simulations from the assumed distributions and no analytical evaluations are required. Furthermore, in this framework, the statistician can take full advantage of the expert knowledge in the specification of the prior distributions and the definition of a model that is as close as possible to reality.

The major drawback of the method is that it still requires a large number of calls to the FEM to build a training set for the ANN. However, in our example, the size of the training set turned out to be far lower than the number of simulations needed in the PMC-ABC procedure. Furthermore, the use of the trained emulator is not limited to the implementation of the PMC-ABC algorithm since it can be integrated into the health monitoring system which, otherwise, would call the FEM many times.

Future work should focus on the application of the method to real data and on the extension to more complex structures such as bridges, towers, etc. Moreover, we plan to define a more sophisticated model that allows inferring also the presence/absence and the entity of the damage.

DRAFT

## References


[1] M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[2] J. O. Berger and L. A. Smith. On the statistical formalism of uncertainty quantification. *Annual review of statistics and its application*, 6:433–460, 2019.

[3] S.-E. Fang, S. Chen, Y.-Q. Lin, and Z.-L. Dong. Probabilistic damage identification incorporating approximate bayesian computation with stochastic response surface. *Mechanical Systems and Signal Processing*, 128:229–243, 2019.

[4] Z. Feng, Y. Lin, W. Wang, X. Hua, and Z. Chen. Probabilistic updating of structural models for damage assessment using approximate bayesian computation. *Sensors*, 20(11):3197, 2020.

[5] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[6] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization 3rd int. In *Conf. for Learning Representations, San*, 2014.

[7] Y. Lei, Y. Su, and W. Shen. A probabilistic damage identification approach for structures under unknown excitation and with measurement uncertainties. *Journal of Applied Mathematics*, 2013.

[8] J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[9] R. Rocchetta, M. Broggi, Q. Huchet, and E. Patelli. On-line bayesian model updating for structural health monitoring. *Mechanical Systems and Signal Processing*, 103:174–195, 2018.

[10] S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[11] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

# Estimation of scientific productivity with a hierarchical Bayesian model

Maura Mezzetti[a] and Ilia Negri[b]

[a]Department of Economics and Finance, Università "Tor Vergata";
maura.mezzetti@uniroma2.it
[b]Department of Economics, Statistics and Finance "Giovanni Anania", Università della Calabria;
ilia.negri@unical.it

**Abstract**

A new measure for the scientific productivity of researchers is introduced. Through a Bayesian hierarchical model, the productivity of Italian researchers is reconstructed and used to understand whether within Italian academics gender difference can be assessed. The new measure takes into account not only the quantity and quality of the scientific production but also the time that elapses between two successive research products, obtaining a productivity curve for each male and female scientist. The results show a persistence of women in the lower levels of the university career even if their scientific productivity is greater than that of their male colleagues.

***Keywords:*** Piecewise Bayesian linear model, Research evaluation, Gender gap

DRAFT

## 1. Introduction

In gender study there exists a major stream of literature that demonstrates the presence of a so-called *productivity gap* in favor of men (1). The lesser productivity of female researchers has been underlined for different disciplines and countries (6; 4). Moreover women academics tend to advance in career at a slower pace and are rewarded less than males (10). Differences in the high tail of the productivity distribution, the so call *top* or *star* scientists, have been studied by (3) and (2). From these studies emerges the idea that to asses about gender differences in productivity, not only the mean value of the productivity has to be considered but the entire distribution along the entire period of productivity has to be investigated. Moreover, not only the quantity, but also the quality of the research output needs to be taken into account. In the next section a measure of individual research productivity is introduced. A curve of productivity is defined that takes in account not only the quantity and quality for any publication but also the time elapse between two successive publications along all the academic career of any scholar. The obtained raw curve is smoothed through a Bayesian hierarchical model that estimates the productivity curve using a piecewise linear regression model as a function of time. The different slopes correspond to different levels of productivity of the researchers, while the instants of time where the slope changes correspond to the different phases that can be supposed the individuals face along the entire period of observation. The curves estimated in this way allow to observe what happens at the lowest levels of the university career. In fact, we are going to investigate what happens among permanent researchers, a figure of tenured assistant professor present only in Italy, where we observe a sort of stagnation of researchers who do not progress to the higher level of the academic hierarchy. To better understand the reasons of this stagnation the reconstruction of the productivity along

the entire career of all academicians at any level is done through the proposed hierarchical Bayesian model. The hierarchical Bayesian model permits to estimate the entire distribution of the productivity and furthermore to analyze the scientific productivity by gender and by position. The estimated model seems to confirm that gender differences are present in the lower tail of productivity distribution, giving rise to a sort of pocket of stagnation.

## 2. A researcher productivity's measure

Evaluation of research productivity and its impact on the scientific community is extremely difficult to achieve, because the multifaceted nature of evaluation, the lack of standard terminology, and the heterogeneity of research fields make hard to identify a preferred model of measurement. As a matter of fact, research productivity in academia is defined differently among academic fields but most of the measures involved to this purpose are related to publications in books and journals. A commonly used measure of research productivity is publication rate (9). Publication rate takes into account only quantity of publications and not quality. As a measure of scientific impact, or quality, a wide range of indicators and metrics are available and their choice depends on considerations of their strengths and limitations. The most widely used research output indicators are based on bibliometric indices and citation parameters (e.g. number of publications in peer-reviewed journals, impact factor and H-index). In this study the productivity for each scholar is measured not only through the number of publications but also considering the impact of the publication on the scientific community. In this direction, for sake of simplicity, let us measure the productivity considering only publications, but the same approach can be used for any product that define the scientific production of a scholar. Let $t_1, \ldots, t_n$, denote the time (in year) when the $n$ papers of a scholar are published. To each publication is associated a weight $p_i$ that represent its quality. The productivity curve is given by the cumulative score, defined for $0 \leq t < T$ as $c(t) = \sum_{i=1}^{n} p_i I_{[t_i \leq t \leq t_{i+1}]}$, where $t_1 = 0$ and $t_{n+1} = T$. Here $T$ represents a hypothetical last moment by which all scientists still manage to publish a paper. A hierarchical Bayesian model is defined and estimated (5) as follows. The *productivity* observed can be characterized by different phases and can be reasonably described by a piecewise linear model with time as independent variable and cumulative score as dependent. In our approach a statistical framework based on a piecewise linear regression with random breakpoints is provided for each subject, in order to allow to describe the productivity along all the career life. The basic idea behind this approach is that each scientist experiments different phases during the career. The *production* is not constant in time due to many factors, as for example maternity leaves or periods committed to governance duties, that we do not have information on. Based on the cumulative curve of the production of any scholar, the Bayesian piecewise linear linear model is able to estimate different slope values that represent different rate of productivity for any scholar. Let $Y_{it}$ be the cumulative score for subject $i$ at year $t$, ($t = 0$ for the year of the first publication) we assume the following model: $Y_{it} \sim N(\mu_{it}, \sigma_i^2)$ where $\mu_{it} = \alpha_i + \beta_i^0 t + \sum_{h=1}^{H} \beta_i^h (t - t_i^h) I_{[t > t_i^h]}$. Here for each subject $i$, parameters $\beta_i^h$ are the individual slopes, the parameters $t_i^h$ are the individual breakpoints and $H$ is the fixed number of different phases can be supposed the individuals face during all the period of observation. This is fixed and chosen by the researcher. According to the Bayesian paradigm the priors for the individual regression parameters are chosen appropriately. According to this model the eventual change in *scientific production* speed is estimated from the data and can be different for every subject. For any $i$, the estimated curve $\mu_{it}$ represents a proxy of the *scientific production* of researcher $i$, along the time of his/her entire career $t$. The estimated slopes of the piecewise linear regression represent the speed of the production. An average of the slopes weighted with time is considered the *average speed* of production. An advantage of a hierarchical Bayesian model is the possibility to fit a second level curve. The model is thus producing, on one hand, an individual curve and, on the other hand, an overall curve. The second level can be specified according to a partition of the subjects in $G$ groups. The partition can be defined either a priori o through a cluster analysis. In the second level the same breakpoints for all the subjects as to be considered. So, in the estimated global curve, for each group, not only the number of breaks are the same for all the subject, but also their location. In the application they are chosen reasonably according to the development of a classic academic career.

DRAFT

## 3. Data

Data are extracted from three different data sources and combined. First, for any scholar working in an Italian university, the academic positions, gender, university and disciplines are available from the Ministry of Education, University and Research website, from 2000 to nowaday. Then from Scopus the publications of each researcher during the entire career were downloaded. Data analysed in the paper refer to March 2021, there were 466 individuals covering a position in Statistics. Finally each publication is linked to the database of the Italian National Agency for the evaluation of Universities and Research institutes (ANVUR) that provides a score of the journal where the paper is published according to four indicators. See (8) for the details. The cumulative score $Y_{it}$ for each schoolar $i$ in year $t$ is computed as follows. If $J$ articles were published in the year $t$ by the researcher $i$, a score $p_j$ is associated with each publication and $y_{it}$ is given by the sum of the scores of all the publications published in that year.

## 4. Results

Following the approach of a piecewise linear regression illustrated in Section 2, a line with four linear trends, $(\beta_i^0, \beta_i^1, \beta_i^2, \beta_i^3)$, is estimated with different estimated break points $(t_i^1, t_i^2, t_i^3)$ for each researcher $i$. The posterior estimates are the average values of the 5000 MCMC iterations after 50000 burn-in iterations. The choice of $H = 3$ is done according to the idea that the entire period of productivity can be split in four periods. Precisely: an initial part, in the first years of the career, where the scholar starts to produce, followed by a second period, where the scholar consolidate is productivity, then a third period characterized by the maturity of the scholar, and a final period till the end of the career. The posterior estimates result quite robust to the choice of hyperparameters. When a change in the linear trend is observed, it corresponds to a change in the production, either a deceleration or an acceleration of productivity. The punctual slope can be interpreted as the increase of the *productivity score* in one year.

### 4.1 Tenured Assistant Professors

We focus our analysis on tenured assistant professors of the research area Statistics, classified as SECS/S-01 in Italian university system. In Italy positions for tenured assistant professors were not opening anymore after 2011. At the end of 2021 in the area of Statistics there were 51 tenured assistant professors. The tenured assistant females are more than males but if this gap were small at the early 2000's than it increases till the 40% of 2021. Figure 1 (left) shows the *scientific production* along the entire career of 44 Tenured Assistant Professors, reconstructed with a piecewise linear curve according to model presented in Section 2. We have to exclude 7 of these subjects with null publication score because the model cannot be estimated for them. Observing the curves it is evident that red curves, corresponding to female researchers, tend to reach higher values: female tenured assistant professors obtain higher scores in publications respect to male tenured assistant professors. This is confirmed fitting a second level curve within each group. Indeed, we add a second level in the model, that considers an average model for two groups, men and women. As mentioned at the end of Section 2, we need to fix the change points of time. We assume three fixed change-points in $(5, 10, 20)$. This choice means we suppose that each scientist can experience a change in the speed of his/her production after 5, 10 and 20 years after the start of his/her career. These points are a reasonable choice considering the typical career of an academic scholar, (7). The chosen points represents an average of the time where scholars usual face with a change in production. The estimated second level curves are shown in Figure 1 (right). The two curves estimated for tenured assistant professors are separated, the curve for female is higher than the corresponding curve for male. The curves represent the growth in research performance for male (blue) and female (red). Difference between gender can easily be observed. Women show to better perform in terms of productivity along all the career. At the end of the observed time, female researchers reach an average value of *productivity* of 63 versus an average male values of 25. The top 20% of most productive tenured assistant professors are all women. The posterior density of parameters shows that the intercept is superior for women. The biggest difference is observed in the parameter representing the change of

DRAFT

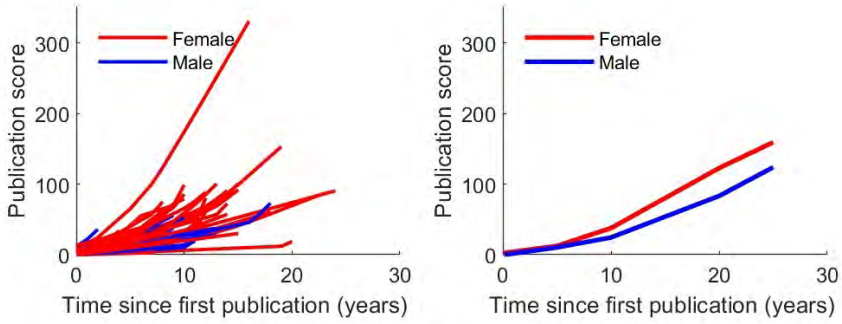



Figure 1: On the left curves of publication score for 44 Tenured Assistant Professors. On the right Second level curves of publication score for the two groups. Red for female, blue for male

slope after 5 years. The mean value of this parameter within men is equal to 0.63, while the mean value of the same parameter within women is 2.97. We analysed data regarding the qualification to associate professor, the necessary title to get promoted to an higher level in academia. The waiting time elapsed from getting the national qualification to be promoted associate professor is about the same between gender, more specifically waiting time from national qualification to associate professor is 1.63 years for men and 1.73 for women. In the residual group of 44 tenured assistant professor, twelve of them (27%) got the habilitation. Between them 10 (83%) are women. Between the 32 that did not take the habilitation 20 (63%) are women. The chi-square test cannot reject the hypothesis of independence between gender and getting the habilitation. A logistic model that relates ability to get the national qualification with the final score within tenured position by gender reveals a significant effect only for women, suggesting that the women should work harder to get rewarded.

DRAFT

## 5. Conclusions

All these facts seem to confirm that the lower tail of distribution of scientific performance is affected by gender difference as pointed out in (1). Women tenured assistant professors continue to work and their productivity is greater than that of men, but they receive promotion less than their male colleagues. So we can argue that it exists a pocket of stagnation for women that cannot leave this position. The analyses for the other levels of the academic career do not show the same differences by gender observed for the tenured assistant professors, both for the second level estimated curves and for the posterior distributions of the parameters. Further analysis could be done including information about application for a higher position.

## References

[1] Abramo, G., Cicero, T., D'Angelo, C. A. Should the research performance of scientists be distinguished by gender?. Journal of Informetrics, 9(1), 25-38. (2015)

[2] Abramo, G., D'Angelo, C. A., & Di Costa, F. A gender analysis of top scientists' collaboration behavior: evidence from Italy. Scientometrics, 120, 405-418. (2019)

[3] Bordons, M., Morillo, F., Fernández, M. T., & Gómez, I. One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. Scientometrics, 57(2), 159-173. (2003)

[4] Fox, M. F. Gender, family characteristics, and publication productivity among scientists. Social Studies of Science, 35(1), 131-150. (2005)

[5] Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. Bayesian data analysis. Chap. & Hall. (1995)
[6] Larivière, V., Ni, C., Gingras, Y., Cronin, B., Sugimoto, C. R. Bibliometrics: Global gender disparities in science. Nature, 504(7479), 211-213. (2013)
[7] Győrffy, B., Csuka, G., Herman, P., & Török, Á. Is there a golden age in publication activity? - an analysis of age-related scholarly performance across all scientific disciplines. Scientometrics, 124, 1081-1097. (2020)
[8] Mezzetti, M., Negri, I. Hierarchical Bayesian model to estimate and to compare research productivity among academics, submitted. 2023.
[9] Rørstad, K. & Aksnes, D. W. Publication rate expressed by age, gender and academic position - A large-scale analysis of Norwegian academic staff. Journal of informetrics, 9(2), 317-333. (2015)
[10] Schneider, B. Z., Carden, W., Francisco, A., Jones Jr, T. O. Women "opting out" of academia: At what cost. In Forum on Public Policy (Vol. 2, No. 1, pp. 1-19). (2011)

DRAFT

# Heat waves and free–knots splines

Gioia Di Credico[a] and Francesco Pauli[a]

[a]Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti",
University of Trieste, Trieste, Italy;
gioia.dicredico@deams.units.it, francesco.pauli@deams.units.it

**Abstract**

We investigate the trend of heat waves in four different locations in Europe. In particular, we employ a flexible semiparametric nonlinear model to detect possible accelerations in the phenomenon's growth. Heat waves are expected to increase in frequency and intensity globally, but the geographical heterogeneity is substantial, thus making local analysis relevant. Our results confirm the expectations in that we detect a significant nonlinear increase in the frequency of heat waves; on the other hand, the intensity appears constant.

***Keywords:*** heat wave, knot locations, TLB, Bayesian inference

DRAFT

## 1. Introduction

The study of climate is the study of the average behaviour of weather. Analyzing the global mean temperature (or a global mean temperature) is a typical example and has been thoroughly discussed (4). The global mean temperature, however, does not exist: it is an artificial construct useful to summarize the climate's state but not something that anyone experiences. Alongside the studies focusing on general behaviour, it is also relevant to investigate more local phenomena that have a direct impact, such as heat waves. Heat waves (HW) can harm the health of the human population, crop yield, energy demand and production, and economic activity. It is also relevant to point out that while the fact that the world is experiencing an increase in the global temperature is clearly recognizable and acknowledged, this does not necessarily translate into an increase of HWs at every location and the same pace. The specificities of the local climate and the fact that the extreme behaviour does not necessarily follow the same trend as the mean may lead to different behaviours (4; 5). In this work, we aim to assess whether some locations in Europe experienced an increase in HW and whether there has been an acceleration in the most recent period. Several solutions exist to model a nonlinear relationship between two quantities. Spline functions, piecewise polynomials with a fixed degree whose joint points are called knots, are one of the most flexible and efficient. Within the various approaches to specifying and estimating a spline function, we employ the method proposed by (2) since it explicitly models the position of the knots, which would correspond to changes in the pace of the trend of the phenomenon. There are no universally agreed criteria to define an HW. We will not discuss the various possibilities and consider a relatively standard definition: a heat wave occurs when the daily maximum temperature is above a daily varying threshold (precisely, the 90th percentile of the daily maximum temperatures observed in a window of 15 days during a reference period, here 1961-1981) for at least three days.

## 2. Methods

The presence of a piecewise linear effect of the covariate on the response variable is evaluated through a linear spline function with free knots. Each knot location, treated as a parameter, detects a change in the slope of the linear spline and a discontinuity on the first–order derivative. In the generalized linear model (GLM) under the Bayesian framework, the probability distribution of the response and the link function is chosen accordingly to the selected response variable. In contrast, the linear predictor specification is common to all the presented models. Without loss of generality, we define the linear predictor for the $i$-th observation, as

$$\eta_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \gamma_k (x_i - \xi_k)_+, \tag{1}$$

for $i = 1, \ldots, n$, where $x_i$ is a continuous predictor, $\beta_0, \beta_1$, and $\gamma_k$, for $k = 1, \ldots, K$, are the spline coefficients, $\xi_k$ represents the location of the $k^{\text{th}}$ knot and $K$ is the total number of knots. The function $(x - \xi_k)_+$ is the truncated linear function defined as $(x - \xi_k) I(x \geq \xi_k)$. By construction, the truncated linear basis (TLB) may be characterized by a high correlation among basis functions, sometimes leading to estimation problems. Therefore we suggest applying this methodology when a limited number of change points can be assumed. In this framework, the ease of interpretability of the knot locations and the spline coefficients is the main advantage of choosing the TLB representation.

Following the default prior definition as in (3), the prior distributions of the spline coefficients and the dispersion parameter are chosen as

$$\beta_0 \sim N(g(m_y), 10 s_y), \quad \beta_1 \sim N(0, 2.5 s_y / s_x), \tag{2}$$

$$\gamma_k \sim N(0, 2.5 s_y / s_{\xi_k}), \quad \sigma \sim \exp(1/s_y) \tag{3}$$

where $g(.)$ is the link function, $m_y$ and $s_y$ are equal to the mean and standard deviation of the response variable in the Gaussian case, and $s_y = 1$ otherwise. Also, $s_x$ is the standard deviation of the predictor, and $s_{\xi_k}$ is computed on the $k$-th truncated linear basis term. These prior distributions allow us to handle the different scales of the variables automatically. The autoscaling characteristic is appealing in the TLB situation; indeed, the closer a knot is to the upper bound of the predictor, the smaller the variance of the associated basis element. In Sect. 3, we consider two GLMs: a Poisson and a Gaussian with canonical link function. The prior distribution of each knot location $\xi_k$ is set as Uniform on the range of the predictor. When more than one knot is assumed, knot locations are subject to order constraints to ensure identifiability.

Lastly, considering the number of knots as a parameter requires applying transdimensional techniques such as the Reversible-jump Markov chain Monte Carlo. However, due to the limited number of assumed knots and the simple model specification, we decided to fit several models with fixed but increasing numbers of knots. The final number of knots is selected through a model's diagnostic checks and comparing the approximate leave-one-out cross-validation (LOO) information criterion, a fully Bayesian equivalent to the AIC (7), and the weights associated with model averaging via stacking of predictive distributions (8).

## 3. Results

We compute yearly summaries for the period 1950-2021 of HW phenomena in four geographic locations; in particular, we considered the following grid cells (long $\times$ lat): $A$: $[26, 27] \times [68, 69]$, located in Finland; $B$: $[-2, -1] \times [53, 54]$, located in UK; $C$ : $[9, 10] \times [45, 46]$, located in Italy; $D$: $[44, 45] \times [56, 57]$, located in Russia. For the purpose of this work, an HW is defined as a period of at least three days in which the maximum daily temperature exceeds the 90–th percentile of the distribution of the maximum temperature in a window of 15 days in a reference period. The temperature time series were

DRAFT

obtained from the Berkeley Earth Surface Temperature dataset (6). We consider summer HWs, that is, HW occurring between May and September of each year.

An HW is a multidimensional phenomenon characterized by duration and intensity; in order to keep into account this multidimensionality, multiple yearly summaries are considered, in particular: the count, i.e., the number of observed heat waves; the duration, i.e., the sum of the duration of all HWs in that year; the maximum duration, i.e., the duration of the longest HW for that year; the average intensity, i.e., the average temperature above the threshold during HWs in that year; the maximum intensity, i.e., the temperature above the threshold in the HW with the maximum observed excess. We model the first and the third using a Poisson regression model with a log link function, and the others assuming a Normal distribution.

We run four chains of 2000 iterations each using the NUTS sampler through Rstan interface (1). Inference is based on the simulation's second half, that is, 4000 final simulations. LOO results and stacking weights agree on the model selection; therefore, we select the model with the lowest expected log pointwise predictive density and highest stacking weight. No divergence transitions appear in the algorithm's sampling step, and the effective sample size parameter and the Rhat statistic do not highlight sampling difficulties for the selected models. As shown in Figure 1, one knot resulted in all the Italian location models, on the count model for UK and Russia, and the maximum duration model for Finland. Two knots are selected in the mean and maximum duration models in the UK and the count model for Finland. The linear predictor shows a slope change at each knot location, mostly of positive signs. The posterior mean and quartiles estimates of the knot locations are reported in Table 1. Figure 2 reports the results for the Poisson models on the count and maximum duration response variables scale.

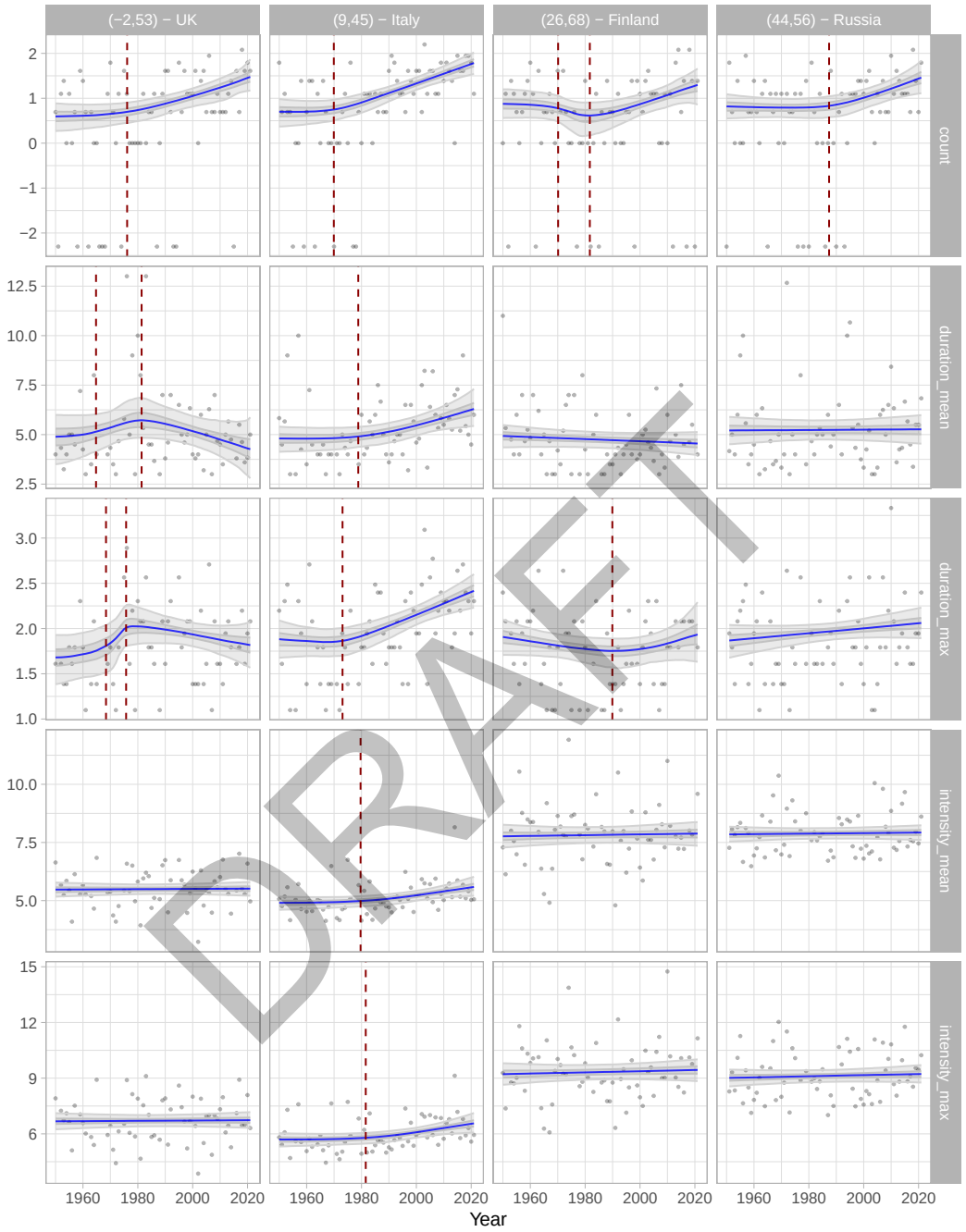
Table 1: Posterior summaries for knots locations.

| | Response | $\xi$ | Posterior Mean | 25% | 50% | 75% |
|---|---|---|---|---|---|---|
| (-2,53) - UK | Count | $\xi_1$ | 1977.30 | 1966.12 | 1976.10 | 1985.09 |
| | Duration | $\xi_1$ | 1968.06 | 1957.61 | 1964.76 | 1974.56 |
| | Duration | $\xi_2$ | 1984.87 | 1977.07 | 1981.38 | 1991.01 |
| | Duration max | $\xi_1$ | 1968.19 | 1961.45 | 1968.40 | 1971.83 |
| | Duration max | $\xi_2$ | 1978.01 | 1973.82 | 1975.73 | 1979.93 |
| (9,45) - Italy | Count | $\xi_1$ | 1969.72 | 1964.54 | 1969.91 | 1975.30 |
| | Duration | $\xi_1$ | 1979.39 | 1970.15 | 1978.82 | 1989.28 |
| | Duration max | $\xi_1$ | 1974.05 | 1968.52 | 1973.08 | 1978.60 |
| | Intensity max | $\xi_1$ | 1979.58 | 1969.50 | 1981.55 | 1989.29 |
| | Intensity mean | $\xi_1$ | 1978.27 | 1967.32 | 1979.66 | 1988.72 |
| (26,68) - Finland | Count | $\xi_1$ | 1972.05 | 1963.44 | 1970.14 | 1977.95 |
| | Count | $\xi_2$ | 1984.31 | 1977.11 | 1981.70 | 1989.68 |
| | Duration max | $\xi_1$ | 1987.12 | 1977.81 | 1989.90 | 1997.74 |
| (44,56) - Russia | Count | $\xi_1$ | 1986.41 | 1980.14 | 1987.38 | 1993.23 |

DRAFT

## 4. Conclusion

One of the effects of climate change is an increase in heat waves frequency and intensity at a global level, with substantial variability in space and time. We study the interannual variability by employing a nonlinear semiparametric model whose estimates suggest that an acceleration in the frequency of the phenomenon occurred since approximately the 80s in all locations considered (see Table 1 and Figure 1). The intensity, on the other hand, does not show a significant increase. These results are broadly coherent with previous analyses of the BEST data. Although the fact that we consider four very specific locations makes the comparison with regional estimates not obvious, we notice that in (5), an increase in frequency was noted for the European area. At the same time, the average intensity had no significant trend. We plan to elaborate on this preliminary exploration by considering more locations to explore the geographical heterogeneity of the phenomenon in its various manifestations (exploring the different measures of frequency and intensity proposed in the literature) and its spatial structure. In particular, we plan to implement models allowing for spatial correlation and the pattern of variation across space.




DRAFT

Figure 1: Results of the fitted model on the linear predictor scale. Locations are on the columns with longitude and latitude between brackets, while response variables are on the rows. Grey points represent the observed values. The solid blue line is the linear predictor computed using the posterior median estimates of the spline coefficients and the knot locations (vertical red dashed lines).

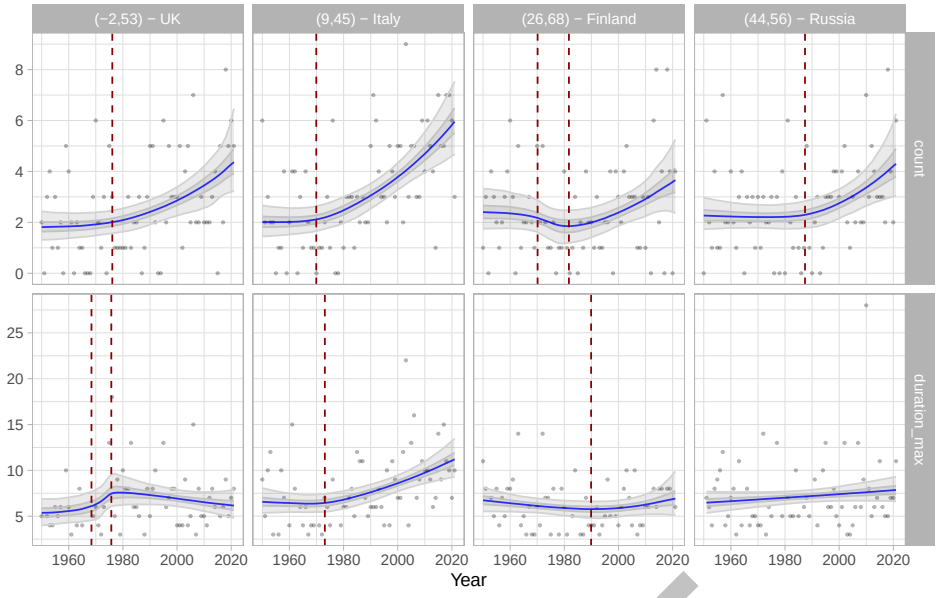



Figure 2: Results of the fitted Poisson models on the response scale. Locations are on the columns with longitude and latitude between brackets, while response variables are listed on the rows. Grey points represent the observed values. The solid blue line is the linear predictor on the exponential scale computed using the posterior median estimates of the spline coefficients and the knot locations (vertical red dashed lines).

DRAFT

## References


[1] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language, J. Stat. Softw., **76**, 1–32 (2017)

[2] Di Credico, G., Edefonti, V., Polesel, J., Pauli, F., Torelli, N., Serraino, D., Negri, E. et al.: Joint effects of intensity and duration of cigarette smoking on the risk of head and neck cancer: A bivariate spline model approach, Oral oncology **94**, 47–57 (2019)

[3] Gelman, A., Hill, J., Vehtari A.: Regression and other stories. Cambridge University Press (2020)

[4] IPCC. Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the IPCC [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Camb. Univ. Press (2022) doi:10.1017/9781009325844.

[5] Perkins-Kirkpatrick, S. E., and S. C. Lewis. Increasing trends in regional heatwaves. Nature communications **11.1**, 3357 (2020).

[6] Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C., Mosher S.: Berkeley earth temperature averaging process, Geoinformatics Geostatistics Overview, **1**, 20–100 (2013)

[7] Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. (2017) doi: 10.1007/s1122201696964

[8] Yao, Y., Vehtari, A., Simpson, D., Gelman A.: Using Stacking to Average Bayesian Predictive Distributions (with Discussion). Bayesian Anal. **13(3)**, 917–1007 (2018)

# The Hierarchical Beta-Bernoulli Process as Out-of-Scope Query Detector

Marco Dalla Pria[a] and Silvia Montagna[a]

[a]Università degli Studi di Torino, C.so Unione Sovietica 218/bis, Torino;
marco.dallapria@unito.it, silvia.montagna@unito.it

**Abstract**

Task-oriented dialog systems are computer systems that interact with humans in natural language. The system receives a query, converts the sequence of words into a semantic representation to be used by the dialog manager, decides the best response for the user, and manages the task. Occasionally, the system may receive an *out-of-scope* query, namely, a query that falls outside the range of the system's capabilites. In this work, we focus on out-of-scope query prediction, and show how the hierarchical Beta-Bernoulli process outperforms state-of-the-art machine learning classifiers.

***Keywords:*** Hierarchical Beta-Bernoulli processes, nonparametric Bayesian modelling, task-oriented dialog systems, out-of-scope query prediction

DRAFT

## 1 Introduction

The increasing sophistication of machine learning algorithms in the last years has led to a revolution in task-oriented dialog systems: nowadays people can ask Amazon's Alexa what is their bank account balance while they are cooking, and will get a satisfactory answer from her. Any dialog system is designed to support a fixed number of intents only. For example, a task-driven system designed to support personal finance queries cannot answer the question "What is the weather like tomorrow?". Queries falling outside the range of intents which the dialog system is designed to work upon are defined *out-of-scope* queries (hereafter, OOS). Correctly identifying that a query is OOS is of paramount importance for the system to avoid performing wrong actions. Thus far, however, little attention has been given to evaluating the performance of state-of-the-art, dialog system machine learning classifiers in OOS prediction. An exception is given by [1], who evaluate and compare the performance of a range of benchmark classifier models focusing on OOS prediction relying on a novel dataset. Whilst the tested models work well in predicting known intents, the authors show that all methods struggle with identifying OOS queries.

The hierarchical Beta-Bernoulli process is a well known Bayesian nonparametric process that has shown good performance in document classification tasks [2]. Informally speaking, documents are a collection of words, thus queries themselves can be seen as documents. However, unlike long-text documents, the fact that queries consist of only a few words is an obstacle towards distinguishing an OOS from an in-scope query, which indeed become indistinguishable if a couple of key words were removed from the OOS query; see Table 1. In face of these difficulties, we believe that the flexibility of a nonparametric model could be instrumental in detecting OOS queries. In this work, we fit a Beta-Bernoulli process to the classification data in [1], and show that it outperforms benchmark machine learning classifiers in OOS query prediction.

The remainder of this paper is organised as follows. In Section 2, we introduce the discrete form of the hierarchical Beta-Bernoulli process [2], which we leverage on in this work, and explain how this process can serve as nonparametric Bayesian prior in document classification tasks. Section 3 illustrates the inferential procedure leading to the classification of an unlabelled document. In Section 4, we analyse the dataset in [1] by means of the hierarchical Beta-Bernoulli process, and Section 5 presents conclusions and directions for future work.

## 2 Methods

The (discrete) Beta process, denoted $BP(c, B_0)$, is a Lévy process over a space $\Omega$ whose Lévy intensity is defined by:

$$\nu(d\omega, dp) = \sum \text{Beta}(cq_i, c(1-q_i))(dp)\delta_{\omega_i}(d\omega)$$

where the *base measure* $B_0 = \sum_i q_i \delta_{\omega_i}$ is discrete, the positive real constant $c$ is the *concentration parameter*, and the total mass $\gamma = B_0(\Omega)$ of the base measure is called *mass parameter*; $\gamma$ is required to be finite. Note that each $q_i$ must lie in $(0, 1)$ in order for the Lévy intensity to be well defined.

The (discrete) Bernoulli process, denoted $BeP(B)$, is the Lévy process characterised by the Lévy intensity:

$$\mu(d\omega, dp) = \sum \text{Bernoulli}(p_i)(dp)\delta_{\omega_i}(d\omega)$$

where the base measure $B = \sum_i p_i \delta_{\omega_i}$ is discrete. Note that the masses $p_i$ must lie in $(0, 1]$ in order for the Lévy intensity to be well defined. The probability of a particular realisation of a $BeP(B)$ with $B$ discrete is:

$$\mathbb{P}\left(X = \{\omega_{j_1}, \omega_{j_2}, ..., \omega_{j_K}\}\right) = \prod_{k=1}^{K} \mathbb{P}\left(\omega_{j_k} \in X\right) = \prod_{k=1}^{K} \int_{[0,1]} \text{Ber}\left(p_{j_k}\right)(dp) = \prod_{k=1}^{K} p_{j_k}$$

Given a discrete base measure $B_0$ and a positive real constant $c$, we can combine the two processes above and obtain the Beta-Bernoulli process (BBp):

$$B \sim BP(c, B_0), \quad X|B \sim BeP(B)$$

and conjugacy holds, that is, given $n$ conditionally *iid* samples $X_1, ..., X_n|B \sim BeP(B)$,

$$B|X_1, ..., X_n \sim BP\left(c + n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i\right).$$

Indeed, by independence over disjoint subsets of $\Omega$ and by the discreteness of $B_0$, inference can be carried out separately for each atom $\omega_i$. The result follows from the well-known Beta-Binomial conjugacy.

We follow [2] and embed the BBp into a hierarchical model, leading to the hierarchical BBp (hBBp):

$$B \sim BP(c_0, B_0), \quad B_1, ..., B_J|B \sim BP(c_j, B), \quad X_{1,j}, ..., X_{n_j,j}|B_j \sim BeP(B_j) \quad j = 1, ..., J$$

An helpful analogy to better understand the model is the following. Consider a corpus $X$ of $n$ documents, each of which is associated with one of $J$ subjects, such that there are $n_j$ documents $X_{1,j}, ..., X_{n_j,j}$ belonging to topic $j$. Any given word in the English vocabulary is more or less likely to appear in a document depending on the subject: some technical terms will be exclusively used in certain domains, other terms are likely to appear in affine subjects, and some other will be ubiquitous. Let us model a document as a subset of words in some vocabulary, or, equivalently, as a binary vector whose components are indexed by the words in the vocabulary. At a given component, $0/1$ stands for the absence/presence of that word in the document, respectively.

Taking the underlying space $\Omega = \{\omega_1, \omega_2, \omega_3, ...\}$ as the vocabulary, which is potentially infinite (think of every possible misspelled word) but countable, let $B_0 = \sum_i b_0(\omega_i)\delta_{\omega_i}$ be a discrete measure

DRAFT

over $\Omega$ such that $b_0(\omega_i) \in (0,1)\ \forall i$, and let $c_0$ be a real constant. Then, $B$ is also discrete, and $B(\omega_i) \sim \text{Beta}(c_0 b_0(\omega_i), c_0(1-b_0(\omega_i)))\ \forall i$. In this setting each $B_j$ is discrete, supported by $\Omega$ and such that $B_j(\omega_i) \sim \text{Beta}(c_j B(\omega_i), c_j(1-B(\omega_i)))$. Then, $B_j(\omega_i)$ gives the probability that word $\omega_i$ appears in a document with topic $j$, while $B(\cdot)$ encodes the sharing of information between topics. Parameters $c_0, c_j$ encode the semantic richness overall and within each topic, respectively. Specifically, if $c_j$ is small then documents of subject $j$ contain the same few words, whereas if $c_j$ is large documents of topic $j$ are dissimilar. Similarly, a small $c_0$ induces a lot of shared terms among the different topics, while a large $c_0$ means that each topic has its own set of specialised terms. Concentration parameters $c_0, c_j$'s will play a key role for the performance of the hBBp.

## 3 Inference

We rely on the hBBp presented above to assign a topic to an unlabelled document $X_{\text{new}}$. Here we explain the inferential procedure.

Because Lévy processes have independent increments over disjoint sets, it is legitimate to carry out inference separately for the set of observed words (the bag-of-words obtained by the union of all the words in every document) and for its complement (the words in $\Omega$ which never appeared in any document of the corpus). Let $\Omega_{\text{obs}} \subset \Omega$ be the collection of unique words that appeared at least once in some document, and let $\Omega_0$ be its complement. The procedures discussed in this Section are summarised in Algorithm 1 and Algorithm 2.

**Inference over $\Omega_{\text{obs}}$** Fix a $\omega \in \Omega_{\text{obs}}$ and define $x_{i,j} = X_{i,j}(\omega)$, $x = \{X_{i,j}(\omega) | j \leq J; i \leq n_j\}$, $b_j = \{B_j(\omega) | j \leq J\}$, $b = B(\omega)$, and $m_j = \sum_{i=1}^{n_j} x_{i,j}$. It is possible to show that:

$$\mathbb{P}\left(x_{n_j+1,j} = 1 \middle| x\right) = \mathbb{E}\left[\mathbb{E}\left[b_j | b, x\right] | x\right] = \mathbb{E}\left[\frac{c_j b + m_j}{c_j b + m_j + c_j(1-b) + (n_j - m_j)} \middle| x\right] = \frac{c_j \mathbb{E}\left[b|x\right] + m_j}{c_j + n_j}.$$

The posterior expectation $\mathbb{E}\left[b|x\right]$ is not available in closed form analytically, thus we will rely on its Monte Carlo approximation. In particular, it is possible to show that the density of $b|x$ is bounded above by the unnormalised density of a Gamma$(\alpha, \beta)$, where

$$\alpha = c_0 b_0 + \sum_{j=1}^{J} 1(m_j > 0),\ \beta = \frac{c_0(1-b_0) - 1}{1-b^*} - \sum_{j=1}^{J} \sum_{i=1}^{m_j - 1} \frac{c_j}{c_j b^* + i} + \sum_{j=1}^{J} \sum_{i=0}^{n_j - m_j - 1} \frac{c_j}{c_j(1-b^*) + i}$$

where $b^*$ is the mode of the density of $b|x$, which can be easily obtained by any appropriate numerical optimisation method, being such a density concave in $(0,1)$. Relying on the Gamma$(\alpha, \beta)$ as proposal distribution, we generate $T$ samples $b_1, b_2, \ldots, b_T$ via Metropolis-Hastings and then approximate $\mathbb{E}\left[b|x\right]$ via the empirical mean. After some testing, we realised that the Gamma approximation is very precise: $20 \leq T \leq 30$ samples yield a satisfactory approximation in the application discussed hereafter.

**Inference over $\Omega_0$** Fix some $\omega \in \Omega_0$. Adapt the notation used above to this new $\omega$. As before, we would like to compute the probability $\mathbb{P}\left(x_{n_j+1,j} = 1 \middle| x = 0\right)$, but this turns out to be more challenging. Given $W$ words $\{\omega_1, ..., \omega_W\} \subset \Omega_0$, and defining $\lambda_j := \sum_{k=1}^{K} \frac{c_0 B_0(\Omega_0)}{c_0 + k - 1} p_{k,j}$,

$$\mathbb{P}\left(\{\omega_1, ..., \omega_W\} \subset X_{n_j+1,j} \middle| x = 0\right) \approx \text{Pois}\left(\lambda_j\right)(W) \prod_{i=1}^{W} b_0(\omega_i)$$

where $p_{k,j} = \mathbb{P}_k(x_{n_j+1,j} = 1, x = 0)$ and $\mathbb{P}_k$ is the probability over the slice of the hBBp corresponding to $\omega$, and is approximated via simulation. The larger $K$, the better the approximation, which is in fact exact for $K \to \infty$. See Algorithm 2 and [2] for further details.

DRAFT

Combining the above, given a new document $X_{\text{new}} = \{\omega_1, ..., \omega_W\}$ whose subject has to be inferred, we compute

$$\mathbb{P}(X_{n_j+1,j} = X_{\text{new}}|X) \approx \prod_{\omega \in X_{\text{new}} \cap \Omega_{\text{obs}}} \frac{c_j \mathbb{E}\left[B(\omega)|X\right] + m_j}{c_j + n_j} \times \text{Pois}\left(\lambda_j\right)(W) \prod_{\omega \in X_{\text{new}} \cap \Omega_0} b_0(\omega).$$

We compute this probability for all topics $j = 1, \ldots, J$, and then assign $X_{\text{new}}$ to the topic $j^*$ that maximises the probability above.

**Algorithm 1:** HBBp training

**Data:** corpus $X$ of $n$ documents, $n_j$ documents for each topic $1 \leq j \leq J$ ; $\Omega$ ; $c_1, \ldots, c_J$ ; $B_0$
$\gamma \leftarrow$ mean number of unique words in a query ;
$B_0 \leftarrow \frac{B_0}{B_0(\Omega)}\gamma$ ;
$\Omega_{\text{obs}} \leftarrow$ unique words in the corpus ;
$F \leftarrow |\Omega_{\text{obs}}|$ ;
$c_0 \leftarrow$ solution of $c_0 = \frac{F-\gamma}{\gamma \log\left(\frac{c_0+n}{c_0+1}\right)}$ ;
**for** $\omega \in \Omega_{\text{obs}}$ **do**
  $M_{\omega,j} \leftarrow$ number of documents of topic $j$ having $\omega$, for $1 \leq j \leq J$
  $b^*_\omega \leftarrow$ mode of the posterior density of $B(\omega)|X(\omega)$
  $\alpha_\omega \leftarrow c_0 b_0(\omega) + \sum_{j=1}^{J} 1(M_{\omega,j} > 0)$, with $b_0(\omega) := 0$ if $\omega \notin \Omega$
  $\beta_\omega \leftarrow \frac{c_0(1-b_0(\omega))-1}{1-b^*_\omega} - \sum_{j=1}^{J}\sum_{i=1}^{M_{\omega,j}-1} \frac{c_j}{c_j b^*_\omega + i} + \sum_{j=1}^{J}\sum_{i=0}^{n_j - M_{\omega,j}-1} \frac{c_j}{c_j(1-b^*_\omega)+i}$
**end**

**Algorithm 2:** Document classification via hBBp

**Input:** New document $X_{\text{new}} = \{\omega_1, \ldots, \omega_W\}$ ; trained hBBp ; $T_1, T_2, K \in \mathbb{N}$
**Output:** Most likely topic $j$ s.t. $X_{\text{new}} = X_{n_j+1,j}$
**for** unique $\omega \in X_{\text{new}} \cap \Omega_{\text{obs}}$ **do**
  $b_1, \ldots, b_{T_1} \leftarrow$ Metropolis-Hastings with Gamma($\alpha_\omega$, $\beta_\omega$) proposal and target $B(\omega)|X(\omega)$
  $\bar{b} \leftarrow \frac{1}{T_1}\sum_{i=1}^{T_1} b_i$
  $P_{1,j} \leftarrow P_{1,j} \frac{c_j \bar{b} + M_{\omega,j}}{c_j + n_j}, \ \forall\, 1 \leq j \leq J$
**end**
**for** $1 \leq k \leq K$ **do**
  $b_1, \ldots, b_{T_2} \leftarrow$ sample from Beta(1, $c_0 + k - 1$)
  **for** $1 \leq j \leq J$ **do**
    $r_{i,j} \leftarrow \frac{c_j b_i}{c_j + n_j} \prod_{j'=1}^{J} \frac{\Gamma(c_{j'})\Gamma(c_{j'}(1-b_i)+n_{j'})}{\Gamma(c_{j'}(1-b_i))\Gamma(c_{j'}+n_{j'})}, \forall\, 1 \leq i \leq T_2$
    $p_{k,j} \leftarrow \frac{1}{T_2}\sum_{i=1}^{T_2} r_{i,j}$
  **end**
**end**
$\lambda_j \leftarrow \sum_{k=1}^{K} \frac{c_0 B_0(\Omega_0)}{c_0+k-1} p_{k,j}, \ \forall\, 1 \leq j \leq J$ ;
$P_{2,j} \leftarrow \text{Pois}(\lambda_j)(W) \times \prod_{\text{unique } \omega \in X_{\text{new}} \cap \Omega_0} b_0(\omega), \ \forall\, 1 \leq j \leq J$ ;
**return** $j^* \leftarrow 1 \leq j \leq J$ maximising $P_{1,j} \times P_{2,j}$

DRAFT

## 4 Data Analysis and Results

We fit the hBBp to the CLINC150[1] data. The dataset contains a training set made of 15000 in-scope queries, 100 for each of 150 intents, 100 OOS training queries, and a test set made of 4500 in-scope queries and 1000 OOS queries. A snapshot of the CLINC150 data is provided in Table 1.

Table 1: A snapshot of in-scope and OOS queries from the CLINC150 dataset.

| Query | Intent |
|---|---|
| what is the temperature in costa mesa | weather |
| does france have their own version of a visa | international_visa |
| where can i pick up my w2 to do my taxes | w2 |
| pay my gas bill from my saving account | pay_bill |
| what do i have on my calendar for march 2 | calendar |
| who are some notable alumni of ucsd | OOS |
| when was nintendo created | OOS |
| when was the theory of evolution first considered | OOS |
| why do males want to be alpha | OOS |
| what are van gogh's best pieces | OOS |

We consider as our space $\Omega$ the set of the most common words in Wikipedia[2], which contains more that 280000 terms. Despite its size, such a vocabulary does not contain many frequent terms appearing in the dataset. Indeed CLINC150 queries are full of misspelled words, symbols, numbers and proper names. However this is not an issue for the hBBp in that its nonparametric nature allows $\Omega$ to grow as the data is observed. Here the underlying assumption is that if $\omega$ is observed in a training query but it is not in $\Omega$, then we treat it as if $\omega$ belongs to $\Omega$ with $b_0(\omega) = 0$. The prior distribution over such $\omega$ is improper, but becomes proper after the Bayesian update. Therefore, misspelled words have been retained in the dataset. Further, we did not remove stop words, which indeed appear to be informative predictors especially when appearing in clusters (e.g., "how would I...in...?", often appears under the intent "translate"), and no stemming has been applied.

Note that a test query might include a word that has never been observed in the training set and is also not present in $\Omega$. Indeed, this is quite common, especially if the test query is OOS. Assuming $b_0(\omega) = 0$ is not a good choice in such case since this would translate into a zero probability of observing such query under every intent, and the classification would not be possible. To overcome this issue, we add a special out-of-training (OOT) feature to the vocabulary of Wikipedia's most common terms. Specifically, if a test query contains a word that has never been observed in the training set and is not present in Wikipedia's vocabulary, then such a word is mapped to OOT and is interpreted as a feature of the query at hand.

To choose an appropriate $B_0$, we rely on a power-law determined by the ranking in the list of Wikipedia's most frequent terms. Having shifted down in the ranking each word by one position (the most frequent word becomes the second most, the second most frequent becomes the third most, and so on) and having put OOS on top of the ranking in the first position, the chosen power-law is $\text{rank}^{-0.1}$, the exponent close to zero to avoid the tail from becoming too thin. The total mass of $B_0$ has also to be coherent with the data. One can show that $\gamma = B_0(\Omega)$ is the mean number of unique words per document, which amounts to 8.31 in CLINC150.

Choosing the concentration parameters is challenging. Besides $c_0$, which is computed as the fixed point solution of a real valued function (see Algorithm 1 and [2]), the crucial point is the choice of the concentration parameters associated to the 151 topics. Unfortunately, an exhaustive grid search for the optimal combination for these hyperparameters is unfeasible given our computational resources. After some trial and error, a tuning "by hand" led to choices giving a good trade-off between in-scope accuracy and OOS recall, and is the chosen setting leading to the results presented hereafter.

[1] https://github.com/clinc/oos-eval
[2] https://en.lexipedia.org/

DRAFT

Table 2: In-scope and out-of-scope performance comparison between the hierarchical Beta-Bernoulli process and benchmark machine learning methods in [1]: FastText, CNN, MLP, BERT neural networks; SVM, a linear support-vector classifier; Google's DialogFlow and Rasa's NLU conversational AIs.

| Classifier | In-Scope Accuracy | Out-of-Scope Recall |
|---|---|---|
| FastText | 89.0 | 9.7 |
| SVM | 91.0 | 14.5 |
| CNN | 91.2 | 18.9 |
| DialogFlow | 91.7 | 14.0 |
| Rasa | 91.5 | 45.3 |
| MLP | 93.5 | 47.4 |
| BERT | **96.9** | 40.3 |
| Hierarchical Beta-Bernoulli process | 86.5 | **79.5** |

A comparison with benchmark machine learning methods is displayed in Table 2, where results referring to models from FastText to BERT are taken from [1] (see [1] for details on these models). These results are promising: although slightly underperforming in terms of in-scope accuracy, the hBBp outperforms in terms of OOS recall, which is the goal of our application. Further, these results should be treated as preliminary results for our work and we expect both in-scope and OOS perfomance to improve with more accurate tuning of the model hyperparameters, as done for the machine learning competitors instead.

## 5 Conclusions

In this paper, we proposed a hierarchical Beta-Bernoulli process for OOS query prediction. The methodology outperforms state-of-the-art machine learning techniques used by task-based dialog systems, and its in-scope performance is in line with that of existing techniques. Moreover, it can handle misspelled words in a straightforward and appealing manner.

Possible future work could be the estimation of the number of different topics within the OOS class via Kingman's *Coalescent*, which plays the role of a nonparametric prior over the dendrogram governing the clustering structure of OOS queries.

DRAFT

## References

[1] Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., and Mars, J.: An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In: *Proc. of the 2019 Conf. on Empir. Methods in Nat. Lang. Process. and the 9th Int. Jt. Conf. on Nat. Lang. Process. (EMNLP-IJCNLP)*. Assoc. for Comput. Linguistics, 2019, pp. 1311–1316.

[2] Thibaux, R. and Jordan, M. I.: Hierarchical Beta Processes and the Indian Buffet Process. In: *Proc. of the Eleventh Int. Conf. on Artif. Intell. and Stat.* Vol. 2. Proc. of Mach. Learn. Res. PMLR, 2007, pp. 564–571.

# A novel definition of comorbidity based on the Global Burden of Diseases project weights

Angela Andreella[a], Lorenzo Monasta[b], and Stefano Campostrini[a]

[a]Deparment of Economics, Ca' Foscari University of Venice, Italy;
angela.andreella@unive.it, stefano.campostrini@unive.it
[b]Institute for Maternal and Child Health - IRCCS Burlo Garofolo, Italy;
lorenzo.monasta@burlo.trieste.it

**Abstract**

Understanding comorbidity characteristics is essential for policymakers and healthcare providers to allocate resources accordingly. However, several definitions of comorbidity can be found in the literature. The main reason for these differences lies in the available information about the analyzed diseases (i.e., the target population analyzed), how to define the burden of diseases, and how to aggregate the occurrence of the detected health conditions. In this manuscript, we focus on the data from the Italian surveillance system PASSI proposing a definition of comorbidity based on the disability weights coming from the Global Burden of Disease project. Thanks to that, we can explore the level of comorbidity based on the presence of ten different non-communicable diseases across socioeconomic sub-populations in Italy.

***Keywords:*** Health-statistics, comorbidity, disability weights, surveillance system PASSI, global burden of disease project

DRAFT

## 1. Introduction

The term comorbidity indicates the simultaneous presence of two or more diseases in the same person. These conditions can be related or unrelated and co-occur one after the other. Comorbidity is common, especially in older people and those with chronic conditions. It can complicate the diagnosis and treatment of individual conditions and impact the person's overall health, and well-being (12).

Understanding morbidity is vital for supporting decision-making health processes. It helps policymakers, healthcare providers, and other stakeholders identify the most pressing health issues facing a population and allocate resources accordingly. Morbidity data can help researchers and scientists identify trends and patterns in the occurrence of different diseases and conditions, which can inform the development of new treatments and prevention strategies (5; 10).

The concept of comorbidity is complex and multidimensional. For that, several definitions can be found in the literature (15). Three main reasons can be identified: (i) analyzing comorbidity depends on the type of available data (e.g., the type of detected diseases in the surveillance/study considered, study objective, target population), (ii) the burden of disease on a person's health depends on the severity or duration of the diseases or conditions, (iii) the presence of multiple morbidities must be aggregated in some way to define the concept of comorbidity. For example, Pastore et al.(11) defines the concept of comorbidity as a binary variable describing the presence of at least one disease over ten detected diseases. The authors of (11) used data from the Italian surveillance system PASSI (1), a monthly cross-sectional study where self-declared health status, diagnosed diseases, and socio-demographic variables are recorded.

Therefore, the definition proposed by Pastore et al.(11) is pretty simple. The burden of disability associated with each disease is not taken into account, as is the burden of having at least one disease versus having no disease. The authors themselves, in fact, suggest using some weights to take into account the level of possible disability coming from each detected disease. For that, in this work, we propose a new definition of comorbidity, focusing on the Italian framework. We then analyze the same surveillance system used by Pastore et al.(11) (i.e., PASSI), considering as disease weights the ones coming from the Global Burden of Diseases (GBD) project (9; 7). These weights, called disability weights, reflect the magnitude of health loss linked with specific health conditions (13). The disability weights are computed using data from surveys based on paired comparison questions. The respondents must consider two hypothetical individuals with different health state names (randomly selected), and they must indicate which one is considered healthier (13). Many factors can influence the computation of the disability weights, i.e., health state description, the panel of judges, valuations methods for health states, time presentation, and surveying techniques (4). However, this paper focuses on defining the comorbidity rather than the disability weights. We decided then to use the GBD disability weights, having been tested and validated several times over time.

The outline of the paper is as follows. Subsect. 2.1 shows the steps to create the comorbidity index based on the disability weights coming from the GBD and non-communicable diseases declared in the Italian surveillance system PASSI. An example of how using this novel comorbidity index (i.e., random forest (2)) is provided in Subsect. 2.2. This analysis permits understanding the level of comorbidity in Italian sub-populations characterized by different socio-economic statuses. Finally, Sect. 3. shows the results of the application of the random forest.

# 2. Methods

## 2.1 Comorbidity index

We use data from the Italian surveillance system PASSI which includes the following question: “Has a doctor ever diagnosed or confirmed you with one or more of the following diseases?”. The surveillance system then measures the following self-reported non-communicable diseases: diabetes, kidney failure, bronchitis/emphysema/respiratory failure, myocardial infarction/cardiac ischemia/coronary artery disease, tumor (including leukemias and lymphomas), chronic liver disease/cirrhosis, stroke/cerebral ischemia, heart diseases (e.g., valvulopathy decompensation), bronchial asthma, and arthrosis/arthritis (e.g., rheumatoid, arthritis, gout, lupus, fibromyalgia). In addition, socio-economic variables are collected and used in this analysis, such as sex, age, level of education (low and medium-high), and the economic difficulties (yes and no) of the respondent. For further details about these socio-economic variables, please refer to the work of Pastore et al.(11).

When the aim is to analyze the comorbidity of a population, one must take into account that the impact on a person’s life of a given disease depends on the severity of this disease. In addition, since the same individual can declare more than one disease, we must define a way to aggregate multiple severities in order to define the concept of comorbidity. In order to take these two aspects into account, we use the disability weights coming from the GBD (13). So, we must associate each non-communicable disease measured by the surveillance system PASSI with one disability weight coming from the GBD. These weights are measured on a scale from $0$ to $1$, where $0$ equals a state of full health, and $1$ equals a state of death. However, we must deal with several problems in this step. First, the GBD provides weights for $440$ diseases, whereas PASSI only examines ten non-communicable diseases. Secondly, for each disease analyzed, the GBD provides different weights depending on the severity of the disease. In order to solve these two problems, we moved in two steps.

As a first step, we selected the diseases considered by the GBD that recall the non-communicable diseases detected by PASSI. For example, focusing on diabetes, we selected through a text mining process all those weights that refer to diseases containing the words ”diabet”, ”diabetes”, ”diabetic”, ”diabeetus”, ”diabetes mellitus”, ”hypertension”, ”obesity”, and ”insulin”. We deal with singular and plural, and the keywords include synonymous coming from the Cambridge English dictionary (6).

DRAFT

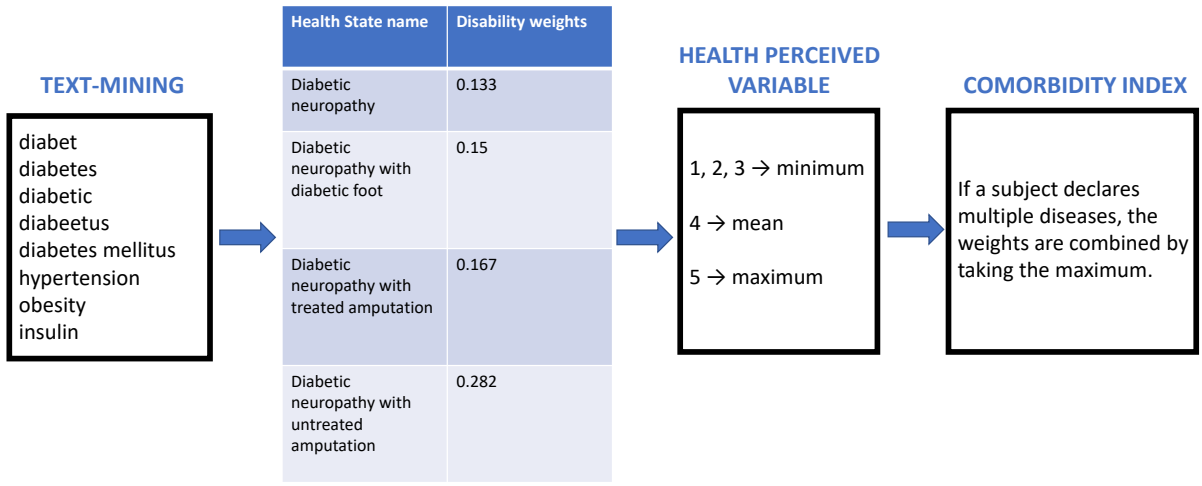



Figure 1: Steps to associate the weights coming from the GBD to the diseases declared in the Italian surveillance system PASSI.

From the first step, for example, still focusing on diabetes, we found 4 weights with a standard deviation equal to 0.06. In order to choose which of these 4 weights to associate with the individual who declared to have diabetes in PASSI, we use the perceived health variable detected in PASSI. The perceived health variable is an ordinal categorical variable that takes values between 1 and 5, where 1 means excellent self-reported health and 5 means very bad self-reported health. Thus, if perceived health is between 1 and 3, we use the minimum value of the weights. If it equals 4, we use the average of the disability weights selected, and if it equals 5, we use the maximum value of these weights. Figure 1 summarizes this process.

Finally, if an individual declares more than one disease, the maximum value between the disability weights selected is considered. In this way, we consider the most impactful disease within the life of the individual who declared more than one disease.

DRAFT

## 2.2 Model

This section proposes a naive utilization of the comorbidity index defined in Sect. 2.1. First of all, the response variable we want to analyze has a particular distribution. The comorbidity index described in Subsect. 2.1, in fact, is a "semi-continuous" multimodal skewed nonnegative variable with several zero values. We decided then to use nonparametric methods, such as machine learning methods, that can handle any functional form of the analyzed response variable. Here, we report the results coming from the random forest approach (2). However, other methods can be used and compared (e.g., Tweedie regression, support vector machine), but it is beyond the scope of this paper.

# 3. Results

Table 1 shows the importance of the covariates analyzed, i.e., age, sex, educational level, and economic problems. The importance is calculated as follows: the method permutes the feature values of each variable and computes the out-of-bag error (mean squared error in this case). The importance score, defined by Strobl et al.(14), is then calculated by averaging the difference in out-of-bag error before and after the permutation over all trees. If the prediction error does change consistently, the related variable is defined as important inside the random forest model. The permutation-based importance measures are then scaled to have a maximum equal to 100 and a minimum equal to 0. Finally, this importance score is conditional in the sense of coefficients in regression models considering both the main and interaction

effects of the variable (14).

| | Importance |
|---|---|
| Age | 100.000 |
| Educational level | 10.718 |
| Sex | 6.148 |
| Economic problems | 0 |

Table 1: Variable importance measure from random forest model.

We can note that age is the main variable that impacts the split of the random forest tree, having an importance score equal to 100. In contrast, the economic problems variable has a minimal effect on the results of the model, i.e., the importance score equals 0. This is probably due to the presence of a strong association between the economic and educational level variables.

Figure 2 shows the predicted values of the GBD disability weights across age, analyzing 4 sub-populations characterized by different economic (no economic problem, economic problem) and educational (low, medium-high) status levels divided by sex. As expected, We can note how the disability weights increase as age increases. We can note a great difference between males and females, particularly in the elderly ages, if the sub-population characterized by a high educational level (i.e., left and right top plots of Figure 2). More interestingly, in older ages, the comorbidity index is lower in the sub-population characterized by high educational level and no economic problems (i.e., left top plot of Figure 2) than in sub-population having economic problems and low educational level (i.e., right bottom plot of Figure 2). For example, focusing on the elderly population (i.e., age equals 69), the comorbidity index equals 0.109 for the females and 0.072 for the males if the sub-population with a high educational level and no economic problem is analyzed. In contrast, it equals 0.13 for the females and 0.113 for the males if the sub-population with a low educational level and no economic problem is considered. In addition, we can note how the difference in terms of comorbidity index is substantial also in adult ages, not only in elderly ages if the sub-population characterized by economic problems and low educational level is analyzed (e.g., the index equals 0.066 for females and 0.043 for males at age 49). According to the literature, these statements support the presence of a difference in terms of comorbidity in socioeconomic class (8; 3).

Finally, the predicted values reported in Figure 2 are in line with the analysis of the Years Lived with Disability (YLD) index coming from the GBD (http://ihmeuw.org/5z0s, http://ihmeuw.org/5z0t) if only the division by age and sex variables is considered which are the only one available from the GBD project. Therefore, thanks to the proposed new comorbidity index, we can also analyze the level of comorbidity of the Italian population characterized by different educational levels and economic status, which the GBD project does not detect.

DRAFT

## 4. Discussion

In this paper, we proposed an alternative definition of comorbidity by analyzing the Italian surveillance system PASSI data and the disability weights given by the GBD project. The diseases detected in PASSI were associated with the disability weights of the GBD by several steps: a text mining one to extract the related GBD weights and the utilization of the perceived health variable reported in PASSI to filter the extracted GBD weights.

We finally proposed a naive analysis of this comorbidity index considering sub-populations characterized by sex, age, and different levels of education and economic situation of the subjects. Therefore, the comorbidity definition proposed permits exploring two novel analyses: the level of comorbidity from the PASSI data and the socio-economic population structure of the GBD data.

However, we made some assumptions in order to create this comorbidity index. These assumptions must be explored in more detail in the future. For example, the process of selecting the weights was done automatically by text mining. It would be appropriate in future works to select these weights with experts.

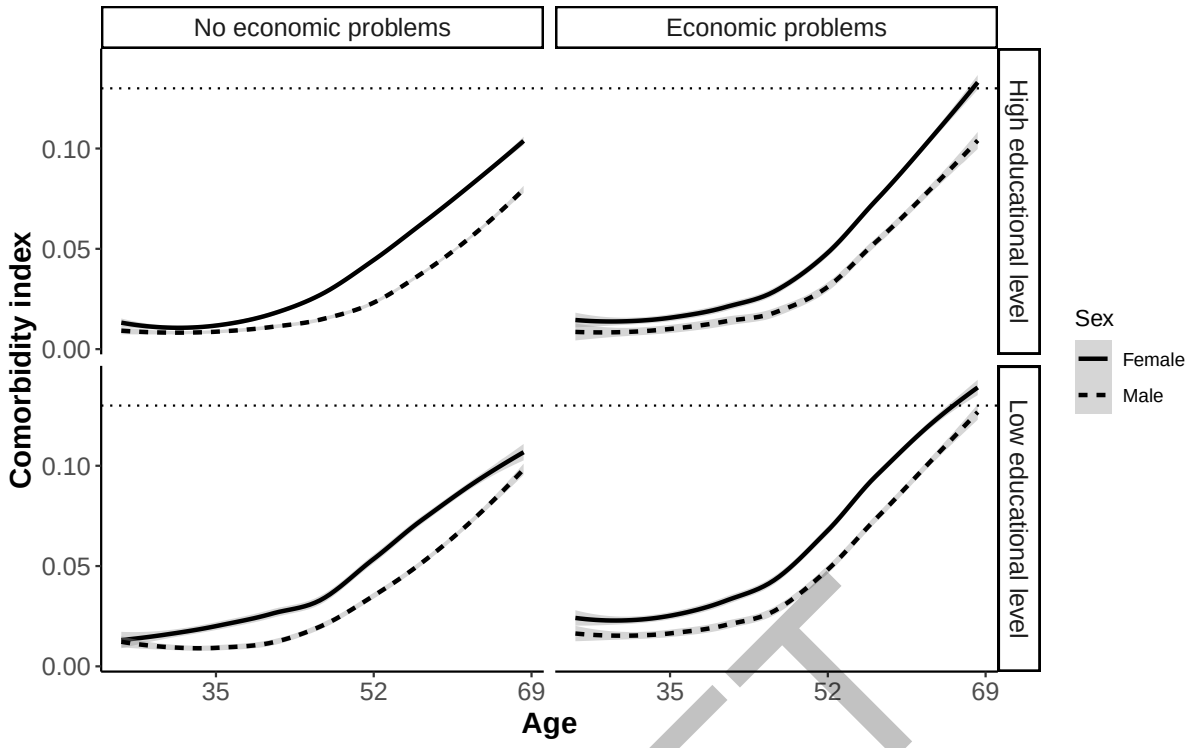



Figure 2: Predictions of the comorbidity index across age considering sub-populations characterized by different levels of education (low-high) and economic status (no economic problems-economic problems). The grey area represents the prediction interval at level 0.95.

In fact, some of the weights analyzed consider combinations of diseases that should be analyzed in more detail. In addition, we examined the maximum disability weights across declared non-communicable diseases while a different propriety could be chosen (i.e., the sum of disease weights, the mean, the minimum, or a novel combination). Finally, analyzing the comorbidity trend would be interesting by comparing different years of the PASSI survey as done in Pastore et al.'s work(11).

**Acknowledgments** Angela Andreella gratefully acknowledges funding from the grant PON 2014-2020/DM 1062 of the Ca' Foscari University of Venice, Italy.

DRAFT

## References

[1] Baldissera, S., Campostrini, S., Binkin, N., Minardi, V., Minelli, G., Ferrante, G., Salmaso, S.: Features and initial assessment of the Italian behavioral risk factor surveillance system (PASSI), 2007-2008. Prev. Chronic Dis., **8**(1) (2011).

[2] Breiman, L.: Random forests. Mach. Learn., **45**(1):5–32 (2001).

[3] Campostrini, S., McQueen, D. V.: Inequalities: the "gap" remains; can surveillance aid in closing the gap? Int. J. Public Health, **56**:219–220 (2014).

[4] Charalampous, P., Polinder, S., Wothge, J., von der Lippe, E., Haagsma, J. A.: A systematic literature review of disability weights measurement studies: evolution of methodological choices. Arch. Public Health, **80**(1):1–16 (2022).

[5] Hernandez, J. B., Kim, P.: Epidemiology morbidity and mortality. In: StatPearls. StatPearls Publishing, Treasure Island (FL). (2022).

[6] Jones, D.: Cambridge English pronouncing dictionary. Cambridge University Press. (2011).

[7] Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., Murray, C. J.: Global and regional

burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet, **367**(9524):1747–1757 (2006).

[8] Minardi, V., Campostrini, S., Carrozzi, G., Minelli, G., Salmaso, S.: Social determinants effects from the Italian risk factor surveillance system PASSI. Int. J. Public Health, **56**:359–366 (2011).

[9] Monasta, L., Abbafati, C., Logroscino, G., Remuzzi, G., Perico, N., Bikbov, B., Tamburlini, G., Beghi, E., Traini, E., Redford, S. B., Ariani, F., Borzì, A. M., Bosetti, C., Carreras, G., Caso, V., Castelpietra, G., Cirillo, M., Conti, S., Cortesi, P. A., Damiani, G., D' Angionella, L. S., Fanzo, J., Fornari, C., Gallus, S., Giussani, G., Gorini, G., Grosso, G., Guido, D., La Vecchia C., Lauriola, P., Leonardi, M., Levi, M., Madotto, F., Mondello, S., Naldi, L., Olgiati S., Palladino, E., Raggi, A., Rubino, S., Santalucia, P., Vacante, M., Vidale, S., Violante S. F., Naghavi, M., Ronfani, L.: Italy's health performance, 1990–2017: findings from the global burden of disease study 2017. Lancet Public Health, **4**(12):e645–e657 (2019).

[10] Murray, C. J., Lopez, A. D.: Evidence-based health policy?lessons from the global burden of disease study. Science, **274**(5288):740–743 (1996).

[11] Pastore, A., Tonellato, S. F., Aliverti, E., Campostrini, S.: When does morbidity start? an analysis of changes in morbidity between 2013 and 2019 in Italy. Stat. Methods Appt., pages 1–15 (2022).

[12] Prados-Torres, A., Calderón-Larrañaga, A., Hancco-Saavedra, J., Poblador-Plou, B., van den Akker, M.: Multimorbidity patterns: a systematic review. J. Clin. Epidemiol., **67**(3):254–266 (2014).

[13] Salomon, J. A., Haagsma, J. A., Davis, A., de Noordhout, C. M., Polinder, S., Havelaar, A. H., Cassini, A., Devleesschauwer, B., Kretzschmar, M., Speybroeck, N., Murray, C. J. L., Vos, T.: Disability weights for the global burden of disease 2013 study. Lancet Global Health, **3**(11):e712–e723 (2015).

[14] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. *BMC Bioinform.*, **9**(1):1–11 (2008).

[15] Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., Roland, M.: Defining comorbidity: implications for understanding health and health services. Ann. Fam. Med., **7**(4):357–363 (2009).

DRAFT

# An Age-Period-Cohort model of gender gap in youth mortality

Giacomo Lanfiuti B aldi[a] and Andrea Nigri[b]

[a]Department of Statistics, Sapienza University of Rome, Rome, Italy

giacomo.lanfiutibaldi@uniroma1.it

[b]Department of Economics, Management and Territory, University of Foggia, Foggia, Italy;
andrea.nigri@unifg.it

DRAFT

**Abstract**

In this paper, we propose an Age-Period-Cohort (A-P-C) model leveraging Skew-Normal distribution, aiming at modelling the gender gap in youth mortality.
We noticed that gender differences in youth mortality are largest around the age of 20, but these differences are not symmetrical with respect to the peak. Following this evidence, we perform the APC analysis using a mixed-effects model in which the response variable follows the Skew-Normal distribution. We adopt a sex ratio approach, using the ratio of age-specific mortality rates of men and women as the response variable. We test and compare different models in which we consider age as a fixed effect and different combinations of period and cohort as fixed or random effects.
Our research focuses on the population under age 45 in the United States between 1960 and 2020. In the results, we see a decrease in gender differences in youth mortality over the last 25 years, and it is evident that cohorts born between the mid-1940s and the 1960s had the highest excess in male mortality at younger ages.

***Keywords:*** Mortality modelling, Skew-Normal, Age-Period-Cohort Model

## 1. Introduction

Gender differences in mortality are increasingly discussed and studied in the social and demographic fields. Modelling and analysing gender differences in mortality can tell us a lot about the social context of a population or a country [9]. Thus, a better understanding of sex differences in mortality can guide public health efforts to reduce overall mortality rates and promote greater equity in health outcomes.
Gender differences are not constant at all ages and are driven by different causes of death. Many studies focus on gender differences in mortality at adult ages, but there is less literature on differences at younger ages. At the same time, gender differences in mortality are greater at younger ages [7] and this is mainly due to the different behaviour of the two groups. Men in particular run a much higher risk of accidental death around the age of 20 [8]. In addition to age, mortality differences change over time because the behaviour that generates them changes over the years [2].
We aim to study how the gender gap in mortality at younger ages (under 45) varies concerning individual age groups and over time. We want to define and interpret the effects of different ages and social, cultural and behavioural changes in society [14]. To do this, we work in an Age-Period-Cohort (APC) framework

leveraging mixed model effects based on a Skew-Normal distribution. The Skew-Normal distribution is not widely used in the APC framework.
In particular, we are interested in studying in the United States (US), where more than the 30% of deaths at young ages are due to external causes (unintentional injuries) [3] and the issue of road accidents, and violent and risky behaviour among young people is often at the centre of public debate.

## 2. Data and Measure

We use a sex-ratio approach to study gender differences: we analyse the ratio of age-specific mortality rates between males ($m^M_{x,t}$) and females ($m^F_{x,t}$) over time.

$$SR_{x,t} = \frac{m^M_{x,t}}{m^F_{x,t}} \tag{1}$$

This measure 1 is useful for several reasons: it allows us to use a single variable to study the two sexes, it is less sensitive to the general level of mortality than the absolute difference in deaths [2], and finally, it has a well-defined and known shape [8] (Fig.1). We use age-specific mortality rates per sex and single year of age (0 - 100) from the Human Mortality Database [12] of the United States in the study period.

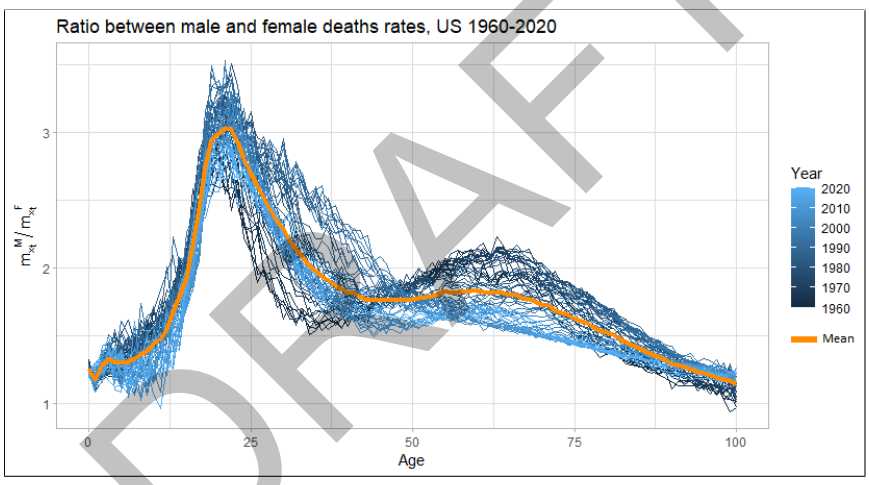



Figure 1: Sex-Ratio of the Age-Specific Mortality Rates in US between 1960 and 2020. Data source: HMD

Generally, this sex-ratio over the ages is characterized by a Peak and a Hump. The peak, which is the highest and most concentrated, coincides with youthful ages and is generally attributed to the highest male mortality due to riskier behaviours [8]. The hump (Fig.2) corresponds to the adult ages and it was primarily caused by excess male mortality from cancer [2]. According to [8] we set the threshold age (between the peak and the hump) at 45 ages and we will focus only on the peak, in order to study the gender gap in mortality at young ages.

The differences between the two sex in infant mortality are very low in all the considered periods, the differences start to increase in the years of adolescence. Male mortality at the peak comes to more than 3 times that of females in most years and there is no trend in the shift of the peak age over the years.
We have noticed that the sex-ratio at younger ages has a Gaussian shape around the peak, but in most of the years, it is not symmetrical around the peak.

DRAFT

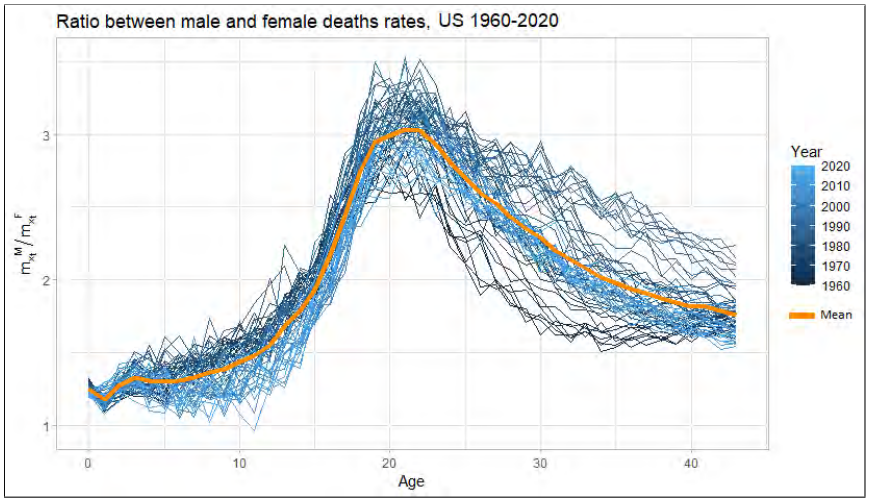



Figure 2: Sex-Ratio of the Age-Specific Mortality Rates in US between 1960 and 2020 for young (under 45) population. Data source: HMD

DRAFT

## 3. Model

In order to study the age, period and cohort effects we suggest leveraging a mixed model effects model based on a *Skew-Normal* distribution. Here is the density function of a Skew-Normal distribution as presented by Azzalini [1]:

$$f(y;\boldsymbol{\theta}) = f(y;\mu,\sigma^2,\delta) = \frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{\sqrt{y-\mu}}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{\sqrt{y-\mu}}{\sqrt{\sigma^2+\delta^2}}\right) \tag{2}$$

Following the framework proposed by Klein et al.[6], we can generally set up the relationship between distribution parameters and the elements of the linear predictor as:

$$g(\varphi) = \sum_{j=1}^{J} f_j(\boldsymbol{\nu}),$$

Where $f$ may comprise various forms, defined on basis of the covariate structure, as follows: a linear function $f_j(\nu) = \mathbf{X}\beta_j$, that represents the fixed effects; and a random effects part $\gamma_g$, in which $g$ is a cluster variable that groups the observations. This can be written in the generic matrix notation $\mathbf{Z}\gamma$, where $\mathbf{Z}$ is the random effects design matrix and $\gamma$ is a correspondent coefficient.

Specifically, let's consider $y^{\mathrm{T}} = (y_1, y_2, \ldots, y_n)$ as the vector of the response variable and $f(y;\boldsymbol{\varphi})$, a density function with $k$ parameters $\boldsymbol{\varphi}^{\mathrm{T}} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ modelled by using linear additive models. We assume that observations $y_i$ are independent conditional on $\boldsymbol{\varphi}$, with density function $f(y_i;\varphi_i)$, where ${\varphi_i}^{\mathrm{T}}$ is a vector of $k$ parameters related to explanatory variables and random effects. Let $g(\cdot)$ be a known monotonic link function relating $\varphi_k$ to explanatory variables and random effects through an additive model given by:

$$g(\varphi) = \eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

where: $\beta^{\mathrm{T}} = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'k})$ is a parameter vector of length $J'$, $\mathbf{X}$ is the design matrix of fixed effects of order $n \times J'$, $\mathbf{Z}$ is a $n \times q_j$ random effects design matrix and $\gamma_j$ is a correspondent $q_j$-dimensional coefficient. In our study, for the skewed normal family distribution: $\varphi = \mu$ and $g(.)$ is the identity function. So, we have the following model:

$$\mu = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

Thereby, the components of the model are: $\mathbf{y}$, is the response vector of length $n$; $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ are the design matrices for the fixed effects and $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_p^{\mathrm{T}})$ are the linear parameters. The

random effects part is given by the design matrices $\mathbf{Z} = (\mathbf{Z}_{11}, \mathbf{Z}_{21}, \ldots, \mathbf{Z}_{J_11})$ and the coefficients: $\gamma^{\mathrm{T}} = \left(\gamma_{11}^{\mathrm{T}}, \gamma_{21}^{\mathrm{T}}, \ldots, \gamma_{\mathcal{L}_11}^{\mathrm{T}}, \ldots, \gamma_{p}^{\mathrm{T}}, \gamma_{2p}^{\mathrm{T}}, \ldots, \gamma_{L_pp}^{\mathrm{T}}\right)$.

## 4. Model specification as an Age-Period-Cohort model

Here, we provide the model to an APC Skew-Normal mixed model to be applied to the study of the gender gap while including the Age-Period-Cohort (A-P-C) structure as predictor. We aim to provide the broadest possible overview of solutions, taking into account model specifications with classical constraints, and using random effects.

Let $\mathcal{A} = \{a_0, a_1, \ldots, a_\omega\}$, $\mathcal{P} = \{p_0, p_1, \ldots, p_n\}$ and $\mathcal{C} = \{c_0, c_1, \ldots, c_m\}$ be the set of age, year and cohort categories, respectively. The APC model describes the gender ration of death rates at age $a \in \mathcal{A}$, time $p \in \mathcal{P}$, and cohort $c \in \mathcal{C}$.

**Constrained Model**

We start by introducing the constrained model as the classic way and thus adapting the constraints to our framework in the following formulation:

$$\mu = \beta_{(a)} + \beta_{(p)} + \beta_{(c)} \tag{3}$$

where, $\beta$s are coefficients of fixed effects using categorical coding for Age, Period and Cohort respectively, imposing the following constraints:

$$\sum_{\omega=1}^{\Omega} \beta_{(a\omega)} = \sum_{n=1}^{N} \beta_{(p_n)} = \sum_{m=1}^{M} \beta_{(c_m)} = 0$$

**Random Effects - Period, Cohort**

We, then, introduce the mixed effects model. We treat Age as fixed effects according to Reither et al. [10], who justify the treatment of period and cohort random effects.

$$\mu = \beta_{(a)} + \gamma_p + \gamma_c \tag{4}$$

In this case, $\gamma_p$ and $\gamma_c$ are random effects for the period and cohort level, and $\beta$ provide fixed effects using categorical coding for Age.

**Random Effects - Cohort**

An alternative is to consider only cohorts as random effects:

$$\mu = \beta_{(a)} + \beta_{(p)} + \gamma_c \tag{5}$$

Where, $\gamma_c$ is the random effect for cohort level and $\beta$s fixed effects using categorical coding for Age and Period.

For all tree models, we specify a $\sigma^2 \sim \Gamma(0.01, 0.01)$ and a $\delta \sim \Gamma(0.01, 0.01)$. Furthermore, we use flat priors for random effects coefficients.

## 5. Results

In our analysis, we estimate the models described in the previous section (4) with a Bayesian approach. Samples from the posterior distributions of the parameters and effects were drawn by using Hamiltonian Monte Carlo sampling and specifically using the `stan` software package [5].

DRAFT

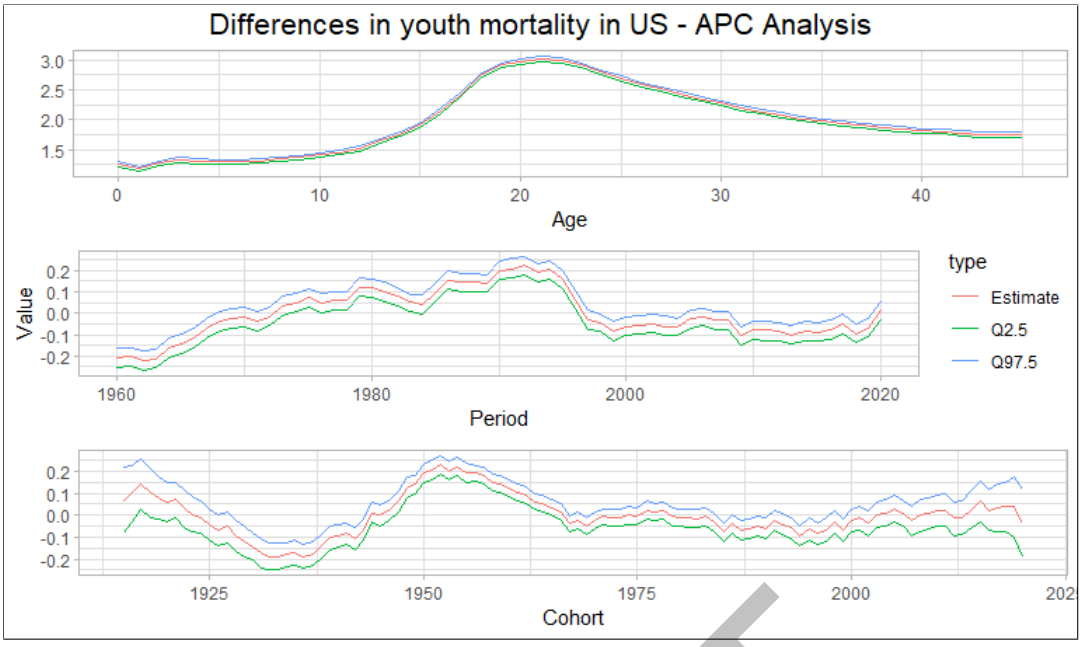



Figure 3: APC model: Age fixed effect, Period and Cohort random effects. Data Source: HMD

Here we report the results (Fig.3) for the model in which we coded age as a categorical variable and treat it as a fixed effect; period and cohort are random effects (Eq. 4).

The age parameters provide the structure of the gender gap in youth mortality, which can be found on average in all years of the period and for all cohorts. The value of the estimated parameters for age increases with adolescence. Researchers usually do not look at biological factors to explain the excess mortality among men at these ages, but rather at individual and social reasons [4] [11]. These reasons become even bigger and more important in the peak years of gender differences (21-22 years): car accidents, suicides and violence are by far the most important causes.

From the two graphs of period and cohort parameters, we can suppose to observe the impact that AIDS has had on US society. AIDS-related mortality is higher for men than for women [13]. Thus, as the pandemic in the US diminished during the 1990s, mortality differences decreased. Cohorts born between the 1940s and 1970s are those on which the AIDS pandemic had the greatest impact.

DRAFT

## 6. Discussion

We observed the gender gap in youth mortality in the US between 1960 and 2020. For this aim, we used a sex-ratio approach in the Age-Period-Cohort framework, leveraging a mixed-effects model based on a Skew-Normal distribution. We tested different models considering age as a fixed effect and various combinations of period and cohort as fixed or random effects.

The parameter estimates were performed adopting a Bayesian approach and using `stan` software.

The innovation of this work is to implement an Age-Period-Cohort analysis, assuming that the target variable has an asymmetric distribution: the Skew-Normal distribution is not widely used in the APC framework, which usually is based on the normal distribution in the demographic field.

Observing sex differences in mortality and how these vary across ages and over time is useful for understanding society and the behaviours that determine them. Moreover, the knowledge of the mortality dynamics and of the differences between the sexes can be an excellent tool in the hands of policymakers. Preliminary results show that over the past 25 years in the US, we have observed a decrease in sex differences in youth mortality. Younger cohorts are benefiting from societal changes and the attention that the topic of youth mortality is receiving in the public debate.

## References


[1] Adelchi Azzalini and A Dalla Valle. "The multivariate skew-normal distribution". In: *Biometrika* 83.4 (1996), pp. 715–726.

[2] Marie-Pier Bergeron-Boucher et al. "Modeling and forecasting sex differences in mortality: a sex-ratio approach". In: *Genus* 74 (2018), pp. 1–28.

[3] M. Heron. *Deaths: Leading causes for 2019*. Vol. 70. National Vital Statistics Reports 9. Hyattsville, MD: National Center for Health Statistics, 2021. DOI: 10.15620/cdc:107021.

[4] Patrick Heuveline and Gail B Slap. "Adolescent and young adult mortality by cause: age, gender, and country, 1955 to 1994". In: *Journal of Adolescent Health* 30.1 (2002), pp. 29–34.

[5] Jason Hilton et al. "Projecting UK mortality using Bayesian generalised additive models". In: *arXiv preprint arXiv:1802.03242* (2018).

[6] Nadja Klein et al. "Bayesian structured additive distributional regression for multivariate responses". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.4 (2015), pp. 569–591.

[7] Hsiang-Ching Kung et al. "Deaths: final data for 2005". In: (2008).

[8] France Meslé. "Life expectancy: a female advantage under threat". In: *Population and Societies* 402.4 (2004), pp. 1–4.

[9] Constance A Nathanson. "Sex differences in mortality". In: *Annual review of sociology* 10.1 (1984), pp. 191–213.

[10] Eric N Reither et al. "Should age-period-cohort studies return to the methodologies of the 1970s?" In: *Social science & medicine* 128 (2015), pp. 356–365.

[11] Susan B Sorenson. "Gender disparities in injury mortality: consistent, persistent, and larger than you'd think". In: *American journal of public health* 101.S1 (2011), S353–S358.

[12] Berkeley (USA) University of California and Max Planck Institute for Demographic Research (Germany). *Human Mortality Database*. http://www.mortality.org. 2021.

[13] I. Waldron, ed. *Contributions of changing gender differences in behavior and social roles to changing gender differences in mortality*. SAGE Publications, Inc., 1995. DOI: 10.4135/9781452243757.

[14] Yang Yang et al. "The intrinsic estimator for age-period-cohort analysis: what it is and how to use it". In: *American Journal of Sociology* 113.6 (2008), pp. 1697–1736.


DRAFT

# Kinlessness in adult and old age across Europe

Marta Pittavino[a], Bruno Arpino[a], and Elena Pirani[a]

[a]Department of Statistics, Computer Science, Applications "Giuseppe Parenti", University of Florence
marta.pittavino@unifi.it, bruno.arpino@unifi.it, elena.pirani@unifi.it

**Abstract**

In this work we estimate the prevalence of older adults aged 50 and more without close kin in several European countries. Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), we examine the prevalence of lacking different types and combinations of living kin, considering how kinlessness vary over time and at different ages. In 2019-2020, the prevalence of adults aged 50 and above who lacked a partner/spouse ranged between 22% and 47% across countries, while the prevalence of childless individuals between 4% and 14%. We detected a large variation of kinlessness across countries and age groups. This is of interest to policy makers because kinlessness is associated with poorer economic and health conditions, living alone, and unmeet care needs. Aging research should address the implications of kinlessness for public health, social isolation, and the demand for institutional care.

***Keywords:*** SHARE data, Family structures, Population aging, Social support

DRAFT

## 1. Introduction

Kinlesness is the lack of close kin. Different definitions have been used in previous studies that vary because of the (combination of) specific kinship ties considered. When studying older adults, it is particularly relevant to focus on the absence of a partner/spouse and children (5) because they are the main providers of care and emotional support, as well as the main agents of social control. Recently, there has been a growing interest on kinlessness. Several studies have focused on the estimation of the prevalence and demograhic characteristics of individuals who lack a specific type of kin (e.g., grandchildren) (3). Other studies, have estimating the prevalence of kinlessness, i.e. lack of more than one kin, especially among older adults (5; 6; 9; 10) and its consequences on health, loneliness, care needs, etc. For example, research has shown that kinless individuals tend to report worse wellbeing and health conditions (although this varies considerably across countries and type of health outcome (1; 4; 7) and show higher likelihood of engaging in unhealthy behaviors (2).

In this work we document the size of the population of older adults (aged 50 and more) without close kin in several European countries. Only (9) provided estimates of kinlessness for several European countries in 2015. We contribute to this literature by providing more recent estimates and by showing the variability in kinlessness across age groups and over time.

## 2. Data and Methods

For this analysis we rely on the data from the Survey of Health, Ageing and Retirement in Europe (SHARE), especially wave 8 carried out in 2019/2020. In each wave, SHARE data cover several key

areas of life - health, socio-economic status, social and family networks - of more than 60,000 individuals aged 50 or over, enabling us to detect kinship ties of individuals. We considered all the 25 countries included in the survey. Additional analyses will use wave 2, 4 and 6 of SHARE to examine changes over time. We will not use all waves because in wave 1 fewer countries than in the other waves participated, while the other waves collected only (wave 3) or mostly (wave 7) retrospective information.

In terms of methods, this study is purely descriptive. We present estimated prevalences (with 95% confidence intervals (CI)). We use cross-sectional calibrated weights that account for sampling design and attrition.

## 3. Results

Tables 1 and 2 report the (weighted) estimates (and 95% confidence intervals (CI)) of the prevalence of individuals aged 50 and over without a given type of kinship tie, by country. On average, about one third of 50+ individuals lack a living partner/spouse, ranging from the minimum of Spain and Finland (respectively 22 and 25%) to a maximum of 47% for Luxemburg (values around 40% are registered also for Poland and Latvia). The absence of children is depicted for one older individual out of ten on average. The lowest values appear for Eastern European countries (e.g., Hungary, Romania, Lithuania around 4-7%) and Northern ones (e.g., Denmark and Sweden with 7-8%), and the highest values of the prevalence (13-14%) are found for countries like Belgium, Croatia, Czech Republic and Spain. Overall, a large variability is found across this group of countries. Italy occupies an intermediate position: about 29% of individuals aged 50 and over is without a partner and just under 12% lacks children.

An even larger variation has been found with respect lacking grandchildren. About 33% of older Europeans declare to have no grandchildren, with the lowest values (15-17%) found for Greece and Hungary. In this case, Italy is - somewhat unsurprisingly - the European country with the largest share of older individuals without grandchildren, almost 53%, but values around 50% are found also for France, Finland and Croatia.

The absence of living parents is clearly higher due to the age group considered. 3 out of 4 respondents aged 50 and over no longer have parents (62-63% in Italy and Romania, 80% in Poland and Sweden). As expected, due to the higher male mortality, the prevalence of individuals without a living father is higher to that of individuals without a living mother (in some countries the difference reaches 15-20 percentage points).

Finally, also the horizontal kinships - existence of living siblings - proves a large variability across countries. Less than 10% of individuals aged 50 and over has no brothers nor sisters in Croatia, Romania and Spain, but the prevalence raises to 30 or even 40% in Luxembourg, Switzerland, Latvia and Slovenia. The average stands around 20%.

It is straightforward to imagine that this overall picture, already highly differentiated by country, varies greatly on the basis of the age considered and the kinship types. If the presence of ascendants is going to reduce as individuals get older, that of descendents (grandchildren) can somewhat compensate for it, especially in those countries where family formation is postponed. In addition, the absence of one type of kinship may be replaced by the presence of other relatives. To get more insights, we considered various combinations of lack of kinship types, differentiating by the level of absence of kinshp ties (k1: no partner AND no children; k2: no partner AND no children AND no sibling; k3: no partner AND no children AND no siblings AND no grandchildren; k4: no partner AND no children AND no siblings AND no grandchildren AND no parent).

Figure 1 displays the estimate of the four considered kinlessness types for 3 age groups, considering four countries which represent different models of mortality and fertility. First, it is worthwhile noting that the lack of various types of kinship (k2-k4) remains low in all the countries, especially for younger older individuals (50-64). Nevertheless, in France a large variation across age groups is depicted: the lack of kinships progressively increases by age, and reaches a prevalence of almost 10% for the oldest individuals. On the contrary, in Czech Republic the absence of various types of kinships is rather similar regardless the age class considered, from 0 to less than 5%. Somewhat surprisingly, the picture in Italy and Denmark - two countries differing from a demographic and socio-cultural point of view - appears

DRAFT

Table 1: Prevalence and 95% confidence intervals [CI] (below) of adults aged 50 or over, without partner, child, grandchild and sibling, from the SHARE study 2019-2020

| Countries | no partner | no child | no grandchild | no brother | no sister | no sibling |
|---|---|---|---|---|---|---|
| Austria | 0.281 | 0.101 | 0.322 | 0.393 | 0.381 | 0.180 |
| | [0.243,0.319] | [0.073,0.130] | [0.271,0.373] | [0.346,0.439] | [0.334,0.428] | [0.151,0.209] |
| Belgium | 0.320 | 0.131 | 0.386 | 0.461 | 0.438 | 0.210 |
| | [0.297,0.342] | [0.114,0.149] | [0.361,0.410] | [0.437,0.485] | [0.414,0.461] | [0.192,0.228] |
| Bulgaria | 0.319 | 0.086 | 0.309 | 0.455 | 0.407 | 0.169 |
| | [0.286,0.351] | [0.063,0.108] | [0.270,0.348] | [0.419,0.490] | [0.373,0.442] | [0.148,0.190] |
| Croatia | 0.284 | 0.149 | 0.482 | 0.239 | 0.284 | 0.080 |
| | [0.231,0.337] | [0.099,0.200] | [0.417,0.547] | [0.202,0.275] | [0.224,0.345] | [0.063,0.097] |
| Cyprus | 0.288 | 0.093 | 0.445 | 0.342 | 0.329 | 0.134 |
| | [0.232,0.345] | [0.064,0.122] | [0.376,0.513] | [0.285,0.399] | [0.273,0.384] | [0.108,0.160] |
| Czech Republic | 0.270 | 0.149 | 0.488 | 0.426 | 0.440 | 0.193 |
| | [0.235,0.304] | [0.115,0.183] | [0.451,0.526] | [0.390,0.463] | [0.403,0.477] | [0.165,0.220] |
| Denmark | 0,320 | 0.088 | 0.310 | 0.376 | 0.346 | 0.152 |
| | [0.294,0.346] | [0.072,0.104] | [0.281,0.340] | [0.349,0.403] | [0.320,0.373] | [0.134,0.170] |
| Estonia | 0.312 | 0.087 | 0.310 | 0.398 | 0.382 | 0.142 |
| | [0.288,0.337] | [0.071,0.103] | [0.283,0.337] | [0.373,0.423] | [0.357,0.406] | [0.125,0.158] |
| Finland | 0.253 | 0.117 | 0.490 | 0.448 | 0.471 | 0.217 |
| | [0.235,0.271] | [0.102,0.131] | [0.469,0.511] | [0.427,0.469] | [0.450.0.493] | [0.201,0.234] |
| France | 0.298 | 0.158 | 0.510 | 0.363 | 0.351 | 0.112 |
| | [0.259,0.336] | [0.126,0.190] | [0.467,0.552] | [0.323,0.404] | [0.311,0.391] | [0.093,0.130] |
| Germany | 0.308 | 0.127 | 0.355 | 0.369 | 0.366 | 0.165 |
| | [0.275,0.340] | [0.102,0.152] | [0.318,0.391] | [0.337,0.402] | [0.333,0.399] | [0.143,0.186] |
| Greece | 0.364 | 0.079 | 0.175 | 0.342 | 0.257 | 0.109 |
| | [0.264,0.465] | [0.017,0.141] | [0.097,0.252] | [0.254,0.431] | [0.204,0.311] | [0.083,0.134] |
| Hungary | 0.288 | 0.038 | 0.150 | 0.484 | 0.496 | 0.234 |
| | [0.243,0.333] | [0.016,0.060] | [0.100,0.200] | [0.431,0.537] | [0.443,0.550] | [0.196,0.272] |
| Israel | 0.379 | 0.102 | 0.254 | 0.418 | 0.350 | 0.161 |
| | [0.352,0.405] | [0.081,0.122] | [0.228,0.280] | [0.393,0.444] | [0.326,0.375] | [0.142,0.179] |
| Italy | 0.291 | 0.117 | 0.529 | 0.404 | 0.393 | 0.174 |
| | [0.241,0.341] | [0.082,0.151] | [0.478,0.581] | [0.352,0.456] | [0.343,0.443] | [0.137,0.211] |
| Latvia | 0.399 | 0.112 | 0.367 | 0.668 | 0.624 | 0.429 |
| | [0.322,0.476] | [0.062,0.163] | [0.275,0.458] | [0.585,0.751] | [0.549,0.699] | [0.346,0.512] |
| Lithuania | 0.307 | 0.066 | 0.273 | 0.446 | 0.426 | 0.180 |
| | [0.278,0.336] | [0.047,0.085] | [0.236,0.311] | [0.413,0.478] | [0.393,0.460] | [0.157,0.203] |
| Luxembourg | 0.468 | 0.101 | 0.237 | 0.612 | 0.536 | 0.319 |
| | [0.443,0.492] | [0.084,0.118] | [0.212,0.261] | [0.588,0.636] | [0.511,0.560] | [0.297,0.342] |
| Malta | 0.344 | 0.103 | 0.292 | 0.521 | 0.456 | 0.229 |
| | [0.310,0.378] | [0.080,0.126] | [0.259,0.326] | [0.487,0.555] | [0.422,0.490] | [0.201,0.256] |
| Netherlands | 0.353 | 0.078 | 0.272 | 0.497 | 0.416 | 0.203 |
| | [0.326,0.381] | [0.061,0.095] | [0.244,0.300] | [0.468,0.526] | [0.388,0.445] | [0.181,0.225] |
| Poland | 0.425 | 0.074 | 0.256 | 0.592 | 0.574 | 0.354 |
| | [0.387,0.463] | [0.052,0.095] | [0.219,0.292] | [0.554,0.629] | [0.537,0.612] | [0.318,0.390] |
| Romania | 0.314 | 0.051 | 0.260 | 0.239 | 0.240 | 0.074 |
| | [0.256,0.372] | [0.025,0.077] | [0.201,0.319] | [0.190,0.288] | [0.186,0.294] | [0.041,0.107] |
| Slovakia | 0.383 | 0.132 | 0.412 | 0.401 | 0.357 | 0.154 |
| | [0.335,0.432] | [0.098,0.167] | [0.366,0.459] | [0.354,0.448] | [0.310,0.404] | [0.114,0.194] |
| Slovenia | 0.379 | 0.088 | 0.281 | 0.636 | 0.581 | 0.413 |
| | [0.342,0.416] | [0.067,0.108] | [0.245,0.317] | [0.598,0.674] | [0.543,0.619] | [0.375,0.450] |
| Spain | 0.224 | 0.155 | 0.363 | 0.193 | 0.180 | 0.043 |
| | [0.187,0.261] | [0.124,0.186] | [0.320,0.406] | [0.161,0.225] | [0.147,0.214] | [0.025,0.061] |
| Sweden | 0.327 | 0.073 | 0.237 | 0.441 | 0.420 | 0.207 |
| | [0.293,0.360] | [0.054,0.092] | [0.206,0.268] | [0.406,0.475] | [0.386,0.455] | [0.178,0.236] |
| Switzerland | 0.359 | 0.120 | 0.318 | 0.526 | 0.524 | 0.328 |
| | [0.323,0.395] | [0.096,0.144] | [0.286,0.350] | [0.492,0.560] | [0.490,0.558] | [0.295,0.361] |

DRAFT

Table 2: Prevalence and 95% confidence intervals [CI] (below) of adults aged 50 or over, without father, mother and parent, from the SHARE study 2019-2020

| Countries | no father | no mother | no parent |
|---|---|---|---|
| Austria | 0.890 | 0.754 | 0.710 |
| | [0.847,0.932] | [0.705,0.803] | [0.658,0.762] |
| Belgium | 0.893 | 0.756 | 0.720 |
| | [0.875,0.911] | [0.732,0.779] | [0.696,0.744] |
| Bulgaria | 0.882 | 0.761 | 0.727 |
| | [0.851,0.914] | [0.723,0.799] | [0.689,0.766] |
| Croatia | 0.921 | 0.762 | 0.729 |
| | [0.889,0.953] | [0.703,0.821] | [0.669,0.790] |
| Cyprus | 0.896 | 0.772 | 0.713 |
| | [0.843,0.950] | [0.703,0.840] | [0.639,0.787] |
| Czech Republic | 0.887 | 0.717 | 0.690 |
| | [0.857,0.916] | [0.676,0.759] | [0.648,0.731] |
| Denmark | 0.872 | 0.732 | 0.691 |
| | [0.848,0.895] | [0.704,0.760] | [0.662,0.720] |
| Estonia | 0.876 | 0.767 | 0.717 |
| | [0.854,0.897] | [0.741,0.793] | [0.690,0.744] |
| Finland | 0.895 | 0.775 | 0.753 |
| | [0.879,0.911] | [0.755,0.795] | [0.732,0.774] |
| France | 0.886 | 0.719 | 0.676 |
| | [0.851,0.921] | [0.673,0.765] | [0.630,0.723] |
| Germany | 0.880 | 0.744 | 0.695 |
| | [0.851,0.910] | [0.706,0.782] | [0.656,0.734] |
| Greece | 0.926 | 0.732 | 0.709 |
| | [0.874,0.978] | [0.628,0.836] | [0.606,0.811] |
| Hungary | 0.902 | 0.838 | 0.777 |
| | [0.853,0.950] | [0.795,0.880] | [0.721,0.833] |
| Israel | 0.918 | 0.776 | 0.746 |
| | [0.899,0.937] | [0.751,0.800] | [0.720,0.771] |
| Italy | 0.825 | 0.705 | 0.631 |
| | [0.768,0.881] | [0.650,0.760] | [0.573,0.689] |
| Latvia | 0.971 | 0.793 | 0.787 |
| | [0.948,0.993] | [0.703,0.883] | [0.697,0.877] |
| Lithuania | 0.909 | 0.773 | 0.746 |
| | [0.883,0.936] | [0.736,0.810] | [0.709,0.782] |
| Luxembourg | 0.952 | 0.814 | 0.794 |
| | [0.938,0.966] | [0.791,0.837] | [0.770,0.818] |
| Malta | 0.947 | 0.808 | 0.790 |
| | [0.929,0.965] | [0.778,0.838] | [0.759,0.821] |
| Netherlands | 0.953 | 0.811 | 0.788 |
| | [0.938,0.968] | [0.786,0.836] | [0.762,0.815] |
| Poland | 0.938 | 0.826 | 0.806 |
| | [0.918,0.959] | [0.793,0.859] | [0.772,0.840] |
| Romania | 0.819 | 0.673 | 0.619 |
| | [0.764,0.874] | [0.608,0.738] | [0.555,0.684] |
| Slovakia | 0.880 | 0.756 | 0.710 |
| | [0.845,0.915] | [0.712,0.801] | [0.664,0.756] |
| Slovenia | 0.959 | 0.804 | 0.791 |
| | [0.941,0.977] | [0.769,0.838] | [0.755,0.826] |
| Spain | 0.900 | 0.776 | 0.750 |
| | [0.872,0.927] | [0.737,0.816] | [0.710,0.789] |
| Sweden | 0.936 | 0.840 | 0.806 |
| | [0.917,0.954] | [0.813,0.868] | [0.777,0.836] |
| Switzerland | 0.908 | 0.792 | 0.769 |
| | [0.889,0.926] | [0.765,0.819] | [0.741,0.797] |

DRAFT

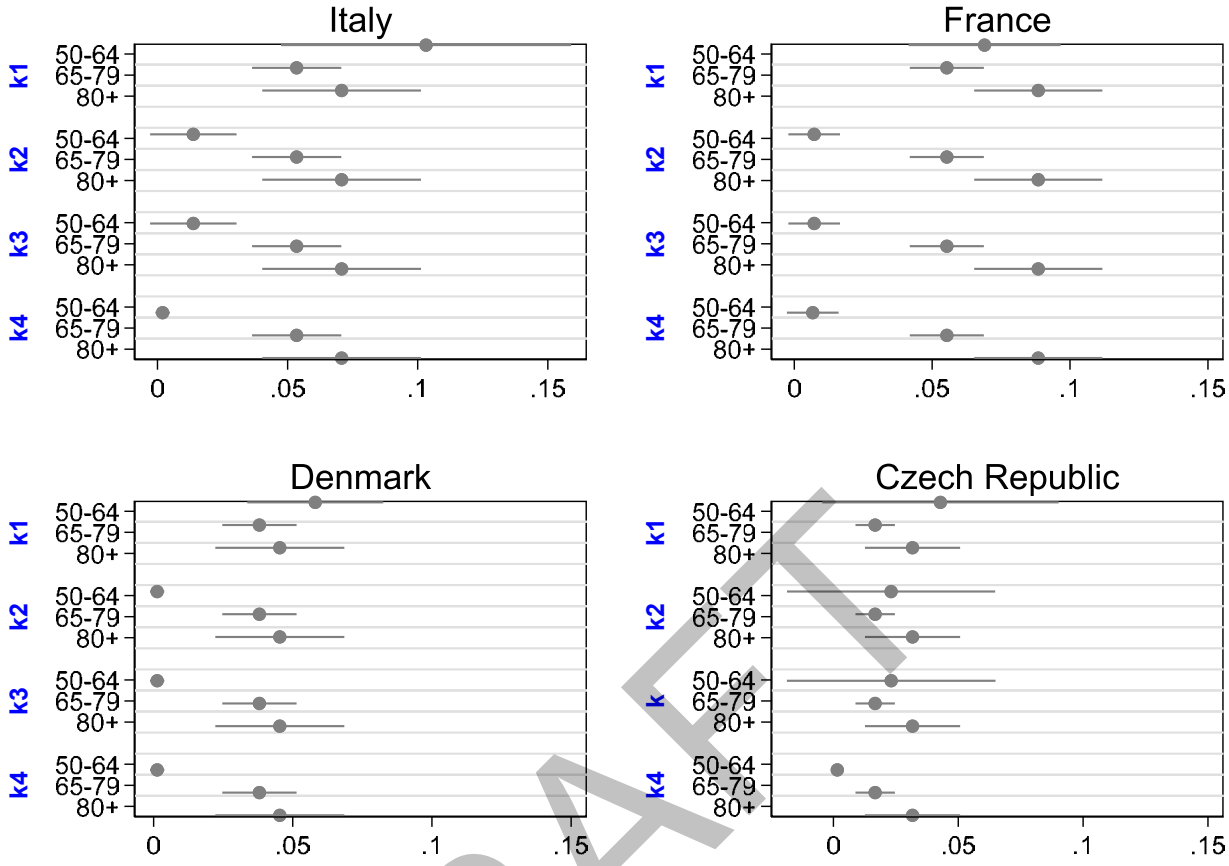



Figure 1: Estimate of kinlessness: $k1, k2, k3$ and $k4$, by 3 age groups (50-64, 65-79, 80+) considering countries which represent different models of mortality and fertility. $k1$ represents no partner AND no children; $k2$ indicates no partner AND no children AND no sibling; $k3$ means no partner AND no children AND no siblings AND no grandchildren; $k4$ represents no partner AND no children AND no siblings AND no grandchildren AND no parent

DRAFT

more similar than foreseen. For instance, the percentage of those who lack both living partner and children (k2) equals to 6-10% for younger respondents, but the lack of also other relatives is extremely reduced less then 2%. For 65-79 and 80+ individuals, the prevalence slightly increases (around 4-7%), but without showing a significant deterioration for the oldest group.

## 4. Discussion

Our analyses show that the prevalence of kinlessness vary considerably across European countries as the result of different socio-demographic dynamics in the past. Due to the recent and current (decreasing) trends in fertility and (increasing) trends in longevity across European countries, it is not difficult to envisage a progressive shrinking in kinship ties. A deeper investigation of the future progression in kinlessness trends is needed.

In the next steps of the analyses we will also provide estimates of kinlessness over a period of about 12 years (2006/7-2019/20). Preliminary estimates show a considerable increase in kinlessness among the majority of the investigated countries.

## 5. Acknowledgements

This paper uses data from SHARE Wave 8 (DOIs: 10.6103/SHARE.w8ca.800) see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N.211909, SHARE-LEAP: GA N.227822, SHARE M4: GA N.261982, DASISH: GA N.283646) and Horizon 2020 (SHARE-DEV3: GA N.676536, SHARE-COHESION: GA N.870628, SERISS: GA N.654221, SSHOC: GA N.823782, SHARE-COVID19: GA N.101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

This publication was produced with the co-funding European Union - Next Generation EU, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-IT, CUP: B83C22004800006.

## References

[1] Albertini M. and Arpino B. Childlessness, parenthood and subjective wellbeing: The relevance of conceptualizing parenthood and childlessness as a continuum. SocArXiv. https://osf.io/preprints/socarxiv/xtfq6/. (2018).

[2] Arpino, B., Bordone, V., and Di Gessa, G. COVID-19 precautionary behaviors and vaccine acceptance among older individuals: The role of close kin. Forthcoming in Proceedings of the National Academy of Sciences (PNAS). (2023).

[3] Arpino B., Gumà, J. and Julià A. Family histories and the demography of grandparenthood. Demographic Research, **39(42)**, 1105–1150 (2018).

[4] Arpino B., Mair C. Quashie N., and Antczak R. Loneliness Before and During the COVID-19 Pandemic: Are Unpartnered and Childless Older Adults at Higher Risk? European Journal of Ageing. **19**, 1327–1338 (2022).

[5] Margolis, R. and Verdery, A. M.: Older Adults Without Close Kin in the United States. J Gerontol. B Psychol. Sci. Soc. Sci., **72 (4)**, 688–693 (2017).

[6] Margolis, R. and Wright L.: Older Adults With Three Generations of Kin: Prevalence, Correlates, and Transfers. J Gerontol. B Psychol. Sci. Soc. Sci., **72 (6)**, 1067–1072 (2017).

[7] Quashie N., Arpino B., Antczak R. and Mair C. Childlessness and Health among Older Adults: Variation across 5 Outcomes and 20 Countries. The Journal of Gerontology: Series B, **76(2)**, 348–359 (2021).

[8] Verdery, A. M. and Margolis, R.: Projections of white and black older adults without living kin in the United States, 2015 to 2060. PNAS, **114 (42)** (2016).

[9] Verdery, A. M., Margolis, R., Zhou Z., Chai X., and Rittirong, J.: Kinlessness Around the World. J Gerontol. B Psychol. Sci. Soc. Sci., **74 (8)**, 1394–1405 (2019).

[10] Zhou Z., Verdery, A. M. and Margolis, R.,: No Spouse, No Son, No Daughter, No Kin in Contemporary China: Prevalence, Correlates, and Differences in Economic Support. J Gerontol. B Psychol. Sci. Soc. Sci., **74 (8)**, 1453–1462 (2019).

DRAFT

# Parameter orthogonalization for the Siler mortality model

Claudia Di Caterina[a] and Lucia Zanotto[b]

[a]University of Verona; claudia.dicaterina@univr.it
[b]University of Bologna; lucia.zanotto@unive.it

**Abstract**

Correlation and, in general, close relationships between parameters can cause problems in the estimation of a model and the consequent fluctuation in the trend of its coefficients. We show the connections existing between parameters in the Siler model, one of the most widely used in demography to approximate mortality over the entire life span, and propose a method to reduce them. Parameter orthogonalization via the Gram-Schmidt-Fisher scoring algorithm seems a promising technique for limiting identification issues and numerical instabilities often encountered when maximizing the likelihood.

***Keywords:*** Siler model, collinearity, orthogonal parameters.

DRAFT

## 1. Introduction

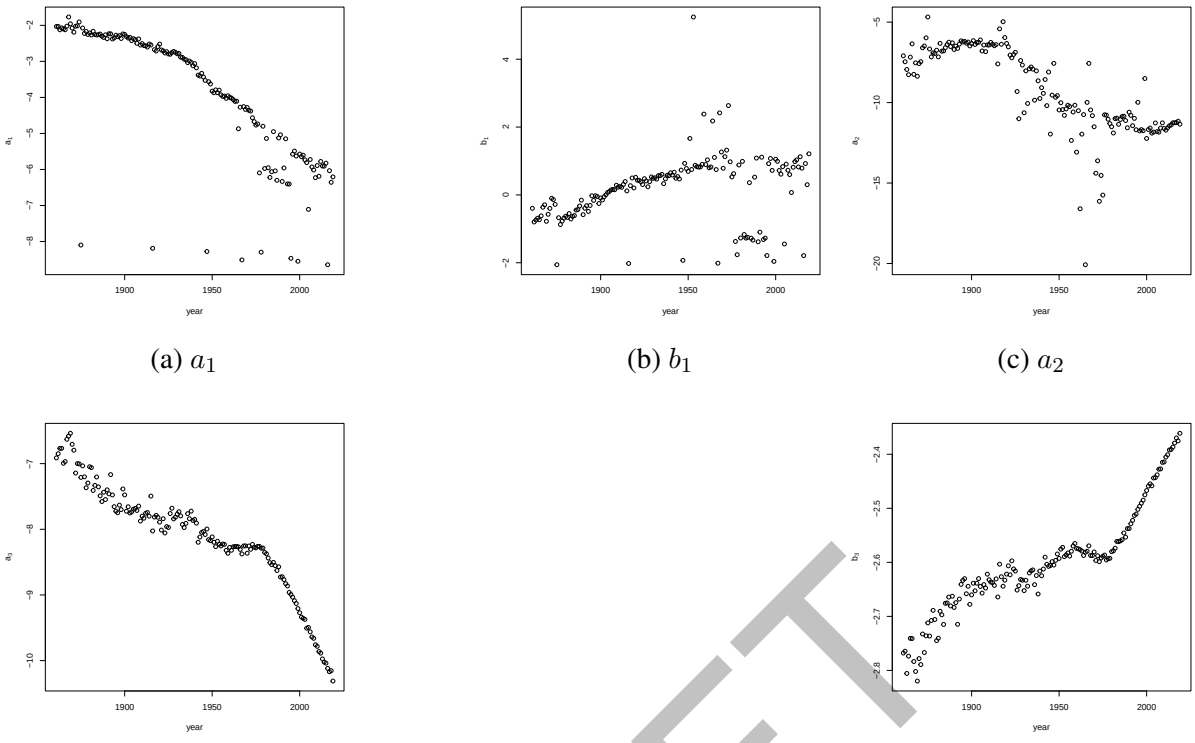
Historically, there are two models mainly used in demography to approximate mortality over the entire lifespan. The first was proposed by Siler (21; 22) to estimate the death rates and consists of three components (five parameters): a negative Gompertz function for infant mortality, a constant representing deaths occurring randomly with respect to age (15), and a Gompertz function (9) for the latter part of the curve. Subsequently, to estimate the probability of death, Heligman and Pollard (10) introduced an 8-parameter model that also captures accidental mortality (i.e., a hump often observed in the death function for males, corresponding to teenage or young adult ages). Rogers (19) and Gage and Mode (7) noted that estimation of the Heligman-Pollard model is difficult, due to model overparameterization and numerical issues. Such numerical difficulties create large fluctuations in parameter estimation over time and space, resulting because of inconsistencies in the dispersion matrix (6). These estimation problems led Dellaportas et al. (6) and Sharrow et al. (20) to use a Bayesian estimation approach.

Although the Siler model has fewer coefficients, both its least squares and maximum likelihood (ML) estimates are quite unstable. This is due to the strong dependence between some of the five parameters. Since the use of parametric models is often preferable (4), in attempt to solve or at least contain the identification problems, we suggest to apply an algorithm which simultaneously orthogonalizes and estimates the parameters of the Siler model (see Section 3) so as to facilitate inferences.

## 2. Siler Model and correlation between its parameters

The model proposed by Siler (21; 22) represents the death rate at age x using the formula:

$$\mu(x) = a_1 e^{b_1 x} + a_2 + a_3 e^{b_3 x}, \quad x \geq 0, \tag{1}$$




(a) $a_1$ (b) $b_1$ (c) $a_2$

(d) $a_3$ (e) $b_3$

Figure 1: Estimation of the Siler model parameters for Sweden 1861–2019.

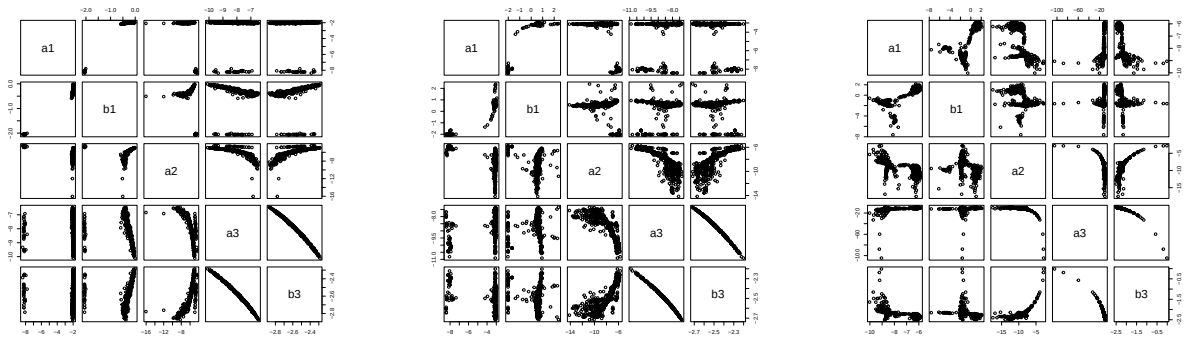
DRAFT

where $a_1 e^{b_1 x}$ is a negative exponential function used to approximate the first part of the curve, $a_2 > 0$ is the Makeham's constant and $a_3 e^{b_3 x}$ is the Gompertz model (9), fitting senescent death rates.

Two methods are usually adopted to estimate the five positive model parameters: least squares or ML. However, least squares are less convenient for this model because nonlinear least squares do not have the same desirable properties of ordinary ones (2). Instead, the ML approach has been tried since 1970 (8) with good results. In the latter case, we can assume that the number of deaths in a given age interval follows a Poisson distribution, where its rate equals the assumed hazard function (for details, see the Section 3 "Estimation methods" in 2). ML estimates are then obtained by numerically maximizing the resulting likelihood function (see Section 3.2), for example using the `optim` routine in R (18).

For illustrative purposes, we focus on Swedish male mortality rates from 1861 to 2019 downloaded from the Human Mortality Database (11). This country offers a long time series covering different mortality scenarios: from before the demographic transition (very high infant mortality, low life expectancy at birth) to the present day (very low infant mortality, compression and shifting of adult deaths resulting in the exceptional increase in life expectancy). As shown in Figure 1, parameter estimates for the Siler model have an irregular behavior over time. This is not due to the quality of the data, but to identification problems caused by their intercorrelation.

A simulation study was implemented in order to better understand cross-correlation among parameters. Three years were selected to represent the whole time series for Sweden: 1861, characterized by pre-demographic transition mortality, 1939, a period of changing from high to low mortality, and the pre-pandemic 2019, marked by high life expectancy. For every year, 1,000 samples of 100,000 observations were drawn from a multinomial distribution with parameters equal to the death probabilities in each age group. These probabilities were derived from the mortality table for the corresponding year. The Siler model was fit on each sample and the correlations between parameter estimates were computed. Figure 2 reports, for all three years, very strong relationships between the parameters $a_1$ and $b_1$, related to infant mortality (correlation greater than 0.88) and between $a_3$ and $b_3$, related to adult mortality (correlation lower than -0.86 and, for 1861 and 1939, lower than -0.96) (16). Also not negligible is the correlation




(a) Sweden 1861 (b) Sweden 1939 (c) Sweden 2019

Figure 2: Correlation plots between estimates of the Siler model for Swedish mortality data.

between $a_2$ and $b_3$, always higher than 0.77.

## 3. Parameter Orthogonalization

When the parameters of a statistical model are orthogonal, it is easier to quantify and distinguish the amount of information carried by the data on the different aspects of the phenomenon under investigation. The fact that resulting estimates are not strongly correlated allows optimization algorithms to converge faster and interpretation of results to be more reliable.

### 3.1 Likelihood setting

The problem of parameter orthogonality in ML estimation dates back to (13) and was later studied, among others, by (12), (1) and (5). Suppose the data $y = (y_1, \ldots, y_n)$ have distribution $f(y; \theta)$ depending on the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. We denote by $l(\theta) = \log f(y; \theta)$ the log-likelihood and by $u(\theta) = \partial l(\theta)/\partial\theta$ the score, and assume that usual regularity conditions are met by the parametric model (17). Two parameter components, $\theta_1$ of dimension $p_1$ and $\theta_2$ of dimension $p_2$, with $p = p_1 + p_2$, are said to be (globally) orthogonal if the entries of the Fisher information $i(\theta)$ matrix satisfy

$$i_{st} = E_\theta\left(\frac{\partial l}{\partial\theta_s}\frac{\partial l}{\partial\theta_t}\right) = E_\theta\left(-\frac{\partial^2 l}{\partial\theta_s\partial\theta_t}\right) = 0\,, \quad \forall\theta \in \Theta, \quad s = 1, \ldots, p_1;\ t = p_1 + 1, \ldots, p.$$

The major consequence of such condition is that the ML estimators of $\theta_1$ and $\theta_2$ are asymptotically independent. This implies that the ML estimate of one parameter is not affected or changes slowly with that of the other, which is helpful to reduce numerical instabilities during the estimation process.

Cox and Reid (5) proposed a general strategy to achieve orthogonalization between a scalar parameter of interest and a set of $q$ nuisance parameters. However, that procedure is based on the solution of $q$ partial differential equations, which is often computationally infeasible or not unique. A more practical solution consists in applying the Gram-Schmidt orthogonalization process (3) via the iterative Fisher scoring algorithm presented in (14). More specifically, a reparameterization $\Theta^*$ is sought through a linear transformation of $\Theta$ with unity Jacobian such that the parameters in $\theta^* = (\theta_1^*, \ldots, \theta_p^*)^\top$ are mutually uncorrelated. As shown in (14, Sect. 4.1), solving a system of linear equations leads to $\theta^* = \theta^*(\theta) = B\theta$, where $B = B(\theta)$ is a lower triangular $p \times p$ matrix with ones on the main diagonal and $(j, k)$-th element

$$b_{jk} = \sum_{r=1}^{j-1} i^{rj} i_{r,j-k}, \qquad j = 2, \ldots, p;\ k = 1, \ldots, j-1, \tag{2}$$

where $i^{rj}$ equals the $(r, j)$-th entry of the inverse Fisher information matrix $i(\theta)^{-1}$. Starting from the update $(m + 1)$ of the classical Fisher scoring algorithm, since the transformation matrix $B$ is non-

DRAFT

singular by construction, it is possible to summarize the Gram-Schmidt-Fisher scoring algorithm (14, Sect. 4.3) by the iterative routine

$$\theta^*_{(m+1)} = \theta^*_{(m)} + i(\theta^*_{(m)})^{-1} u(\theta^*_{(m)}) \,, \tag{3}$$

where $\theta^*_{(m)} = B\theta_{(m)}$ is the orthogonal parameter vector at the $m$-th iteration, $i^{-1}(\theta^*_{(m)}) = Bi(\theta_{(m)})^{-1}B^\top$ is asymptotically diagonal and $u(\theta^*_{(m)}) = (B^\top)^{-1}u(\theta_{(m)})$ is the score function in the parameterization $\Theta^*$.

## 3.2 Gram-Schmidt-Fisher scoring algorithm for the Siler model

As the Gram-Schmidt-Fisher scoring algorithm seems particularly convenient for non-linear ML estimation of models with a small parameter space dimension (14, Sect. 6), we propose to apply it as an approximate orthogonalization technique for the parameter $\theta = (\theta_1, \ldots, \theta_p)^\top = (a_1, b_1, a_2, b_2, b_3)^\top$ of the Siler model (see Section 2).

Consider the dataset of mortality rates $m_x$ at age $x$ ($x = x_0, \ldots, \omega$) and Siler's mortality hazard function $\mu(x) = \mu_x(\theta)$. Assuming the number of deaths are independent across age groups and follow a Poisson distribution, the log-likelihood of the model can be expressed by

$$l(\theta) = \sum_{x=x_0}^{\omega} \{m_x \log \mu_x(\theta) - \mu_x(\theta)\} \,,$$

and the 5-dimensional score function $u(\theta)$ has $j$-th entry

$$u_j = \frac{\partial l}{\partial \theta_j} = \sum_{x=x_0}^{\omega} \left\{ \frac{m_x}{\mu_x(\theta)} - 1 \right\} \frac{\partial \mu_x(\theta)}{\partial \theta_j} \,, \qquad j = 1, \ldots, 5 \,,$$

where $\partial\mu_x(\theta)/\partial\theta_1 = e^{-\theta_2 x}$, $\partial\mu_x(\theta)/\partial\theta_2 = -\theta_1 x e^{-\theta_2 x}$, $\partial\mu_x(\theta)/\partial\theta_3 = 1$, $\partial\mu_x(\theta)/\partial\theta_4 = e^{\theta_5 x}$ and $\partial\mu_x(\theta)/\partial\theta_5 = \theta_4 x e^{\theta_5 x}$. Then, as $E_\theta(m_x) = \mu_x(\theta)$, it is possible to show that the $(j,k)$-th element of the Fisher information matrix $i(\theta)$ equals

$$i_{jk} = \sum_{x=x_0}^{\omega} \frac{1}{\mu_x(\theta)} \frac{\partial \mu_x(\theta)}{\partial \theta_j} \frac{\partial \mu_x(\theta)}{\partial \theta_k} \,, \qquad j = 1, \ldots, 5 \,;\; k = 1, \ldots, 5.$$

Since the parameters of the Siler model must be positive, i.e. $\theta_j > 0$ ($j = 1, \ldots, 5$), it is safer to maximize numerically the log-likelihood in the unconstrained parameter space $\mathbb{R}^5$ under the reparameterization $\psi = \psi(\theta) = \log\theta$. The likelihood quantities we need to perform the Gram-Schmidt-Fisher scoring algorithm in the parameterization $\Psi$ can be readily obtained by (17)

$$u^\Psi(\psi) = J_\theta(\psi)^\top u(\theta(\psi)) \quad \text{and} \quad i^\Psi(\psi) = J_\theta(\psi)^\top i(\theta(\psi)) J_\theta(\psi) \,,$$

where $J_\theta(\psi)$ is the diagonal Jacobian matrix of the inverse transformation $\theta(\psi) = e^\psi$, with $j$-th diagonal entry $\partial\theta_j/\partial\psi_j = e^{\psi_j}$.

Once we derive analytically $u^\Psi(\psi)$ and $i^\Psi(\psi)$, the implementation of the Gram-Schmidt-Fisher scoring algorithm exposed in (14, Sect. 4.3.1) can be adapted as follows:

1. Start with an initial estimate $\psi_{(0)}$ of $\psi = \log\theta$.
2. Set $m = 0$.
3. Estimate the $5 \times 5$ transformation matrix $B = B(\psi_{(m)})$ with entries given in (2).
4. Calculate the current orthogonal parameter estimate $\psi^*_{(m)} = B\psi_{(m)}$.
5. Update $\psi^*$ by (3): $\psi^*_{(m+1)} = \psi^*_{(m)} + i^\Psi(\psi^*_{(m)})^{-1} u^\Psi(\psi^*_{(m)})$.
6. Update $\psi$: $\psi_{(m+1)} = B^{-1}\psi^*_{(m)}$.
7. Stop if $||\psi^*_{(m+1)} - \psi^*_{(m)}||_2 / ||\psi^*_{(m)}||_2 < \varepsilon$.
8. Set $m = m + 1$ and repeat from Step 3.

DRAFT

Table 1: Correlation matrix for estimates of the Siler model obtained in the original parameterization via `optim` (left) and in the orthogonal parameterization via the Gram-Schmidt-Fisher scoring algorithm (right) for Sweden 2019.

| | Original parameterization $\theta$ | | | | | | Orthogonal parameterization $\theta^*$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $a_1$ | $b_1$ | $a_2$ | $a_3$ | $b_3$ | | $a_1^*$ | $b_1^*$ | $a_2^*$ | $a_3^*$ | $b_3^*$ |
| $a_1$ | 1.000 | 0.890 | -0.676 | 0.279 | -0.508 | $a_1^*$ | 1.000 | -0.160 | -0.150 | 0.024 | -0.293 |
| $b_1$ | 0.890 | 1.000 | -0.530 | 0.189 | -0.387 | $b_1^*$ | -0.160 | 1.000 | 0.362 | -0.186 | -0.177 |
| $a_2$ | -0.676 | -0.530 | 1.000 | -0.492 | 0.773 | $a_2^*$ | -0.150 | 0.362 | 1.000 | -0.296 | -0.283 |
| $a_3$ | 0.279 | 0.189 | -0.4924 | 1.000 | -0.865 | $a_3^*$ | 0.024 | -0.186 | -0.296 | 1.000 | 0.731 |
| $b_3$ | -0.508 | -0.387 | 0.773 | -0.865 | 1.000 | $b_3^*$ | -0.293 | -0.177 | -0.283 | 0.731 | 1.000 |

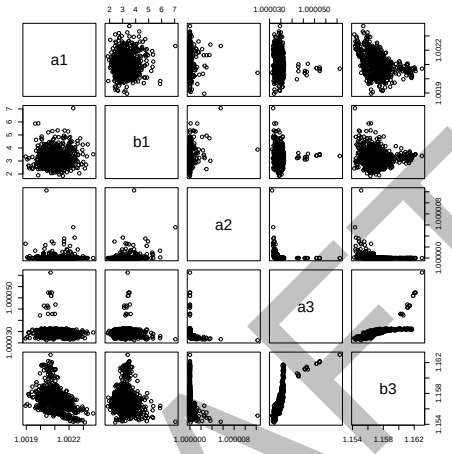



Figure 3: Correlation plots of the Siler model for Sweden 2019 with the orthogonal parametrization.

DRAFT

## 3.3 Preliminary results and future work

R is used to program the Gram-Schmidt-Fisher scoring algorithm described above. Partial results are shown in Figure 3, which refers to year 2019. In the original parameterization, as discussed in Section 2, estimates of $a_1$ and $b_1$ are highly positively correlated, while $a_2$ and $b_3$ are highly negatively related. When switching to the orthogonal parameterization, instead, the corresponding correlations are smaller. We note an overall drop in the intercorrelation between parameters under the orthogonal parameterization, except for that between $a_3$ and $b_3$ whose reduction is only partial (see Table 1.

It is worth pointing out that the Gram-Schmidt orthogonalization affects differently the single components of the parameter vector, so the estimates in the first positions are less prone to suffer from round-off errors (14, Sect. 4). This might explain the inflation observed in the correlation between the parameters $a_3$ and $b_3$, in positions $j = p - 1, p$ of $\theta$. Another important remark concerns the interpretation of the orthogonal estimates with respect to the original parameterization: $\theta_j^* = \theta_j - \sum_{k=1}^{j-1} b_{j,j-k}\theta_{j-k}$ $(j = 1 \ldots, 5)$ where the $b_{jk}$s can be viewed as coefficients of the multiple linear regression of $\theta_j$ on the previous parameters $\theta_1, \ldots, \theta_{j-1}$ (14, Sect. 4).

Finally, the current implementation of the Gram-Schmidt-Fisher scoring algorithm for the Siler model is at a very early stage. We plan to study more in depth its numerical stability, sensitivity to starting values and convergence times. The ultimate goal is to reduce identification problems and obtain more stable ML estimates, smoothing out the trend of the parameters.

## References


[1] Amari, S. *Differential Geometrical Methods in Statistics*. New York: Springer-Verlag, 1985.

[2] Canudas-Romo, V., Mazzuco, S., and Zanotto, L. Measures and models of mortality. In *Handbook of Statistics*, vol. 39. Elsevier, 2018, pp. 405–442.

[3] Clayton, D. Algorithm as 46: Gram-schmidt orthogonalization. *J. R. Statist. Soc. C 20* (1971), 335–338.

[4] Congdon, P. Statistical graduation in local demographic analysis and projection. *J. R. Statist. Soc. A 156* (1993), 237–270.

[5] Cox, D. R., and Reid, N. Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B 49* (1987), 1–39.

[6] Dellaportas, P., Smith, A. F., and Stavropoulos, P. Bayesian analysis of mortality data. *J. R. Statist. Soc. A 164* (2001), 275–291.

[7] Gage, T. B., and Mode, C. J. Some laws of mortality: how well do they fit? *Human Biology* (1993), 445–461.

[8] Garg, M. L., Rao, B. R., and Redmond, C. K. Maximum-likelihood estimation of the parameters of the gompertz survival function. *J. R. Statist. Soc. C 19* (1970), 152–159.

[9] Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115 (1825), 513–583.

[10] Heligman, L., and Pollard, J. H. The age pattern of mortality. *Journal of the Institute of Actuaries 107* (1980), 49–80.

[11] Human Mortality Database. University of California, Berkeley (USA), and max Planck Institute for demographic research (Germany). *Available: www.mortality.org (Data downloaded: February 2023*.

[12] Huzurbazar, V. S. Sufficient statistics and orthogonal parameters. *Sankhyā 17* (1956), 217–220.

[13] Jeffreys, H. Theory of probability. *Theory of Probability* (1939).

[14] Kwagyan, J., Apprey, V., and Bonney, G. E. Gram–schmidt–fisher scoring algorithm for parameter orthogonalization in mle. *Cogent Mathematics 3* (2016), 1159847.

[15] Makeham, W. M. On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries 8*, 6 (1860), 301–310.

[16] Mazzuco, S., Scarpa, B., and Zanotto, L. A mortality model based on a mixture distribution function. *Population Studies 72* (2018), 191–200.

[17] Pace, L., and Salvan, A. *Principles of statistical inference: from a Neo-Fisherian perspective*. World scientific, 1997.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[19] Rogers, A. Parameterized multistate population dynamics and projections. *Journal of the American Statistical Association 81* (1986), 48–61.

[20] Sharrow, D. J., Clark, S. J., Collinson, M. A., Kahn, K., and Tollman, S. M. The age pattern of increases in mortality affected by hiv: Bayesian fit of the heligman-pollard model to data from the agincourt hdss field site in rural northeast south africa. *Demographic research 29* (2013), 1039.

[21] Siler, W. A competing-risk model for animal mortality. *Ecology 60* (1979), 750–757.

[22] Siler, W. Parameters of mortality in human populations with widely varying life spans. *Statistics in medicine 2* (1983), 373–380.


DRAFT

# Pseudo-observations in survival analysis

Marta Cipriani[a], Alfonso Piciocchi[b], Valentina Arena[b], and Marco Alfo'[a]

[a]Sapienza University of Rome; marta.cipriani@uniroma1.it, marco.alfo@uniroma1.it
[b]GIMEMA Foundation, Rome; a.piciocchi@gimema.it, v.arena@gimema.it

**Abstract**

In recent years, pseudo-observations have gained much interest as they address one of the main issues with survival data, i.e. incomplete observations due to right-censoring, and they allow for the analysis of such data by standard statistical methods. In this work, a method based on pseudo-values will be illustrated for direct regression modelling of the survival function, the restricted mean survival time, and the cumulative incidence function in a competing risks situation. The method based on pseudo-observations will be compared to the traditional Cox proportional hazard and Fine-Gray models.

***Keywords:*** survival data, pseudo-observations, right-censoring, GEE, informative missing

DRAFT

## 1. Introduction

The existence of missing data is a topic of crucial interest in survival analysis. Frequently, we do not have complete observations for some individuals in the study due to the inherent structure of survival data. In this context, a source of information loss is the occurrence of censoring, a technicality which must be dealt with when examining survival data. Thus, standard statistical models cannot be used to analyse this kind of data.

In the classical theory of survival analysis, the censoring is assumed to be independent of the survival time (8), i.e. censoring is non-informative on the *true* survival time. In this perspective, censoring is essentially due to delayed entry. Namely, units that enter late in the study remain under observation for a shorter time and, reasonably, are less likely to be associated with an event, given the individual features. This assumption implies that the velocity of occurrence of the event of interest can be estimated by considering only the survival experience of non-censored units.

In more complex situations, when units drop out prematurely, censoring cannot be directly assumed to be independent of the survival experience and, therefore, it may be informative. In this case, the issue is to account for the information that could not be observed. Therefore, under the *missing at random* (MAR) assumption (12), the censored observations can be imputed on the basis of the observed ones.

The so-called pseudo-observations estimate a set of pseudo-values of the (time-related) response variable for each individual in the study at a given set of time points. Hence, they allow to model the effect of covariates on the event times by methods usually restricted to longitudinal data with full information.

## 2. Methods

The pseudo-observation approach was first suggested by Andersen *et al.* (3) to perform a generalized linear regression analysis of survival data, albeit it is a more general method with many fields of application. The theory is based on the pseudo-observations drawn from the jackknife method (11). The following definition can be used to introduce the concept.

**Definition 1.** *Let $\phi$ be some function of the survival time $T$ for which we are interested in the parameter $\theta(T) = E(\phi(T))$ and let $\widehat{\theta}(T)$ be an unbiased (or approximately unbiased) estimator of $\theta(T)$. The pseudo-observations for subject $i$ is then defined as*

$$\widehat{\theta}_i(t) = n\widehat{\theta}(t) - (n-1)\widehat{\theta}^{-i}(t)$$

*where $\widehat{\theta}^{-i}$ is the estimator derived from the sample of size $n-1$ obtained by ignoring subject $i$.*

In general, one may then think of the pseudo-observation $\theta_i(t)$ as a replacement for the, possibly incompletely observed, random variable $\phi_i(t)$. Once the pseudo values are obtained they can be used in a standard generalized estimation equation (GEE) or generalized linear model (GLM) to obtain regression estimates:

$$g(E(\phi_i(t)|\mathbf{Z}_i)) = \alpha + \beta'\mathbf{Z}_i$$

with $g(.)$ a *link function* and $\mathbf{Z}_i$ the subject-specific covariates vector.
The GEE approach based on pseudo-values is frequently proposed in the literature to fit a model for functions of the event time. The justification for using pseudo-observations in such approach relies on some neat results by Graw *et al.* (7) concerning the estimated regression parameters.
The main advantage of this technique is that it enables modelling the survival data by a wider number of regression methods. Consequently, it allows to take into account very useful parameters estimates in the context of survival analysis (e.g. a regression model for the restricted mean survival time).
In addition, as discussed by Andersen *et al.* (4), the so-called classical survival methods require that several assumptions about the models used are tested, such as the propotional hazard hypothesis in the Cox model. By contrast, the hypotheses needed to use pseudo-observations are less restrictive. Simple uses of the method do, however, require that censoring is independent of covariates, although this assumption might be relaxed (5).

## 2.1 Pseudo-values for the survival function

If the target parameter $\theta(t)$ is the survival function, $\widehat{\theta}(t)$ can be obtained by the Kaplan-Meier estimator $\widehat{S}(t)$. The pseudo-observations for subject $i$ is then

$$\widehat{S}_i(t) = n \cdot \widehat{S}(t) - (n-1) \cdot \widehat{S}^{-i}(t)$$

As usual in survival analysis in discrete time (2), we may use a *complementary log-log* link function and define the estimation model as follows

$$g(S(t|\mathbf{Z}_i)) = \log(-\log(S(t|\mathbf{Z}_i))) = \log H_0(t) + \beta'\mathbf{Z}_i$$

This model corresponds to the Cox model in discrete time, where

$$S(t|\mathbf{Z}_i) = S_0(t)^{\exp(\beta'\mathbf{Z}_i)} = \exp\Big(-H_0(t)\exp\big(\beta'\mathbf{Z}_i\big)\Big)$$

## 2.2 Pseudo-values for other quantities

Pseudo-observations allow us to define a model for other quantities associated to the survival function, as the **restricted mean survival** (RMS). The RMS is defined as $\mu(t) = \int_0^t S_t(u)$ and it can be estimated by the area under the Kaplan-Meier curve up to time $t$:

$$\widehat{\theta}(t) = \widehat{\mu}(t) = \int_0^t \widehat{S}_t(u)du$$

and pseudo-observations are

$$\widehat{\mu}_i(t) = n \cdot \widehat{\mu}(t) - (n-1) \cdot \widehat{\mu}^{-i}(t)$$

DRAFT

In a situation with two competitive risks, associated to hazard functions $h_1(t)$ and $h_2(t)$, respectively, the **cumulative incidence function** is defined by

$$C_k(t) = \int_0^t h_k(u) \exp\left[\int_0^u \big(h_1(v) + h_2(v)\big)dv\right]du \qquad k = 1, 2$$

and it may be estimated by the Aalen-Johansen estimator $\widehat{C}_k(t)$. The $i$'th pseudo-observation corresponding to $C_k(\cdot)$ at time $t$ is then given by

$$\widehat{C}_{ki}(t) = n \cdot \widehat{C}_k(t) - (n-1) \cdot \widehat{C}_k^{-i}(t)$$

Using the *complementary log-log* link function leads to the definition of an estimation model corresponding to the proportional subdistribution hazards model proposed by Fine and Gray (6):

$$g(C_k(t|\mathbf{Z}_i)) = \log\Big(-\log\big(C_k(t|\mathbf{Z}_i)\big)\Big) = 1 - \exp\big[H_0 \cdot e^{\beta' \mathbf{Z}_i}\big]$$

## 3. Case study

The potential of methods based on pseudo-observations is shown through an application to data collected in a phase II experimental trial, the GIMEMA AML1310, and carried out by Fondazione GIMEMA Franco Mandelli Onlus (13). In order to make a comparison, standard methods of survival analysis and pseudo-observations approach have been compared on these data.

### 3.1 Available data

The sample includes 500 acute myeloid leukemia (AML) patients on which several prognostic factors have been collected:

- *demographic characteristics*: age, sex
- *hematological characteristics*: white blood cells, platelets, etc.
- *genetic profiles*: mutations of specific genes
- *risk categories*: identified according to NCCN 2009[1]
- *clinical parameters*: minimal residual disease (MRD), performance status, etc.

The GIMEMA Foundation has developed a risk-adapted, MRD-oriented, prospective clinical trial, the strategy of which consisted of the prognostic integration of pre-treatment cytogenetics and genetics with post-consolidation MRD. Based on this strategy, patients were to receive post-consolidation autologous or allogenic stem cell transplantation, respectively, depending on their risk profile.
The primary trial objective was to evaluate the treatment strategy in terms of Overall Survival (OS) at 24 months.

### 3.2 Results

The number and position of time points, for which the pseudo-observations are calculated, is a choice which must be made prior to the analysis. One time point is enough to obtain estimates of the regression parameter (9); however, more time points may needed to capture the trend in the time to event distribution. In this analysis five time points, equally spaced and concentrated within the most dense region of the time to event distribution, have been chosen.
The pseudo-observation-based models discussed in the following are based on the assumption that the censoring time is independent of the event time and of all covariates in the survival model.
As for the **survival function** estimates, the results obtained using the two methods (the Cox model and

[1]National Comprehensive Cancer Network (NCCN) 2009 recommendations

DRAFT

the model based on pseudo-observations) do not substantially differ in terms of point estimate or significance. However, the pseudo-observation approach helps avoid the assumption of hazard proportionality needed in the former.
Even though the impact of using pseudo-observations for inference on survival function was minor, the strength of this approach arises in situations where no standard models exist. Indeed, the pseudo-observations allow to build regression models for the **restricted mean survival** time.
The R package `survival` does not enable to compute the RMS conditional on more than one covariate. On the contrary, the two packages devoted to modelling pseudo-observations (`pseudo` and `eventglm`) provide the chance to assess the simultaneous effect of a set of covariates on the restricted mean lifetime. In this case the *link function* $g(.)$ used in the estimation model is an identity function, so that the estimated parameters represent the difference between the areas under the estimated survival curves.
As for the **cumulative incidence function** estimates, the relapse has been considered the primary event (death is the competing event) to build the competing risk model. The comparison between the Fine-Gray and the GEE models shows that the first produces a reduced set of significant effects than that obtained using the pseudo-observations.

### 3.2.1 Some clues on censoring distribution

Thanks to the tools available in the R packages to work with pseudo-observations (10), it is possible to evaluate the sensitivity of models to hypotheses on the censoring-data mechanism.
The identification of the assumption which reasonably best fits the analysed data has been carried out by comparing the results obtained according to the different hypotheses. The best fitting is found by looking at the length of the estimated confidence intervals.
The results of this work point out that the censoring times are reasonably described by an additive risk model (1). Thus, assuming that the censoring depends on covariates, the pseudo observations are calculated with the inverse probability of censoring weighted (IPCW) approach, where the censoring probabilities are estimated using Aalen's additive hazards model.

DRAFT

## 4. Conclusions

The characteristic of recovering censored information at an event time makes pseudo-observations particularly advantageous to derive more efficient estimators for quantities of interest in survival analysis. Especially when considering studies meant to assess causal links, methods based on pseudo-observations are going to be widely used to deal with the issue of partially observed information due to censoring.
The main advantage of this approach is that it enables modelling the survival data by a wider number of regression methods, e.g. a regression model for the restricted mean survival time.
The comparison between standard methodologies in survival analysis and techniques based on pseudo-observations may highlight the potentialities of the proposed method and the ease to consider a larger number of models generally precluded to censored data. Furthermore, it is possible to obtain more detailed information and evaluate the sensitivity with respect to the hypotheses about the censoring mechanism.
Over recent years, interest in pseudo-observations has been growing and researchers have focused on new methodologies for producing increasingly precise pseudo-observations. Among these, the use of a semi-parametric approach to computing pseudo-observations proved to be more efficient than a completely non-parametric one.
In the literature about pseudo-observations there are still few unanswered questions which may provide ideas for future research developments in this area. Indeed, the performance of the methods based on pseudo-observations may depend on the amount and type of missing information (missing at random or not) recorded up to the end of the study. For this reason an in-depth research on this issue is needed and may contribute to improving the method.

## References


[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 701-726.

[2] Aitkin, M., Anderson, D., Francis, B., Hinde, J. (1989). *Statistical Modelling in GLIM*, Clarendon Press, Oxford.

[3] Andersen, P.K., Klein, J.P., Rosthoj. S. (2003) Generalized linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, **90**, 15-27.

[4] Andersen, P.K., Syriopoulou, E., Parner, E.T. (2017). Causal inference in survival analysis using pseudo-observations. *Statistics in medicine*, **36**, 2669-2681.

[5] Binder, N., Gerds, T.A., Andersen, P.K. (2014) Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, **20** (2), 303–315.

[6] Fine, J.P., Gray, R.J. (1999) A proportional hazards model for the subdistribution of a competing risk. *JASA*, **94**, 496-509.

[7] Graw, F., Gerds, T.A., Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, **15** (2), 241–255.

[8] Klein, J.P., Moeschberger, M.L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data*. New York: Springer.

[9] Klein, J.P., Andersen, P.K. (2005) Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function. *Biometrics*, **61**(1), 223–229.

[10] Klein, J.P., Gerster, M., Andersen, P.K., Tarima, S., Pohar, M. (2008). SAS and R Functions to Compute Pseudo-values for Censored Data Regression. *Comput Methods Programs Biomed*, **89**, 289–300.

[11] Rupert G. Miller. (1974) The jackknife - a review. *Biometrika*, **61**(1), 1–15.

[12] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

[13] Venditti, A., Piciocchi, A., Candoni, A., Melillo, M., *et al.* (2019). GIMEMA AML1310 trial of risk-adapted, MRD-directed therapy for young adults with newly diagnosed acute myeloid leukemia. *Blood*, vol.134, n.12.


DRAFT

# Sex Gap in Cancer-Free Life Expectancy: The Association with Smoking, Obesity and Physical Inactivity

Alessandro Feraldi[a], Cristina Giudici[b] and Nicolas Brouard[c]

[a] Department of Statistical Sciences, Sapienza University of Rome; alessandro.feraldi@uniroma1.it
[b] MEMOTEF Department, Sapienza University of Rome; cristina.giudici@uniroma1.it
[c] National French Institute for Population Studies (INED), Paris; brouard@ined.fr

**Abstract**

We measured sex-specific total life expectancy, cancer-free life expectancy (CFLE), and years spent with cancer according to the (co-) occurrence of three behavioural risk factors, such as smoking, obesity and physical inactivity. We examined differences between women and men using data from the United States Health and Retirement Study 2008–2018 and we applied multistate lifetable approach for each combination of smoking, obesity, and physical inactivity, controlling for education. Risk factors were associated with shorter CFLE, the shortest observed in current smoker's men and women (-4.7 years at age 50). Reductions of CFLE in physically inactive people was higher in women (-3.3 years) than in men (-2.3 years) and obesity had a significant effect only in women (-3.1 years). Sex differences decreased at older age. The (co-) occurrence of behavioural risk factors reduces the CFLE disadvantage of men compared to women.

**Key words:** sex gap, cancer-free life expectancy, mortality risk factors

DRAFT

## 1. Introduction

Cancer is second leading cause of death globally. According to the United States Cancer Statistics (USCS), around 16.9 million men and women had a cancer history in the United States (US) in 2019 and there are projected more than 22.0 million cases in 2030 (USCS 2019). In 2019, the prevalence of cancer in women and in men was 4.0% and 3.8%, respectively. Prostate, breast, lung and bronchus, and colorectal cancer are the most frequently diagnosed cancers in the US (USCS 2019). Although improvements in cancer survival due to advances in screening technology and implementation and treatments in recent decades, cancer still contributes significantly to years lived with disability and to the risk of mortality.

Studies have shown that modifiable lifestyle factors such as smoking, physical activity, alcohol intake, body weight, and diet quality affect both life expectancy and incidence of chronic diseases, including cancer (Stenholm et al 2016; Leskinen et al 2018). Nevertheless, little research has looked at how multiple behavioral risk factors may affect life expectancy free from the major diseases, especially free from cancer (i.e. Cancer-Free Life Expectancy - CFLE).

Using data from Nurses' Health Study and the Health Professionals Follow-Up Study in 1980–2014, Li and colleagues (Li et al 2020) estimated CFLE according to five healthy behaviours in the US They showed that at age 50, compared to women with healthy behaviours, women with unhealthy behaviours can expect to live 8.3 year less without cancer, whereas this difference was 6.0 years in men. Conversely, a study of Zaninotto and colleagues in the US between 2002 and 2013, observed larger reduction in chronic disease free life expectancy (including cancer) in men than in women in case of risky behaviour (Zaninotto al 2020). Most of the existing studies have some limitations related to non-representative populations (reducing the generalizability of results), short duration of follow-up and study populations aged <75 years (Li et al 2020; Zaninotto al 2020; Leskinen et al; Leskinen et al 2018).

Accordingly, the aim of this study was to examine the extent of the reduction in CFLE due the (co-) occurrence of risk factors such as smoking, obesity and physical inactivity, in a nationally representative

longitudinal survey of older people (aged 50 and over) in the United States. Additionally, we study the sex gap in CFLE and in the association with multiple risk factors, with a follow-up of 10+ years. The number of years lived without cancer is estimated with health expectancy outcomes, a measure that combines incidence and mortality to estimate life expectancy lived with and without cancer. Taking into account both morbidity and mortality, estimates of life expectancy free of cancer provide useful metrics for health professionals and policy makers in order to better estimate future healthcare costs of cancer and to plan for healthcare needs.

## 2. Data and methods

Data were retrieved from the Health and Retirement Study (HRS) in the US, an ongoing nationally representative longitudinal study on health, behavioural risk factors and wealth, in which people have been interviewed approximately every two years, from 1992 to 2018. In the HRS database, mortality follow-up is ascertained through linkages to the National Death Index and reports from survivors. We used data from 2008 (baseline) to 2018 and we included people aged 50+ with valid data on cancer and behavioural risk factors. At each wave of the study respondents were asked 'has a doctor ever told you that you have cancer'. This information was used to assess the presence of cancer at each wave, which includes any cancer conditions reported before the age of 50 from available information on respondents. In this study, all individuals in the sample had information on the presence of cancer at baseline (2008). Participants who had missing lifestyle factors at baseline were excluded (423 individuals, 2.5%). The resulting analytical samples included 16,438 aged 50 years and older (out of the 16,861 HRS members aged 50 years and older in 2008).

Smoking status was dichotomized into "Never or former smoker" and "Current smoker". Obesity was measured according to self-reported Body Mass Index (BMI) and dichotomized as "obese" (BMI $\geq$ 35 Kg/m2) and "not obese" (BMI < 35 Kg/m2). Frequency of moderate physical activity was used to assess physical inactivity, which was dichotomized as "physically inactive" if taking part in moderate physical activity for less than one day a week and "physically active" otherwise. The co-occurrence of multiple behavioural risk factors was defined as reporting 2 or more risks.

We used multistate Markov survival models to estimate how participants moved between no cancer (state 1), cancer (state 2), and death (state 3) states. This model had three possible transitions: no cancer to cancer, no cancer to death, and cancer to death. Participants who developed cancer could only move from the disease state to the death state. Age-specific transition rates were modelled using multinomial logistic regression using as covariates age and behavioral risk factors (smoking, obesity and physical inactivity); years of education was included as controlling variable. Separate models were specified for women and men. Sex-specific transition rates are then applied to a synthetic cohort in order to summarize them into duration: Total Life Expectancy (TLE); Cancer-Free Life Expectancy (CFLE) - expected average number of remaining years of life with no cancer; and Cancer Life Expectancy (CLE) - expected average number of remaining years of life expected to live in cancer states.

Sex gap in Cancer-Free Life Expectancy was calculated as absolute difference: females minus males. Transition rates were estimated with the *msm* R package (Jackson 2011) and R package Estimating Life Expectancies in Continuous Time (*ELECT*) was used in order to estimate state-specific life expectancies conditional on reaching age 50 years (van den Hout, Chan & Matthews 2019). Confidence intervals were estimated using 1000 bootstrap samples.

## 3. Preliminary findings

The sample includes 9,606 women (54.6%) with a mean age of 67.8 years and 6,832 men (45.4%) with a mean age of 66.7 years (p=0.07). At age 50, the average number of years that people can expect to live without cancer was 30.6 years in women and 26.8 years for men. The overall sex gap in CFLE at age 50 was 3.9 years. At age 50, a woman with no cancer at baseline could expect to live on average 36.2 years of remaining life expectancy, whereas remaining life expectancy was shorter to 29.5 years for a woman with cancer. Men with no cancer at age 50 can expect to live 33.3 years, whereas men with cancer can expect to live only 27.7 years.

Figure 1 shows life expectancy free from cancer over age, according to the occurrence of behavioral risk factors, for women (panel a) and for men (panel b). At age 50, CFLE was 33.7 years in women and 28.6 years in men with no behavioral risk factors. Compared to this group, in presence of behavioral

DRAFT

risk factors, years of life expected to leave without cancer were lower in both women and men: respectively, 28.9 and 24.0 for smoking, 30.6, 30.3 and 23.6 for physical inactivity, and 26.1 and 22.8 years for 2 or more risk factors. Compared to people with no risk factors, being current smokers was associated with the shortest CFLE in both sexes: approx. 4.7 fewer years free of cancer at age 50. Obesity had a significant effect on cancer-free life expectancy only in women (3.1 fewer years). Similarly, reductions of CFLE in physically inactive people was higher in women (3.3 years) than in men (2.3 years).

Panel c in Figure 1 displays the sex gap in CFLE according to behavioral risk factors. Compared to no risk factors, the presence of risk factors was associated with a smaller sex gaps in CFLE. At age 50, the difference in CFLE between women and men with no risk factors was around 5.0 years. The CFLE sex gap was similar for smoking (4.9 years at age 50), whereas it was smaller to 4.0 years for physical inactivity, 3.4 years for multiple risk factors, and the lowest for obesity, 1.2 years. Differences between women and men were narrower at older ages.

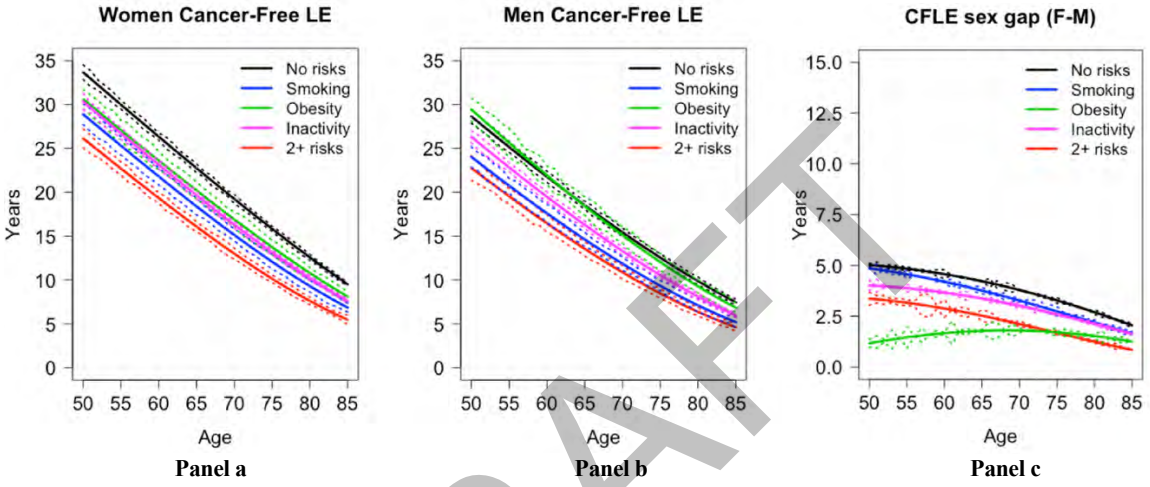



Figure 1: Women (panel a) and Men (panel b) Cancer-Free Life Expectancy (CFLE) and sex gap (Female (F) CFLE - Male (M) CFLE, panel c) in CFLE over age, according to behavioral risk factors.

DRAFT

## 4. Conclusion

Using a nationally representative study of ageing in the US, we showed that behavioural risk factors were associated with reduced remaining number of years spent without cancer. Reducing smoking and obesity, and increasing physical activity among older people could potentially lead not only to longer lives but also healthier lives (free of cancer). Compared to men, women live on average more years free of cancer (about 3.0 years on average between age 50–85). Additionally, the (co-) occurrence of risk factors reduces the cancer-free life expectancy disadvantage of men with respect to women, especially at younger ages. The results of this study provide useful metrics for health professional and policy makers in the quantification of future healthcare costs of cancer and in the plan for healthcare needs.

## References

[1] Leskinen, T., Stenholm, S., Aalto, V., Head, J., Kivimäki, M., & Vahtera, J. (2018). Physical activity level as a predictor of healthy and chronic disease-free life expectancy between ages 50 and 75. *Age and Ageing*, *47*(3), 423-429.

[2] Li, Y., Schoufour, J., Wang, D. D., Dhana, K., Pan, A., Liu, X., ... Hu, F. B. (2020). Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. bmj, 368.

[3] Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort profile: the health and retirement study (HRS). *International journal of epidemiology*, *43*(2), 576-585.

[4] Stenholm, S., Head, J., Kivim¨aki, M., Kawachi, I., Aalto, V., Zins, M., ... Vahtera, J. (2016). Smoking, physical inactivity and obesity as predictors of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. International journal of epidemiology, 45(4), 1260-1270.

[5] U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on 2021 submission data (1999–2019): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; www.cdc.gov/cancer/dataviz, September 2022.

[6] van den Hout, A., Chan, M. S., Matthews, F. (2019). Estimation of life expectancies using continuous-time multi-state models. Computer methods and programs in biomedicine, 178, 11-18.

[7] Zaninotto, P., Head, J., Steptoe, A. (2020). Behavioural risk factors and healthy life expectancy: evidence from two longitudinal studies of ageing in England and the US. Scientific Reports, 10(1), 1-9.

DRAFT

# Women's Exposure to HIV in Africa: the Role of Intimate Partner Violence

Micaela Arcaio[a1], Anna Maria Parroco[a]

[a] University of Palermo; micaela.arcaio@unipa.it,annamaria.parroco@unipa.it

**Abstract**

Sub-Saharan Africa's female population is strongly affected both by HIV and intimate partner violence (IPV). The literature has paid great attention to the association between these two phenomena. This work aims at evaluating the relationship between IPV and HIV using survey data from countries in this region – specifically, data from the Demographic and Health Survey of currently partnered women who were tested for HIV during the survey and were administered the special module on domestic violence. We present the estimates of a logistic regression model with Firth's adjustment, to account for the rarity of being HIV-positive. Among the explanatory variables, a composite indicator of IPV is significantly associated with HIV. This result can be interpreted in the framework of the theory of polyvictimisation.

**Keywords:** HIV, Domestic Violence, Sub-Saharan Africa, Logistic Regression with Firth's adjustment

DRAFT

## 1. Introduction

Recent data about Sub-Saharan Africa (SSA) show that this region is extremely affected both by intimate partner violence[2] (IPV) and HIV. SSA is, indeed, the region most affected by it in the world, since over two-thirds of people living with HIV were residents in these countries in 2021 [2]. IPV is an equally important matter in these countries: 33% of women have been abused at least once in their life by their partner and 20% have been abused in the last year alone [23].

The association between gender dynamics within couples connected to practices of hegemonic masculinity and HIV has been paid great attention to in the literature, especially in countries with high HIV prevalence. Hegemonic masculinity refers to a societal pattern in which stereotypically male traits are idealized as the masculine cultural ideal, explaining how and why men maintain dominant social roles over women and other groups considered to be feminine [4]. Sexual decisions in couples, in particular, are uniformly attributed to men.

A systematic review of these matters shows that there are several ways in which violence makes women more vulnerable to this disease: on the one hand, marital rape by an HIV-positive partner is expected to expose them to the virus; wherever abuse is present, women lack agency in asking for safe sexual practices; revictimization processes put women in vicious cycles of risky sexual practices [15].

As far as we know, in the literature, IPV and HIV are mostly studied via qualitative studies or ad-hoc surveys – or in the clinical environment, which, however, goes beyond the scope of our work. This

[1] Corresponding author: Micaela Arcaio, Department of Psychology, Educational Science and Human Movement, University of Palermo. Micaela Arcaio was supported by the National Operational Programme on Research and Innovation 2014-2020.

[2] The United Nations define intimate partner violence or domestic violence as a behavioural pattern of power and control over an intimate partner [19].

study is the first step in the ongoing project to analyse the relationship between domestic violence and HIV using the DHS survey from African countries, which collects data on several topics such as socio-demographic characteristics, family planning, reproductive health, HIV, and domestic violence. The analysis will be carried out through the lens of ecological models, i.e., accounting for social and contextual factors, as well as personal characteristics that may be associated with HIV.

In this work, Section 2 highlights the current literature on HIV and intimate partner violence. In Section 3 we briefly discuss the data and methods used for this analysis. Sections 4 and 5 respectively contain the results of the analysis and a brief discussion of the results with conclusions.

## 2. Intimate Partner Violence and HIV: the contextualisation of disease

The contextualization of risk factors and the way the social context mediates the relationship between risk factors and disease have been widely discussed in their analysis [9]. When it comes to HIV, gender inequality plays a key role in the development of the infection, especially in Sub-Saharan Africa, where the phenomenon is so relevant. Indeed, in 2021, women accounted for 63% of new HIV infections in this region, with a slower decline in their incidence trend than in men's [22]. Moreover, while in the rest of the world 94% of new infections in 2021 were imputable to key populations[3], in Sub-Saharan Africa, almost half of the new infections are imputable to the general population [22].

Gender dynamics that favour men over women highlight the vulnerability of women and girls to the virus, with young women in Sub-Saharan Africa three times more likely to be HIV-positive than men their age [2]. A cross-sectional study conducted in South Africa in 2001 [6] shows that both physical and sexual partner violence are associated with increased odds of HIV infection, while a longitudinal study based, once again, in South Africa confirms that power inequity inside couples and intimate partner violence contribute to HIV infection in young women [12].

In this framework, ecological models have been proposed both for HIV and IPV to highlight the complex interplay between social and individual characteristics of both phenomena. These models move from societal and community-level factors to relationship and individual characteristics that are associated with the presence of HIV and IPV [3].

At the societal level, as women's empowerment grows, one might expect that women will also experience less abuse from intimate partners since gender equality is supposed to improve their conditions overall [11]. However, the opposite can be true as well: if men perceive female empowerment as a threat, they respond with increased levels of violence [16]. Gender inequality and violence against women and girls have also been shown to exacerbate the risk of HIV infection [6,12].

At the community level, violence seems to be associated with poverty and living in rural areas [13]. Moreover, the perpetration of violence itself is facilitated by the general acceptance of violence within couples [8]. On the contrary, rural areas have been shown to be characterised by a lower prevalence of HIV than urban areas [18]. Furthermore, while the highest HIV prevalence rates are found in low-income countries, more often than not, it is the richest population strata that are the most impacted by this virus [17]. This occurrence is, however, context-specific: the contrary is, indeed, found in rich countries, and it is due to several factors such as recurrence to sex workers [6].

Relationship and individual factors refer to gender dynamics within couples as well as their socio-economic characteristics. People with a higher education are less likely to either perpetrate or suffer IPV [1, 14] and to be affected by HIV [10].

The association between socioeconomic context and gender also plays a key role in the study of HIV. Most of the characteristics of the association between HIV and context here presented, indeed, are amplified in women. Specifically, sex work might become a necessity to women living in poverty, thus exposing them further to HIV infection, in particular for those living in substandard housing in urban environments, in case of unemployment, and when they have a history of intimate partner violence [6, 9].

[3] Key populations are considered to be particularly vulnerable to HIV infection. They include: sex workers; clients of sex workers and sex partners of key populations; people who inject drugs; gay men and other men who have sex with men; transgender women. These populations account for less than 5% of the global population, however, they are disproportionately affected by HIV – 70% of all new global infections were found among these groups in 2021 [21].

DRAFT

## 3. Data and methods

HIV and intimate partner violence are investigated using data from the Demographic and Health Survey (DHS) programme. The DHS is a nationally representative household survey, employed for the collection and analysis of harmonised demographic and health data [5]. Women of reproductive age (aged 15-49) are, indeed, the main subject of this survey.

The focus of this work is the "HIV dataset", which contains the result of ELISA (enzyme-linked immunosorbent assay) tests to determine respondents' serostatus, and the domestic violence module within the questionnaire administered to women. To collect data on these matters, in all households in which men are also included in the survey, blood samples are collected from those who volunteer for HIV testing; furthermore, the domestic violence module is administered to randomly selected ever-partnered women in heterosexual relationships already involved in the survey. Thus, the statistical units in this study are currently-married (or cohabiting) women, aged 15-49, whose serostatus and current experiences of abuse are known.

In this study, we included all the surveys that carried both the HIV dataset and the domestic violence module in the years between 2016 and 2019. Thus, four surveys from Sub-Saharan Africa were included: Burundi 2016-2017, Cameroon 2018, Sierra Leone 2019, and Zambia 2018. The sample, thus, consists of 10,532 currently partnered women, with an average age of 31 years (SD = 8.06).

Country-level estimates are weighted using DHS guidelines and specific cluster weights for HIV prevalence and the proportion of women who have been abused at least once by their current partners.

Table 1: Descriptive statistics

| | **Burundi** | **Cameroon** | **Sierra Leone** | **Zambia** | **Total** |
|---|---|---|---|---|---|
| Number of respondents | 2,058 | 1,059 | 2,464 | 4,951 | 10,532 |
| HIV Prevalence (%) | 0.65 | 3.03 | 1.89 | 13.96 | 7.38 |
| Victims of domestic violence (%) | 48.7 | 43.06 | 60.01 | 43.83 | 48.37 |

The aim of this analysis is to evaluate the relationship between HIV and IPV, controlling for other relevant socio-demographic characteristics. Because of the binary nature of the response variable – HIV status of respondents – the model here used is a logistic regression model. However, to account for the rarity of being HIV-positive in the sample, the Firth adjustment is here used [6]. Using this method, the log-likelihood is penalized with one-half of the logarithm of the determinant of the information matrix, thus reducing the bias seen in generalized linear models.

Domestic violence is assessed via a composite indicator, created using a confirmative factor analysis (CFA), using three binary variables. Indeed, the module on domestic violence primarily assesses IPV via three indicators:

- "Physical violence", referring to being pushed, shook, slapped, punched, or threatened at gunpoint by her partner;
- "Emotional Violence", involving humiliation, threats of physical harm or insults;
- "Sexual Violence", indicating marital rape by the respondent's partner.

The graphical representation of this (reflective) measurement model[4] – with the relative coefficients – can be found in Figure 1.

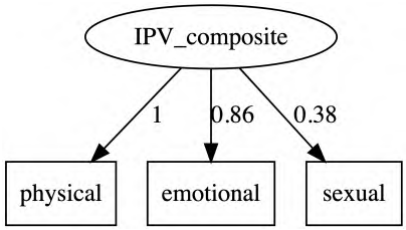



Figure 1: Path diagram of the CFA

[4] Cronbach's alpha = 0.61 $CI_{95\%}$: (0.597-0.624)

DRAFT

A minimum-maximum transformation of the post-estimation results of the CFA will be used within the selected model.

Other covariates in the model are divided into four domains, that reflect the structure of the ecological model:

- *Respondents' characteristics*: the number of justifications respondents give to men using physical violence on their wives; age at first sexual encounter;
- *Household characteristics:* age of respondent; whether the respondent has any children; employment status of respondent; wealth index (categorizing respondents in "poor", "middle", "rich"); the respondent and her husband's educational level ("no education", "primary", "secondary", "higher");
- *Community:* living in rural areas;
- *Context:* number of decisions the respondents take part in in their households; Country dummy variables - present also to account for the data being pooled over four different surveys from different countries.

## 4. Results

The result of the estimations of the selected model can be found in Table 2. A likelihood ratio test was also employed to check whether the composite indicator of violence is relevant in the fit of the model: given that LRT (1 d.f.) = 3.97 (p-value = 0.04), the model that also includes IPV fits significantly better than the model without it. The approximation of the LRT statistic with the Wald test – of the selected model vs. the null model – returns Wald (19 d.f.) = 654.45 (p-value ≈ 0), thus pointing to an overall good fit of the model.

Table 2: Model estimates

| Domain | Covariate | | Odds Ratio |
|---|---|---|---|
| Respondents' characteristics | IPV composite | | 1.002** |
| | No. justifications to wife beating | | 0.943** |
| | Age at first intercourse | | 0.897*** |
| | Age | | 1.056*** |
| Household characteristics | Respondent has children | | 0.684* |
| | Respondent is employed | | 0.978 |
| | Wealth index (ref. poor) | | |
| | | Middle | 1.347** |
| | | Rich | 1.955*** |
| | Husband's education level (ref. No education) | | |
| | | Primary | 0.950 |
| | | Secondary | 1.512** |
| | | Higher | 1.274 |
| | Respondent's education level (ref. No education) | | |
| | | Primary | 1.187 |
| | | Secondary | 1.263 |
| | | Higher | 0.874 |
| Community | Living in rural areas (vs. urban) | | 0.544*** |
| Societal level | No. decisions she takes in her HH | | 1.077** |
| | Country (ref. Burundi) | | |
| | | Cameroon | 2.128** |
| | | Sierra | 1.259 |
| | | Zambia | 8.283*** |
| | Constant | | 0.0141*** |

*** p < 0.01, ** p < 0.05, * p < 0.1

DRAFT

The IPV composite indicator resulted significantly and positively associated with the HIV status of respondents, even though the variation measured by the odds' size is small (OR = 1.002). However, when we consider the number of justifications respondents give to men using physical violence on their wives, each additional justification decreases the odds of being HIV positive by almost 6%. Age at first intercourse is also significantly associated with HIV: indeed, for each year of delay in age at first sex the odds of being HIV positive decrease by more than 10%.

Household characteristics also matter. Wealth, as stated in the literature, is positively associated with HIV: as the household of respondents gets richer, their odds to be HIV positive also increase. However, neither the respondent's nor her husband's education seem to be significantly associated with her HIV status. Furthermore, while being employed is not associated with respondents' serostatus, having children is negatively associated with their mothers being HIV positive, given the access to testing and anti-retroviral therapy once they enter antenatal care.

The community and the context in which respondents live are also associated with their HIV status. As it has been shown in the literature, living away from the city seems to expose respondents less to the HIV virus: people who live in rural areas are, indeed, 54% less likely to be HIV-positive. More power in making decisions within the household – here taken as a proxy of empowerment overall – is however significantly and positively associated with HIV. The coefficients for the Country dummy variables reflect the general condition of their resident populations with respect to HIV prevalence.

## 5. Conclusions

This work is the first step in our ongoing project to analyse the relationship between domestic violence and HIV using survey data from African countries. Our analysis has confirmed some of the results highlighted in the literature – being abused by the current partner makes women more likely to be HIV-positive. This result becomes especially relevant given that while the literature on HIV mostly focuses on physical and sexual violence separately, the literature on domestic violence is progressively moving towards the concept of polyvictimisation – abuse calls for other forms of abuse, thus reinforcing and facilitating the path from one form to another in victims [19].

However, some results from the model deserve further investigation: indeed, there seems to be a twisted association between decision-making power and the number of justifications given for physical violence, and HIV.

This work certainly does not lack limitations. Specifically, given the preliminary stage of this work, alternative models could be chosen for the analysis of the matter at hand – e.g. mixed-effects logistic regression models to account for the nature of the pooled data and Poisson regression models, that are often used in the literature to evaluate HIV. Further work can also be done on the choice of the covariates in the model and introducing interactions between relevant variables.

But this work also offers many possibilities for further development. Mainly, we want to evaluate the hypotheses according to which the context of residence may be relevant to women's likelihood to be HIV positive and of being victims of IPV; then, that IPV can be a relevant factor in the diffusion of HIV at a sub-national level. This will allow us to further check for the interplay between contextual and individual characteristics of respondents, always in the framework of the ecological model already employed in the analysis.

DRAFT

## References

[1] Abrahams, N., Jewkes, R., Laubscher, R., Hoffman, M.: Intimate partner violence: Prevalence and risk factors for men in Cape Town, South Africa. (2006). Violence Vict., 21(2), 247-264.

[2] AIDSinfo | UNAIDS. Aidsinfo.unaids.org. (2023) Available via: https://aidsinfo.unaids.org. Cited 14 Feb 2023.

[3] Centers for Disease Control and Prevention. The social-ecological model: A framework for prevention. (2015)

[4] Connell, R. W., & Messerschmidt, J. W.: Hegemonic masculinity: Rethinking the concept. (2005). Gender & society, **19**(6), 829-859.
[5] Croft, T. N., Marshall, A., & Allen, C.: Guide to DHS - DHS-7. (2018). DHS
[6] Dunkle K. L., Jewkes R.K., Brown H. C., Gray G. E., McIntryre J. A., Harlow S.D.: Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa. (2004) The Lancet **363**, 1415—1421
[7] Firth, D.: Bias reduction of maximum likelihood estimates. (1993). Biometrika, **80**(1), 27-38.
[8] Gage, A. J.: Women's experience of intimate partner violence in Haiti. (2005). Soc. Sci. Med., 61(2), 343—364.
[9] Gillespie, S., Kadiyala, S., & Greener, R.: Is poverty or wealth driving HIV transmission? (2007) Aids, **21**, S5–S16
[10] Hargreaves, J. R., Bonell, C. P., Boler, T., Boccia, D., Birdthistle, I., Fletcher, A., Glynn, J. R.: Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in Sub-Saharan Africa. (2008). Aids, **22**(3), 403–414
[11] Heirigs, M., Moore, M.: Gender inequality and homicide: a cross-national examination. (2017). Int. J. Comp. Appl. Crim. Justice, **42**(4), 273--285. doi: 10.1080/01924036.2017.1322112
[12] Jewkes R.K., Dunkle K. L., Nduna M., Shai N.: Intimate partner violence, relationship power inequity, and incidence of HIV infection in young women in South Africa: a cohort study. (2010) The Lancet **376**, 41--48
[13] Jeyaseelan, L., Kumar, S., Neelakantan, N., Peedicayil, A., Pillai, R., & Duvvury, N.: Physical spousal violence against women in India: some risk factors. (2007). J Biosoc. Sci., **39**(5), 657-670.
[14] Kishor, S., & Johnson, K.: Reproductive health and domestic violence: Are the poorest women uniquely disadvantaged? (2006). Demogr., **43**(2), 293-307.
[15] Maman, S., Campbell, J., Sweat, M. D., Gielen, A. C.: The intersections of HIV and violence: directions for future research and interventions. (2000). Soc. Sci. Med., **50**(4), 459--478.
[16] Meinck, F., Cluver, L. D., Boyes, M. E., & Mhlongo, E. L.: Risk and protective factors for physical and sexual abuse of children and adolescents in Africa: A review and implications for Practice. (2015). Trauma Violence Abus., **16**, 81–107.
[17] Mishra, V., Bignami-Van Assche, S., Greener, R., Vaessen, M., Hong, R., Ghys, P.D., Boerma, J.T., Van Assche, A., Khan, S. and Rutstein, S.: HIV infection does not disproportionately affect the poorer in sub-Saharan Africa. (2007). Aids, **21**, S17-S28.
[18] Nutor, J. J., Duah, H. O., Agbadi, P., Duodu, P. A., & Gondwe, K. W.: Spatial analysis of factors associated with HIV infection in Malawi: indicators for effective prevention. (2020). BMC Public Health, **20**(1), 1–14.
[19] Okumu, M., Orwenyo, E., Nyoni, T., Mengo, C., Steiner, J. J., & Tonui, B. C.: Socioeconomic factors and patterns of intimate partner violence among ever-married women in Uganda: pathways and actions for multicomponent violence prevention strategies. (2021). J. Interpers. violence.
[20] UN: What Is Domestic Abuse?. (2023). Available via: https://www.un.org/en/coronavirus/what-is-domestic-abuse Cited 14 Feb 2023
[21] UNAIDS: Dangerous inequalities: World AIDS day report 2022. (2022). Available via: https://www.unaids.org/en/resources/documents/2022/dangerous-inequalities. Cited 11 Feb 2023
[22] UNAIDS: Executive summary - in Danger: Unaids global aids update 2022. (2022). Available via: https://www.unaids.org/en/resources/documents/2022/in-danger-global-aids-update-summary. Cited 11 Feb 2023
[23] World Health Organization. Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women. (2021)

DRAFT

# An extension of finite mixtures of latent trait analyzers for biclustering bipartite networks

Dalila Failli[a], Maria Francesca Marino[a], and Francesca Martella[b]

[a]Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 59 - 50134 Firenze; dalila.failli@unifi.it, mariafrancesca.marino@unifi.it

[b]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Roma; francesca.martella@uniroma1.it

**Abstract**

Network data analysis has received increasing attention recently. Bipartite networks represent a specific type of network data describing the relationships between disjoint sets of nodes, called sending and receiving nodes. We extend the Mixture of Latent Trait Analyzers (MLTA) specifically tailored for the analysis of bipartite networks to achieve a twofold goal. First, the aim is to perform a joint clustering of sending and receiving nodes, thus partitioning the data matrix into homogeneous blocks, as in the biclustering approach. In addition, a latent trait is used to model the dependence between receiving nodes, as in the latent trait framework. The proposal also admits the inclusion of nodal attributes on the latent layer of the model to understand how they affect cluster formation. An EM algorithm with Gauss Hermite approximation is proposed to estimate the model parameters.

***Keywords:*** Model-based clustering, Network data, Two-mode networks, Nodal attributes, EM algorithm

DRAFT

## 1. Introduction

Over the years, many social, technological, and biological processes have been represented as networks. These are collections of interconnected units (nodes) that can capture interactions within a system. In this context, bipartite networks are a special type of networks that represent the relationships between two disjoint sets of nodes, formally called sending and receiving nodes. A primary characteristic of this type of network is that connections exist only between nodes belonging to different sets, as illustrated in Figure 1.

A relevant aspect of network analysis concerns the simultaneous clustering of sending and receiving nodes aiming at partitioning the data matrix into homogeneous blocks, called biclusters. An example of a block structure is shown in Figure 2, where rows (sending nodes) and columns (receiving nodes) of the data matrix are reordered according to the corresponding class membership, thus returning blocks of sending nodes that connect similarly with subsets of receiving nodes.

A common example of application concerns the field of genetics, where the biclustering approach can be used to identify groups of genes which are co-expressed under subsets of experimental conditions.

Different biclustering approaches are available in the literature, such as the model-based ones (12; 17; 2; 14). In this specific context, several methods based on finite mixtures have been proposed (8; 9; 19; 11; 13; 18; 15).

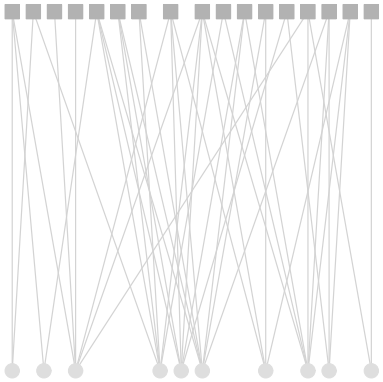


Figure 1: Example of bipartite networks

We start from the MLTA model, introduced by (6; 7). Here, the aim is of clustering sending nodes via a finite mixture specification, while accounting for the dependence between receiving nodes via a continuous latent variable, as in the latent trait framework. Our proposal is to modify the MLTA in two ways. First, allowing for a joint clustering of sending and receiving nodes, where sending nodes are partitioned into clusters called components and, in each of them, receiving nodes are partitioned into clusters called segments. Furthermore, we also allow for the inclusion of nodal attributes on the latent layer in order to understand how they influence component formation.

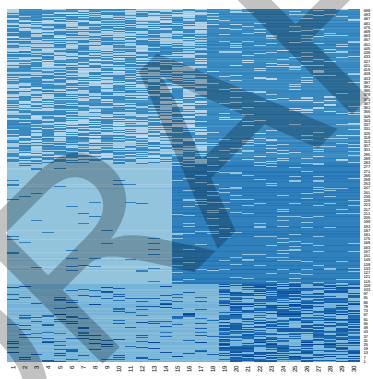


Figure 2: Example of block structure

DRAFT

The paper is organized as follows: in Section 2. we extend the MLTA model, also describing model assumptions, parameter estimation, and model selection. Section 3. shows the results of a simulation study conducted in order to verify the efficacy of the proposed approach. Section 4. contains concluding remarks and details further extensions of the approach.

## 2. Mixture of latent trait analyzers

Let $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$ denote the set of sending nodes and $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$ the set of receiving nodes. In this framework, bipartite networks can be formally described by a random incidence matrix $\boldsymbol{Y} = \{Y_{ik}\}$, with elements

$$Y_{ik} = \begin{cases} 1 \text{ if sending node } n_i \text{ is connected with receiving node } r_k, \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

To obtain a clearer picture of the data at hand, (7) propose to extend the Mixture of Latent Trait Analyzers (MLTA) (6) in the context of bipartite networks. The model combines latent class and latent trait analysis

by assuming that the set of $N$ sending nodes can be divided into $G$ distinct classes (or groups) and that the propensity of each sending node to be connected with the $R$ receiving nodes depends also on a multidimensional continuous latent trait. Our contribution is to further extend the MLTA model by performing a joint clustering of sending and receiving nodes, also taking into account nodal attributes in the latent model structure.

## 2.1 The MLTA model

The MLTA model assumes that every sending node belongs to one of $G$ unobserved groups identified by the latent random variable $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})' \sim \text{Multinomial}(1, (\eta_1, \ldots, \eta_G))$, whose generic element is

$$z_{ig} = \begin{cases} 1 \text{ if sending node } n_i \text{ belongs to group } g, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

The parameter $\eta_g$ denotes the probability that a randomly selected sending node belongs to group $g$, with $g = 1, \ldots, G$, under the constraints that $\sum_{g=1}^{G} \eta_g = 1$ and $\eta_g \geq 0$, $g = 1, \ldots, G$.
Furthermore, the model assumes the existence of a $D$-dimensional continuous latent trait $\boldsymbol{u}_i$ distributed according to a Gaussian density with null mean vector and identity covariance matrix, i.e. $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, which captures the heterogeneity of the connections between sending and receiving nodes. Thus, response variables contained in the $\boldsymbol{y}_i$ vector are assumed to be independent Bernoulli random variables with parameters $\pi_{gk}(\boldsymbol{u}_i)$, $k = 1, \ldots, R$, modelled through the following logistic function:

$$\pi_{gk}(\boldsymbol{u}_i) = p(y_{ik} = 1 \mid \boldsymbol{u}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-(b_{gk} + \boldsymbol{w}'_{gk}\boldsymbol{u}_i)]}. \tag{3}$$

Here, the parameter $b_{gk}$ represents the attractiveness of the $k$-th receiving node for sending nodes belonging to group $g$, while $\boldsymbol{w}_{gk}$ represents the influence of the latent trait $\boldsymbol{u}_i$ on the probability of a connection between sending nodes belonging to the $g$-th group and receiving node $r_k$.

## 2.2 Extending the MLTA model

To perform a joint clustering of sending and receiving nodes, we follow an approach similar to that proposed by (15) and modify the logistic function in (3) as:

$$\pi_{gk}(\boldsymbol{u}_i) = p(y_{ik} = 1 \mid \boldsymbol{u}_i, z_{ig} = 1) = \frac{1}{1 + \exp[-(b_g + \boldsymbol{a}'_{gk}(\boldsymbol{\mu} + \boldsymbol{u}_i)]}. \tag{4}$$

Here, $b_g$ is a component-specific latent effect, $\boldsymbol{\mu}$ is a $D$-dimensional vector of fixed effects, and $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$ is a $D$-dimensional continuous latent trait capturing the residual heterogeneity of connections between sending nodes belonging to the $g$-th component and receiving nodes belonging to the $d$-th segment. Moreover, $\boldsymbol{a}_{gk}$ is a $D$-dimensional row stochastic vector ($D \leq R$) with

$$a_{gkd} = \begin{cases} 1 \text{ if receiving node } r_k \text{ belongs to segment } d, \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

This allows us to select the membership of the $k$-th receiving node to one of the $D$ segments for those sending nodes belonging to component $g$.
Following the strategy adopted in (5), we account for the effect that nodal attributes may have on group membership by letting the parameter $\eta_g$ vary across sending nodes. This is done by considering a latent class regression model based on the vector of nodal attributes $\boldsymbol{x}_i$, as follows:

$$\eta_{ig} = \frac{\exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_g\}}{1 + \sum_{g'=2}^{G} \exp\{\boldsymbol{x}'_i\boldsymbol{\beta}_{g'}\}}, \quad g = 2, \ldots, G, \tag{6}$$

where $\boldsymbol{\beta}_g$ denotes the model coefficient vector for the $g$-th group.

DRAFT

## 2.3 Parameter estimation

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_G, b_1, \ldots, b_G, \boldsymbol{a}_{11}, \ldots, \boldsymbol{a}_{GR}, \mu_1, \ldots, \mu_D)$ represent the vector of all free model parameters. Given the assumptions described in the previous section, the log-likelihood function of the model can be written as:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \Big( \sum_{g=1}^{G} \eta_{ig} \int \prod_{k=1}^{R} p(y_{ik} \mid \boldsymbol{u}_i, z_{ig} = 1) p(\boldsymbol{u}_i) d\boldsymbol{u}_i \Big), \tag{7}$$

where $p(y_{ik} \mid \boldsymbol{u}_i, z_{ig} = 1) = (\pi_{gk}(\boldsymbol{u}_i))^{y_{ik}} (1 - \pi_{gk}(\boldsymbol{u}_i))^{1-y_{ik}}$. The integral to be solved in equation (7) cannot be computed analytically, therefore an EM algorithm with a Gauss-Hermite approximation of the log-likelihood function is proposed. In detail, after the initialization of model parameters and the approximation of the intractable integral with Gauss-Hermite quadrature, the parameters $b_g$ and $\mu_d$ are updated via a standard Newton-Raphson algorithm with augmented data, while the $D$-dimensional row stochastic vector $\boldsymbol{a}_{gk}$ is updated via a classification step, following a similar strategy to that proposed by (15). Finally, parameters $\boldsymbol{\beta}_g$ are estimated via a Newton-Raphson step and $\eta_{ig}$ is updated accordingly.
At convergence, each sending node can be assigned to the $g$-th component via a Maximum a Posteriori (MAP) rule and each receiving node can be assigned to the $d$-th segment according to the vector $\boldsymbol{a}_{gk}$.

## 2.4 Standard errors and model selection

To evaluate the standard errors of the estimates obtained with the EM algorithm, several methods are available, such as the jackknife method (4; 6). Given an incidence matrix $\boldsymbol{Y}$ with $N$ sending nodes and $R$ receiving nodes, this method consists in extracting $N$ samples of size $(N-1) \times R$, obtained by removing one sending node at a time from the original data matrix. However, we found that a more efficient strategy for deriving standard errors relies on the use of a non-parametric bootstrap (3), which consists in extracting with repetition $N$ rows of the incidence matrix, so that each sending node can appear multiple times.
Since the number of components $G$ and the number of segments $D$ are considered as fixed quantities, it is possible to estimate the model for different values of $G$ and $D$, then selecting the optimal model as the one corresponding to the smallest value of the chosen information criterion, such as the Bayesian Information Criterion (BIC) (16) or the Akaike's Information Criterion (AIC) (1).

DRAFT

# 3. Simulation study

The performance of the model in terms of parameters' recovery and clustering is evaluated through a simulation study with a different number of nodes, as described below.

## 3.1 Simulation setup

We simulated 100 samples in three different scenarios based on a varying number of sending nodes ($N = 500,\ N = 1000,\ N = 2000$), while the number of receiving nodes $R$ is kept constant and equal to 30. Furthermore, we considered a fixed number of segments $D = 2$ and components $G = 3$. As regards the latent class variable, block membership is defined via a single nodal attribute $x_i$ which is drawn from a Gaussian distribution with mean and variance equal to 1. In each scenario, a multi-start strategy based on 100 random starts is adopted.

## 3.2 Simulation study: clustering recovery

The ability of the proposal in correctly classifying sending and receiving nodes is evaluated through the Adjusted Rand Index (ARI) (10). The results are shown in Table 1. Looking at these results, we note that, for $N$ greater than 500, the classification is good for both rows (sending nodes) and columns

(receiving nodes), and the Adjusted Rand Index remains stable. The variability in the estimates is due to the intrinsic variability of simulations.

Table 1: Distribution across samples of the Adjusted Rand Index for varying $N$

| | | Adjusted Rand Index | | |
|---|---|---|---|---|
| | | 1st Qu. | Median | 3rd Qu. |
| $N$=500 | Row | 0.9259 | 0.9802 | 0.9944 |
| | Col. | 0.1693 | 1.0000 | 1.0000 |
| $N$=1000 | Row | 0.9369 | 0.9871 | 0.9966 |
| | Col. | 0.1788 | 1.0000 | 1.0000 |
| $N$=2000 | Row | 0.9461 | 0.9823 | 0.9943 |
| | Col. | 0.3599 | 1.0000 | 1.0000 |

## 3.3 Simulation study: latent class parameters

Figure 3 shows the distributions of $\beta_g$ parameters across samples, for different values of $N$. Looking at these figures, it is evident that variability reduces as the size of the network increases. Furthermore, as $N$ increases, we are increasingly able to identify the true values of model parameters.

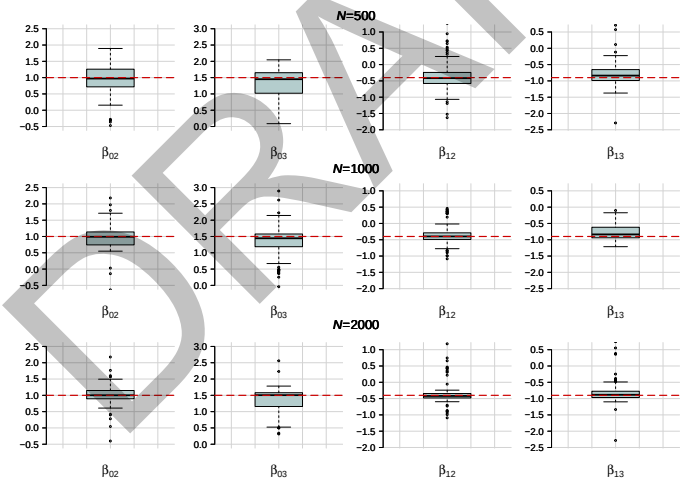



Figure 3: Distribution across samples of $\beta_g$ parameters for varying $N$

## 4. Conclusions

The mixture of latent trait analyzers is modified to achieve a twofold objective for the analysis of bipartite networks: i) performing a joint clustering of sending and receiving nodes; ii) including nodal attributes to study how nodes' characteristics influence the component membership probability.
The simulation study shows that the model can be effectively employed for biclustering bipartite networks. In detail, when the number of sending nodes is large, the variability of estimates is reduced and the classification of nodes is good. However, the simulation study needs to be further extended by letting the number of partitions and receiving nodes vary, as well as reducing the number of sending nodes to investigate how classification performs in smaller networks.

DRAFT

A further development may involve the application of the proposal for the analysis of a large real-world data set, such as a gene-experimental condition network.

## References


[1] Akaike, H.: A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. **19**, 716–23 (1974)

[2] Dhollander, T., Sheng, Q., Lemmens, K., De Moor, B., Marchal, K., Moreau, Y.: Query-driven module discovery in microarray data. Bioinformatics. **23**, 2573–2580 (2007)

[3] Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. **7**, 1–26 (1979)

[4] Efron, B.: Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika. **68** 589–599 (1981)

[5] Failli, D., Marino, M.F., Martella, F.: Extending finite mixtures of latent trait analyzers for bipartite networks. In: Balzanella A., Bini M., Cavicchia C. and Verde R. (eds.) Book of short Paper SIS 2022, pp. 540–550. Pearson (2022)

[6] Gollini, I., Murphy, T.B.: Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. **24**, 569–588 (2014)

[7] Gollini, I.: A mixture model approach for clustering bipartite networks. In: Ragozini, G., Vitale, M.P. (eds.) Challenges in Social Network Research: Methods and Applications, pp. 79–91. Springer International Publishing (2020)

[8] Govaert, G., Nadif, M.: Clustering with block mixture models. Pattern Recognit. **36(2)**, 463–473 (2003)

[9] Govaert, G., Nadif, M.: Block clustering with Bernoulli mixture models: comparison of different approaches. Comput. Statist. Data Anal. **52**, 3233–3245 (2008)

[10] Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

[11] Keribin, C., Brault, V., Celeux, G., et al.: Estimation and selection for the latent block model on categorical data. Stat. Comput. **25**, 1201–1216 (2014)

[12] Lazzeroni, L., Owen, A.B.: Plaid models for gene expression data. Statist. Sinica. **12**, 61–86 (2002)

[13] Martella, F., Alfò M., Vichi, M.: Biclustering of gene expression data by an extension of mixtures of factor analyzers. The Int. J. Biostat. (2008) doi:10.2202/1557-4679.1078

[14] Martella, F., Vichi, M.: Clustering microarray data using model-based double K -means. J. Appl. Stat. **39(9)**, 1853–1869 (2012)

[15] Martella, F., Alfò, M.: A finite mixture approach to joint clustering of individuals and multivariate discrete outcomes. J. Stat. Comput. Simul. 87:11, 2186–2206 (2017)

[16] Schwarz, G.: Estimating the Dimension of a Model. Ann. Stat. **6**, 461–464 (1978)

[17] Sheng, Q., Moreau, Y., De Moor, B.: Biclustering microarray data by Gibbs sampling. Bioinformatics. **19**, 196–205 (2003)

[18] Vicari, D., Alfò, M.: Model based clustering of customer choice data. Comput. Statist. Data Anal. **71**, 3–13 (2014)

[19] Wyse, J., Friel, N.: Block clustering with collapsed latent block models. Stat. Comput. **22**, 415–428 (2012)


DRAFT

# Constrained Mixtures of Generalized Normal Distributions

Pierdomenico Duttilo[a], Alfred Kume[b], and Stefano Antonio Gattone[c]

[a]University "G. d'Annunzio" of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy; pierdomenico.duttilo@unich.it
[b]School of Mathematics, Statistics and Actuarial Sciences, University of Kent, Canterbury CT2 7FS, UK; a.kume@kent.ac.uk
[c]DISFIPEQ, University "G. d'Annunzio" of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy; gattone@unich.it

DRAFT

**Abstract**

In this work, constrained univariate mixtures of generalized normal distributions (CMGND) are introduced. Specifically, mixture parameters are constrained to be equal across mixture components. The expectation conditional maximization (ECM) algorithm is used to estimate the constrained parameters via the maximum likelihood estimation (MLE). In addition, the iterative Newton-Raphson method is applied to handle the non-linear iteration equations of the parameters during the estimation stage. Next, a simulation is performed to assess the parameter estimation performance for a two-component CMGND with the same scale and shape parameters, i.e. with the same variance and kurtosis. Simulation results show that the estimation accuracy of the constrained mixture is higher than the unconstrained mixture.

***Keywords:*** Constrained mixtures of generalized normal distributions, ECM algorithm, Maximum likelihood estimation, Newton-Raphson method

## 1. Introduction

Over time, non-normal mixture distributions have gained increasing attention to analyse datasets characterized by non-normal features like skewness and heavy tails (10).

Among the statistical distributions available in the literature, the generalized normal distribution (GND) is able to model a large variety of statistical behaviours thanks to the additional shape parameter which controls the tail weights (14). Then, finite mixtures of generalized normal distributions (MGND) have the flexibility to fit non-normal data (16).

MGND have been successfully applied in signal processing, computer vision, pattern recognition and other recent statistical tasks that require mixture estimation (13).

Bazi et al. (2006) applied univariate MGND for image processing (5). The estimation of the parameters was performed via the maximum likelihood estimation (MLE), and the expectation maximization (EM) algorithm. Allili (2012) used the univariate MGND for wavelet representation (1). Parameters have been estimated with a Bayesian method which optimizes a minimum message length objective, and the EM algorithm. Nguyen et al. (2014) proposed a univariate bounded generalized Gaussian mixture model defining a bounded support region for each component (16). Recently, Wen et al. (2022) studied a univariate two-component MGND and proposed an expectation conditional maximization (ECM) algorithm for parameter estimation (18).

Mixture distributions with unconstrained parameters may have some problem in the estimation phase. Firstly, in normal mixtures it is well known that when parameters are not restricted the resulting likelihood from a sample is unbounded, *"no maximum likelihood estimator exists in the unconstrained problem"* (7). Thus, it is possible to observe this problem also in MGND, since the GND is a *"natural generalization of the normal distribution"* (14). Secondly, the number of parameters increases with the number of the mixture components and the estimation could result computationally problematic.

As a consequence, different methods have been proposed to overcome these critical issues[1]. These methods can be divided into two main approaches: linear constraints methods, and eigenvalue decomposition methods. The former impose linear restrictions on the mixture parameters. By contrast, the latter exploit the eigenvalue decomposition of the component covariance matrices to impose constraints. Mainly these methods have been applied to constrain mixtures of normals (4; 6; 8; 9; 15; 17) and Student-t (2; 3).

To the best of our knowledge none of the existing studies propose a constrained estimation of the univariate MGND. We aim to fill this gap by proposing constrained univariate mixtures of generalized normal distributions (CMGND) where the parameters are constrained to be equal across mixture components. The ECM algorithm is used to estimate constrained parameters via the MLE together with the Newton-Raphson method. Next, a simulation is performed to assess the parameter estimation performance for a two-component CMGND with the same scale and shape parameters.

The rest of the paper is organized as follows. Section 2. illustrates the methodology. Section 3. illustrates the simulation. Finally, Section 4. provides some conclusions.

DRAFT

## 2. Methodology

A univariate finite MGND is given by the marginal distribution of the random variable $X$

$$f(x|\theta) = \sum_{k=1}^{K} \pi_k p_k(x|\mu_k, \sigma_k, \nu_k), \tag{1}$$

where:

- $\theta = \{\pi_k, \mu_k, \sigma_k, \nu_k\}$, $k = 1, .., K$;
- $K$ is the number of mixture components;
- $\pi_k$ is the $k$-th mixture weight which satisfies $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$;
- $p_k(x|\mu_k, \sigma_k, \nu_k)$ is the $k$-th probability density function of the generalized normal distribution (GND), which is defined as follows

$$p_k(x) = \frac{\nu_k}{2\sigma_k \Gamma(1 \,/\, \nu_k)} \exp\left\{ - \left| \frac{x - \mu_k}{\sigma_k} \right|^{\nu_k} \right\} \qquad \text{with } \Gamma(1 \,/\, \nu_k) = \int_0^{\infty} t^{1/\nu_k - 1} \exp^{-t} dt, \tag{2}$$

  where $\mu_k$ is the $k$-th location parameter ($\mu_k \in$ R), $\sigma_k$ is the $k$-th scale parameter ($\sigma_k > 0$), and $\nu_k$ is the $k$-th shape parameter ($\nu_k > 0$).

It is possible to capture a wide range of statistical distributions by varying the shape parameter $\nu_k$ who determines the tail weights (See Figure 1). The normal distribution is yielded with $\nu_k = 2$, whereas the Laplace distribution is yielded with $\nu_k = 1$. It is noticed that $1 < \nu_k < 2$ yields an "intermediate distribution" between the normal and the Laplace distribution. As limit cases, for $\nu_k \to +\infty$ the distribution tends to a uniform distribution, while for $\nu_k \to 0$ it will be impulsive (5; 13; 18).

[1] (11) and (7) give a more detailed account of what has been done so far.

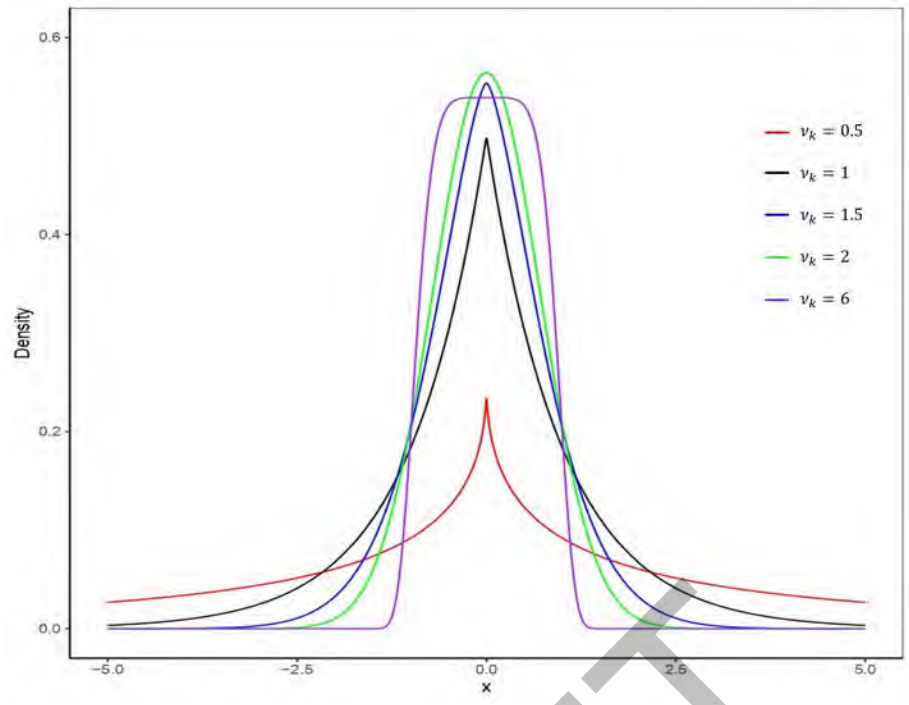



Figure 1: $k$-th probability density function for $\mu_k = 0$, $\sigma_k = 1$ and different shape values.

Constraints are imposed on $\mu_k$, $\sigma_k$ and $\nu_k$ to be equal across the mixture components: $\mu_k = \mu$, $\sigma_k = \sigma$, $\nu_k = \nu$, for $k = 1, ..., K$. Thus, taking all possible combinations of these constraints into consideration would result in a 8-model family[2]. For identifiability purposes, we need to impose that the mixture weights must be different to each other.

Following (18), the ECM algorithm (12) is applied to perform parameter estimation of the CMGND. From Eq. 1 the log-likelihood function is given by

$$\log L(\theta) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \frac{\nu_k}{2\sigma_k \Gamma(1/\nu_k)} \exp\left\{ -\left| \frac{x_n - \mu_k}{\sigma_k} \right|^{\nu_k} \right\} \right\}. \tag{3}$$

The **E-step** involves computing the conditional expected value by using the following equation

$$Q(\theta, \theta^{(m-1)}) = \sum_{k=1}^{K} \left[ \sum_{n=1}^{N} z_{nk}^{(m-1)} \log \left\{ \pi_k \frac{\nu_k}{2\sigma_k \Gamma(1/\nu_k)} \exp\left\{ -\left| \frac{x_n - \mu_k}{\sigma_k} \right|^{\nu_k} \right\} \right\} \right], \tag{4}$$

where

$$z_{nk}^{(m-1)} = \frac{\pi_k p(x_n | \mu_k, \sigma_k, \nu_k)}{\sum_{k=1}^{K} \pi_k p(x_n | \mu_k, \sigma_k, \nu_k)}.$$

The **CM-Step** maximizes $Q(\theta, \theta^{(m-1)})$ with respect to $\theta$ to obtain the $m$-th parameter estimates and increases the expectation of the complete likelihood of the data. The derivatives of the log-likelihood function are set to zero with respect to $\pi_k$ and each constrained parameter, i.e. $\mu$, $\sigma$, and $\nu$:

$$\frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \pi_k} = 0, \quad \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \mu} = 0, \quad \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \sigma} = 0, \quad \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \nu} = 0. \tag{5}$$

It is possible to demonstrate that a non-linear equation is obtained from Eq. (5) for each constrained parameter. In order to compute the constrained parameters values at the ECM iteration $m$-th from the

[2]The iteration equations of the unconstrained parameters are provided by (18).

DRAFT

non-linear equations, the iterative Newton-Raphson method is applied (9; 5; 18) as follows

$$\mu^{(m)} = \mu^{(m-1)} - \frac{f\left(\mu^{(m-1)}\right)}{f'\left(\mu^{(m-1)}\right)}, \quad \sigma^{(m)} = \sigma^{(m-1)} - \frac{h\left(\sigma^{(m-1)}\right)}{h'\left(\sigma^{(m-1)}\right)}, \quad \nu^{(m)} = \nu^{(m-1)} - \frac{g\left(\nu^{(m-1)}\right)}{g'\left(\nu^{(m-1)}\right)}, \tag{6}$$

where

$$f\left(\mu^{(m-1)}\right) = \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \mu}, \quad h\left(\sigma^{(m-1)}\right) = \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \sigma}, \quad g\left(\nu^{(m-1)}\right) = \frac{\partial Q(\theta, \theta^{(m-1)})}{\partial \nu}. \tag{7}$$

## 3. Simulation

Using the **R** software, the simulation is performed for the CMGND with common scale and shape parameter, i.e. with the same variance and kurtosis. The common shape parameter is set to 1.5 in order to test the fitting of the "intermediate distribution" (See Section 2). Samples are generated with the **R**'s function *rgnorm*. Besides, the sampling procedure is repeated $R = 50$ times and sample sizes $N = 500, 2000, 5000$. To assess the estimation performance Bias, MSE and Std are computed as follows:

$$\begin{aligned} Bias\left(\hat{\theta}\right) &= \left| \frac{1}{R} \sum_{s=1}^{R} \hat{\theta}_r - \theta \right|, \\ MSE\left(\hat{\theta}\right) &= \frac{1}{R} \sum_{s=1}^{R} (\hat{\theta}_r - \theta)^2, \\ Std\left(\hat{\theta}\right) &= \sqrt{\frac{1}{R} \sum_{s=1}^{R} \left( \hat{\theta}_r - \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r \right)^2}, \end{aligned} \tag{8}$$

where $\theta$ is the true parameter value and $\hat{\theta}_r$ is the estimate of $\theta$ for the $r$-th simulated data. To avoid the label switching issue, mixtures components are sorted according to the location parameter ($\mu_k$) since $K = 2$ and $\mu_1 \neq \mu_2$.

Tables 1 and 2 show the simulation results. The bias, MSE, and Std of the CMGND are lower than those of the MGND. It can be seen that the estimation accuracy of the CMGND is high from that of the MGND. For $N = 5000$, the bias and MSE are quite similar for both mixtures models. To conclude, Table 3 shows the CPU time in seconds for sample sizes of 500, 2000, and 5000 of the MGND and CMGND. It is found that as the sample size increases the CMGND consumes les CPU time than the MGND.

DRAFT

## 4. Conclusions

In this work, a new constrained univariate mixture model has been introduced. Specifically, this study adds to the literature the CMGND where the parameters are constrained to be equal across mixture components. The ECM algorithm is used to estimate constrained parameters via the MLE. Besides, the iterative Newton-Raphson method is applied to handle the non-linear iteration equations of the parameters during the estimation stage. In brief, simulation results show that the estimation accuracy of the constrained mixture is higher than the unconstrained mixture. The proposed model can be improved in two directions: introducing the multivariate version, and applying a global optimization of the parameters since the solutions strongly depend on the initial starting point.

Table 1: Simulation results for the MGND.

| | $\pi_1$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ | $\nu_1$ | $\nu_2$ | |
|---|---|---|---|---|---|---|---|---|
| $\theta$ | 0.7 | 0 | 5 | 2 | 2 | 1.5 | 1.5 | $N$ |
| Est. | 0.6219 | 0.1026 | 4.3516 | 1.9295 | 2.8049 | 1.8487 | 2.4341 | 500 |
| Bias | 0.0781 | 0.1026 | 0.6484 | 0.0705 | 0.8049 | 0.3487 | 0.9341 | |
| MSE | 0.0301 | 0.1387 | 1.9467 | 0.2672 | 2.7476 | 1.7744 | 3.6243 | |
| Std | 0.1565 | 0.3617 | 1.2480 | 0.5173 | 1.4637 | 1.2986 | 1.6757 | |
| Est. | 0.6998 | -0.0065 | 4.9755 | 2.0022 | 2.0458 | 1.5043 | 1.5572 | 2000 |
| Bias | 2e-04 | 0.0065 | 0.0245 | 0.0022 | 0.0458 | 0.0043 | 0.0572 | |
| MSE | 3e-04 | 0.0041 | 0.0170 | 0.0085 | 0.0329 | 0.0081 | 0.0327 | |
| Std | 0.0169 | 0.0643 | 0.1293 | 0.0930 | 0.1774 | 0.0905 | 0.1734 | |
| Est. | 0.7005 | -0.0041 | 5.0002 | 2.0051 | 2.0070 | 1.5052 | 1.5151 | 5000 |
| Bias | 5e-04 | 0.0041 | 0.0002 | 0.0051 | 0.0070 | 0.0052 | 0.0151 | |
| MSE | 1e-04 | 0.0020 | 0.0040 | 0.0058 | 0.0129 | 0.0040 | 0.0132 | |
| Std | 0.0115 | 0.0448 | 0.0640 | 0.0768 | 0.1144 | 0.0635 | 0.1152 | |

Table 2: Simulation results for the CMGND with the same scale and shape parameter.

| | $\pi_1$ | $\mu_1$ | $\mu_2$ | $\sigma$ | $\nu$ | |
|---|---|---|---|---|---|---|
| $\theta$ | 0.7 | 0 | 5 | 2 | 1.5 | $N$ |
| Est. | 0.6989 | 0.0299 | 5.0201 | 1.973 | 1.4898 | 500 |
| Bias | 0.0011 | 0.0299 | 0.0201 | 0.0270 | 0.0102 | |
| MSE | 8e-04 | 0.0106 | 0.0258 | 0.0344 | 0.0305 | |
| Std | 0.0291 | 0.0996 | 0.1611 | 0.1852 | 0.1762 | |
| Est. | 0.7007 | -0.0034 | 4.9836 | 2.0153 | 1.5136 | 2000 |
| Bias | 7e-04 | 0.0034 | 0.0164 | 0.0153 | 0.0136 | |
| MSE | 2e-04 | 0.0022 | 0.0102 | 0.0058 | 0.0071 | |
| Std | 0.0137 | 0.0471 | 0.1008 | 0.0757 | 0.0843 | |
| Est. | 0.7004 | -0.0045 | 4.9987 | 2.0051 | 1.5048 | 5000 |
| Bias | 4e-04 | 0.0045 | 0.0013 | 0.0051 | 0.0048 | |
| MSE | 1e-04 | 0.0014 | 0.0027 | 0.0031 | 0.0030 | |
| Std | 0.0088 | 0.0370 | 0.0528 | 0.0563 | 0.0553 | |

Table 3: CPU time in seconds for sample sizes of 500, 2000, and 5000.

| | 500 | 2000 | 5000 |
|---|---|---|---|
| MGND | 4.5646 | 14.2914 | 39.8010 |
| CMGND | 4.1522 | 11.5124 | 29.9576 |

DRAFT

## References


[1] Allili, M.S.: Wavelet Modeling Using Finite Mixtures of Generalized Gaussian Distributions: Application to Texture Discrimination and Retrieval. IEEE Trans. Image. Process. **21**, 1452–1464 (2012)

[2] Andrews, J.L., McNicholas, P.D., Subedi, S.: Model-based classification via mixtures of multivariate t-distributions. Comput. Stat. Data Anal. **55**, 520–529 (2011)

[3] Andrews, J.L., Wickins, J.R., Boers, N.M., McNicholas, P.D.: teigen: An R Package for Model-Based Clustering and Classification via the Multivariate t Distribution. J. Stat. Softw. **83**, 1–32 (2018)

[4] Banfield, J.D., Raftery, A.E.: Model-Based Gaussian and Non-Gaussian Clustering. Biom. **49**, 803–821 (1993)

[5] Bazi, Y., Bruzzone, L., Melgani F.: Image thresholding based on the EM algorithm and the generalized Gaussian distribution. Pattern Recognit. **40**, 619–634 (2006).

[6] Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**, 781–793 (1995).

[7] Chauveau, D., Hunter, D.R.: ECM and MM algorithms for normal mixtures with constrained parameters. HAL science ouverte (2013) hal-00625285v2

[8] Hathaway, R.J.: A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. Ann. Stat. **13**, 795–800 (1985).

[9] Kim, D.K., Taylor, J.M.G.: The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Restrictions on the Parameters. J. Am. Stat. Assoc. **90**, 708–716 (1995)

[10] Lee, S.X., McLachlan, G.J.: Model-based clustering and classification with non-normal mixture distributions. Stat. Methods Appt. **22**, 427–454 (2013)

[11] McLachlan, G.J., Peel, D.: Finite mixture models. Wiley Series in Probability and Statistics, New York (2000)

[12] Meng, X.L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**, 267–278 (1993)

[13] Mohamed, O.M.M., Jaïdane-Saïdane, M.: Generalized Gaussian mixture model. 2009 17th Eur. Signal. Process. Conf., pages 2273–2277, (2009).

[14] Nadarajah, S.: A generalized normal distribution. J. Appl. Stat. **32**, 685–694 (2005)

[15] Nettleton, D.: Convergence properties of the EM algorithm in constrained parameter spaces. Can. J. Stat. **27**, 639–648 (1999)

[16] Nguyen, T.M., Jonathan, Wu Q., Zhang H.: Bounded generalized Gaussian mixture model. Pattern Recognit. **47**, 3132–3142 (2014)

[17] Quandt, R.E., Ramsey, J.B.: Estimating Mixtures of Normal Distributions and Switching Regressions. J. Am. Stat. Assoc. **73**, 730–738 (1978)

[18] Wen, L., Qiu, Y., Wang, M., Yin, J., Chen, P.: Numerical characteristics and parameter estimation of finite mixed generalized normal distribution. Commun. Stat. Simul. Comput. **51**, 3596–3620 (2022)


DRAFT

# Mixture-based clustering with covariates for ordinal responses

Kemmawadee Preedalikit[a], Daniel Fernández[b], Ivy Liu[c], Louise McMillan[c], Marta Nai Ruscone[d], and Roy Costilla[e]

[a]University of Phayao; kemmawadee@gmail.com
[b]Universitat Politècnica de Catalunya - BarcelonaTech; daniel.fernandez.martinez@upc.edu
[c]Victoria University Wellington; ivy.liu@vuw.ac.nz, mcmilllo@ecs.vuw.ac.nz
[d]University of Genoa; marta.nairuscone@unige.it
[e]AgResearch NZ; roy.costilla@agresearch.co.nz

DRAFT

**Abstract**

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows and columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. In this study, we have extended mixture-based models to include covariates and test what effect this has on the resulting clustering structures. We focus on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. We fit the models using the Expectation-Maximization (EM) algorithm and assess their performance. Finally, we also illustrate an application of the models to the well-known arthritis clinical trial data set.

***Keywords:*** cluster analysis, mixture models , EM algorithm, ordinal responses, proportional odds model

## 1. Introduction

A well-known definition of an ordinal variable says it is one characterized by a categorical data scale, which describes an order showing differing degrees of dissimilarity (1). Thus, although ordinal variables are affected by the distances among their ordinal categories, those distances are not known. Cluster analysis is the study of techniques to classify a set of related objects into the same cluster (6) and can be applied to identify groups, patterns, or clusters in a data set. Many different approaches to clustering have been developed. The earliest approaches use partition optimization; the most common method is the k-means clustering (11). Several authors have proposed extensions to this approach (see e.g., (24; 9; 21)). Moreover, the objects can be clustered in a hierarchical way, gradually agglomerating objects into larger and larger clusters (25). All approaches listed above are based on mathematical distance metrics and therefore statistical inferences, model selection procedures, and goodness-of-fit assessments cannot be easily applied due to the lack of an underlying probability model (6; 7). Cluster analysis based on finite mixture models (18) assumes that responses in the data matrix arise from mixtures of statistical distributions, with each cluster corresponding to one component of the mixture. The fitted parameters for those distributions are those that have the maximum likelihood based on the observed data. Likelihood-based

methods include those proposed by (18; 5; 15), among others. More recently, (10; 19) proposed an approach via finite mixtures for binary and count data using Bernoulli or Poisson building blocks. Other authors have introduced clustering algorithms specifically for ordinal data: see e.g. (4; 20; 14; 7; 8). (14) proposed a mixture-based two-dimensional clustering solution relying on the proportional odds assumption of the cumulative logit model. (7) developed an equivalent model-based clustering approach using the ordered stereotype model (3). Unlike distance-based methods, which only determine which objects should be clustered together, likelihood-based methods can additionally describe the properties of each cluster, based on the fitted parameters, and can also estimate the probability of each object being allocated to each cluster. Additionally, the mixture-based approaches for ordinal responses introduced above are focused on finding cluster structures based only on the matrix of ordinal responses, and assume that no associated covariates are available. Any available covariates can be analyzed alongside the clustering results, to assist with interpretation of the cluster structures, even though there has been no reference to the covariates during the clustering process, but actually incorporating covariates in the clustering process could lead to different fitted clustering structures, and a different estimate of the number of clusters. Generally speaking, if a model with covariates is fitted, subjects tend to be clustered according to their responses and covariate effects. Therefore, it is desirable to make available covariates endogenous to the clustering process to improve interpretation of the main characteristics of the clusters (17). Our approach to mixture-based clustering involves constructing an additive linear model of parameters, connected to the response data via a link function. Additional terms such as covariates may easily be added to the linear predictor. To the best of our knowledge, (8) introduced this formulation of model-based clustering for ordinal data with covariates, but the performance of these covariate methods and, more importantly, their influence on the resulting clustering structures, have not been documented so far. The main purpose of this article is to extend such models to include covariates and allow them to affect the detection of cluster structures. Moreover, we are also interested in comparing how the resulting clustering structures compare to those obtained without covariates, and how these changes may affect the interpretation of the results. We will focus on extending the one-dimensional clustering approach proposed in (14). This approach models ordinal response data using the proportional odds assumption of the cumulative logit model (from now on "proportional odds model"). We will include covariates directly in the linear predictor. Our approach to clustering follows the constructivist approach described by (12), but with an interest in realist clustering: we think there are many scenarios where patterns in the data can be simplified by identifying clusters of observations that follow similar patterns, but if there is a real structure in the data, then we wish to determine that structure. There are many real-world scenarios that we can model as a response variable being affected by predictor variables, and in some of those scenarios, certain groups of observations may have different patterns of response to the predictors than other groups of observations. If those groups have already been identified, then we might attempt mixed model analysis or multilevel modelling; but if the groups have not already been identified, then the method we propose here provides a pathway to detecting these groupings of response patterns. So our approach could be seen as a bridge between regression modelling and cluster analysis.

DRAFT

## 2. Model formulation

When the data are in matrix form, clustering of rows is called row clustering. We present the row clustering formulation for finite mixtures based on the proportional odds model. This closely follows the model formulations in (14; 8). Clustering can also be applied to the columns of the data matrix, but the formulation of those models is equivalent to the formulation of row clustering models applied to a transposed version of the data matrix, so we only present the row clustering formulation here. We decided to focus on row clustering because it is more common to have covariates linked to observations (rows) than to variables (columns). However, the variable version is very similar; in order to include column covariates instead of row covariates, we can simply change the indices in the formulae. We consider a set of $n$ subjects and $m$ ordinal response variables, each with $q$ possible ordinal response categories. Thus, data can be represented by an $n \times m$ matrix $\mathbf{Y}$ with ordinal entries $y_{ij}$. The row cluster index $r$ ($r = 1, \ldots, R$) represents the number of the row cluster and the symbol $i \in r$ indicates that row $i$ is

allocated to row cluster $r$. We shall assume that all rows belonging to the same row cluster $r$ have ordinal responses driven by the same row cluster effect, i.e. that there are no individual row effects. In the case of the proportional odds model where the effect of rows on the response is considered, the probability that $y_{ij}$ takes category $k$, when row $i$ is in row cluster $r$, is defined by

$$P[y_{ij} = k | i \in r] = \theta_{ijrk},$$

where $i = 1, \ldots, n,\ j = 1, \ldots, m$ and $k = 1, \ldots, q$ with $\sum_{k=1}^{q} \theta_{ijrk} = 1$ for a given $i, j$ and $r$. This can be expressed using linear predictor terms as

$$\text{logit}\left(\sum_{h=1}^{k} \theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + \beta_j + \gamma_{rj}), \tag{1}$$

where the parameters $\{\mu_k\}$ are the cutpoints. $\{\alpha_r\}$ and $\{\beta_j\}$ indicate the effects of row cluster $r$ and column $j$, respectively, and $\{\gamma_{rj}\}$ represent the associations between different row clusters and individual columns. Corner-point or sum-to-zero constraints on $\{\alpha_r\}$, $\{\beta_j\}$ and $\{\gamma_{rj}\}$ must be included to avoid identifiability problems and the monotonically increasing constraint $\mu_1 < \mu_2 < \cdots < \mu_q (= \infty)$ is included to capture the ordinal nature of the responses. The (unknown) proportion of rows in each row group $r$ is defined as $\{\pi_1, \ldots, \pi_R\}$, with $\sum_{r=1}^{R} \pi_r = 1$. In a simpler model with clustering of rows, the rows (observations/subjects) will tend to be clustered if they have similar patterns of responses, without taking into account the information present in the covariates. However, if we include in the clustering process the information from covariates such as type of cancer, treatment dose, initial tumour burden, size of the tumour, gender, and age then the resulting clusters may be different because patients with equal or similar values in the covariates should be *a priori* more likely to co-cluster than others. For instance, patients with larger tumour sizes will tend to be clustered together regardless of their responses to the questionnaire. This motivational example is based on the one given in (16).

We now define the model formulation of row clustering using the proportional odds model, with additional covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, as follows,

$$\text{logit}\left(\sum_{h=1}^{k} \theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + \beta_j + \gamma_{rj} + x_i^T \delta_r), \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are a set of $p$ covariates associated with row $i$ of the data matrix; these covariates can be categorical or continuous. The parameters $\{\delta_r\}$ represent the effects of the covariates; we assume these effects are the same for all rows in the same row cluster $r$. When fitting this model, the subjects will be clustered according to both their response patterns and the values of their covariates, which may lead to different estimates of cluster assignment. Having in mind that $R$ and $C$ are the numbers of row clusters and column clusters, respectively, we will deal with the possible values of $C = m$ (when column effects are different and therefore they are included within the model, without clustering). $C = 1$ when the column effect is the same and it is not included into the model. Considering the simplest row clustering model, without column effects, the proportional odds model without covariates can be expressed as

$$\text{logit}\left(\sum_{h=1}^{k} \theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - \alpha_r, \tag{3}$$

where the number of parameters, including the $R - 1$ independent values of $\pi_r$, is $q + 2R - 3$. Adding $p$ covariates into model (3), we obtain

$$\text{logit}\left(\sum_{h=1}^{k} \theta_{ijrh}\right) = \eta_{ijrk} = \mu_k - (\alpha_r + x_i^T \delta_r), \tag{4}$$

where there are now $q + (p + 2)R - 3$ parameters in the model. Models (3) and (4) will be used in the simulation and application section to compare the clustering structure.

DRAFT

## 3. Application

We applied the models proposed in this article to the *arthritis clinical trial* data set (13), which compares the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis. The data set is obtained from the **R** package *multgee* (23). In this application, the covariate-dependent clustering could help to identify subsets of patients with similar covariate information patterns. This insight would be important because it would provide a flexible approach for identifying potential heterogeneous gender, age, and auranofin treatment effects on the arthritis scores. For instance, if the elderly experience more symptoms and, consequently, tend to be more pessimistic about their arthritis status, our proposed model would allow us to distinguish subsets of older people that tend to report higher/lower arthritis scores. However, we note that this is only an example and we do not advocate the clinical relevance of the covariate-dependent clustering model. In real settings, clinicians and the statisticians together should decide which model, i.e. no clustering, clustering with covariates, or clustering without covariates, is more relevant to answer their research questions. After fitting the models without covariates (3) and with covariates (4), with different number of row clusters, we compared them using the information criteria AIC and BIC (see results in Table 1). AIC indicates that the best model is the version of the row clustering model including age and treatment covariates ($\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$) with $R = 4$ row clusters (AIC = 2136.78), which is better than its counterpart in the model without covariates (AIC=2154.40). However, BIC shows that the model without covariates ($\mu_k - \alpha_r$) and $R = 4$ is the best model (BIC=2202.05). A possible reason is that BIC penalizes higher numbers of parameters more strongly than AIC does, leading to a preference for more parsimonious models. This result might make sense as the model with covariates includes more information in the fitting process. On the other hand, AIC is a standard way in model selection to choose the best model but, as we mentioned above, ideally, clinicians and the statisticians should decide together which is the more relevant model. Additionally, Table 3 shows the results of the comparison of clustering structure agreement of the selected models with and without covariates by using the information theoretic criterion ARI, 1-NVI and 1-NID. The results assume each patient has been allocated to the cluster for which they have the highest posterior probability of membership. Comparing the clustering structure agreement (ARI, 1-NVI, and 1-NDI) of the best model (Model (4) with $R = 4$ including age and treatment covariates) and its counterpart without covariates (Model (3)), the values of the three measures are 0.66 (ARI), 0.47 (1-NVI), and 0.64 (1-NDI), which shows that models (3) and (4) result in different clustering structures.

DRAFT

## References


[1] Agresti, A.: Analysis of Ordinal Categorical Data, Second Edition. Wiley and Sons (2010)

[2] Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd International Symposium on Information Theory, pp. 267-281. Budapest: Akademiai Kiado (1973)

[3] Anderson, J.A.: Regression and ordered categorical variable. Journal of Royal Statistic Society 46, 1-30 (1984)

[4] Biernacki, C., Jacques, J.: Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing 26, 929-943 (2016)

[5] Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V., Walther, G.: Advances in mixture models. Computational Statistics and Data Analysis 51(11), 5205-5210 (2007)

[6] Everitt, B., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. JohnWiley and Sons, New York (2011)

[7] Fernández, D., Arnold, R., Pledger, S.: Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis 93, 46-75 (2016)

[8] Fernández, D., Arnold, R., Pledger, S., Liu, I., Costilla, R.: Finite mixture biclustering of discrete type multivariate data. Advances in Data Analysis and Classification 13, 117-143 (2019)

[9] Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97, 611-631 (2002)

Table 1: Results of row clustering models fitted to the arthritis data set. The best model in each group of models (no covariates, one, two, or three covariates), based on AIC, is shown in bold.

| Model | | $R$ | number of parameter | Log-like | AIC | BIC |
|---|---|---|---|---|---|---|
| $\mu_k - \alpha_r$ | | 2 | 6 | -1096.99 | 2205.99 | 2234.58 |
| | | 3 | 8 | -1077.73 | 2171.46 | 2209.59 |
| | | **4** | 10 | -1067.20 | **2154.40** | **2202.05** |
| | | 5 | 12 | -1067.20 | 2158.40 | 2215.58 |
| $\mu_k - (\alpha_r + x_i\delta_r)$ | $x$= age | 2 | 8 | -1138.18 | 2292.37 | 2330.49 |
| | | 3 | 11 | -1071.88 | 2165.75 | 2218.17 |
| | | 4 | 14 | -1065.18 | 2158.37 | 2225.08 |
| | | 5 | 17 | -1060.84 | **2155.68** | 2236.68 |
| | $x$=treatment | 2 | 8 | -1082.28 | 2180.57 | 2218.69 |
| | | 3 | 11 | -1067.93 | 2157.87 | 2210.28 |
| | | **4** | 14 | -1057.70 | **2143.40** | 2210.11 |
| | | 5 | 17 | -1056.23 | 2146.46 | 2227.46 |
| | $x$= gender | 2 | 8 | -1096.89 | 2209.77 | 2247.89 |
| | | 3 | 11 | -1079.51 | 2181.02 | 2233.44 |
| | | 4 | 14 | -1066.92 | **2161.84** | 2228.55 |
| | | 5 | 17 | -1066.37 | 2166.74 | 2247.74 |
| $\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$ | $x_1$ = age, | 2 | 10 | -1072.54 | 2165.07 | 2212.72 |
| | $x_2$= treatment | 3 | 14 | -1059.23 | 2146.46 | 2213.17 |
| | | **4** | 18 | -1050.39 | **2136.78** | 2222.55 |
| | | 5 | 22 | -1048.53 | 2141.05 | 2245.88 |
| | $x_1$ = age, | 2 | 10 | -1085.83 | 2191.67 | 2239.32 |
| | $x_2$= gender | 3 | 14 | -1068.97 | 2165.95 | 2232.66 |
| | | 4 | 18 | -1061.29 | **2158.58** | 2244.35 |
| | | 5 | 22 | -1059.26 | 2162.52 | 2267.35 |
| | $x_1$ = treatment, | 2 | 10 | -1081.82 | 2183.64 | 2231.29 |
| | $x_2$= gender | 3 | 14 | -1065.99 | 2159.99 | 2226.71 |
| | | 4 | 18 | -1056.73 | **2149.45** | 2235.22 |
| | | 5 | 22 | -1055.06 | 2154.13 | 2258.96 |
| $\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r} + x_{i3}\delta_{3r})$ | $x_1$ = age, | 2 | 12 | -1071.60 | 2167.21 | 2224.39 |
| | $x_2$= treatment, | 3 | 17 | -1060.50 | 2155.00 | 2236.01 |
| | $x_3$= gender | 4 | 22 | -1050.35 | **2144.71** | 2249.54 |
| | | 5 | 27 | -1052.14 | 2158.35 | 2287.00 |

DRAFT


[10] Govaert, G., Nadif, M.: Latent block model for contingency table. Communications in Statistics - Theory and Methods 39(3), 416-425 (2010)

[11] Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. Applied Statistics 28, 100-108 (1979)

[12] Hennig, C.: What are the true clusters? Pattern Recognition Letters 64, 53-62 (2015)

[13] Lipsitz, S.R., Fitzmaurice, G.M., Molenberghs, G.: Goodness-of-fit tests for ordinal response regression models. Journal of the Royal Statistical Society. Series C (Applied Statistics) 45(2), 175-190 (1996)

[14] Matechou, E., Liu, I., Fernández, D., Farias, M., Gjelsvik, B.: Biclustering models for two-mode ordinal data. Psychometrika 81(3), 611-624 (2016)

[15] Melnykov, V., Maitra, R.: Finite mixture models and model-based clustering. Statistics Surveys 4, 80-116 (2010)

[16] Müller, P., Quintana, F., Rosner, G.L.: A product partition model with regression on covariates. Journal of Computational and Graphical Statistics 20:1, 260-278 (2011)

[17] Murphy, K., Murphy, T.B.: Gaussian parsimonious clustering models with covariates and a noise

Table 2: Estimated parameters of two models, the first with no covariates and the second with the covariates age and treatment.

| | model without covariates (3) | model with covariates (4) | | |
|---|---|---|---|---|
| $R$ | $\alpha_r$ | $\alpha_r$ | $\delta_{1r}$ (age) | $\delta_{2r}$ (treatment) |
| 1 | 4.20 (0.17) | 3.64 (0.12) | 0.56 (0.15) | 0.23 (0.14) |
| 2 | 1.26 (0.25) | 1.20 (0.15) | 0.02 (0.25) | -0.84 (0.22) |
| 3 | -1.41(0.14) | -1.50 (0.24) | -0.22 (0.12) | -0.82 (0.25) |
| 4 | -4.04 (0.13) | -3.34 (0.18) | 0.58 (0.13) | -2.00 (0.17) |

Table 3: Arthritis data set: Comparison of clustering structure agreement between models without covariates (3) (left) vs. with covariates (4) (right).

| $R$ | Clustering comparison | ARI | 1-NVI | 1-NID |
|---|---|---|---|---|
| 2 | no covariate VS age | 0.86 | 0.69 | 0.81 |
| | no covariate VS treatment | 0.74 | 0.51 | 0.68 |
| | no covariate VS gender | 0.87 | 0.70 | 0.82 |
| | no covariate VS age, treatment | 0.69 | 0.45 | 0.61 |
| | no covariate VS age, gender | 0.87 | 0.70 | 0.82 |
| | no covariate VS treatment, gender | 0.75 | 0.55 | 0.70 |
| | no covariate VS age, treatment, gender | 0.70 | 0.46 | 0.62 |
| 3 | no covariate VS age | 0.43 | 0.37 | 0.52 |
| | no covariate VS treatment | 0.14 | 0.12 | 0.21 |
| | no covariate VS gender | 0.33 | 0.32 | 0.47 |
| | no covariate VS age, treatment | 0.58 | 0.25 | 0.38 |
| | no covariate VS age, gender | 0.70 | 0.61 | 0.74 |
| | no covariate VS treatment ,gender | 0.27 | 0.23 | 0.36 |
| | no covariate VS age, treatment, gender | 0.14 | 0.17 | 0.27 |
| 4 | no covariate VS age | 0.44 | 0.34 | 0.48 |
| | no covariate VS treatment | 0.85 | 0.68 | 0.80 |
| | no covariate VS gender | 1.00 | 1.00 | 1.00 |
| | no covariate VS age, treatment | **0.66** | **0.47** | **0.64** |
| | no covariate VS age, gender | 0.76 | 0.57 | 0.71 |
| | no covariate VS treatment, gender | 0.84 | 0.67 | 0.80 |
| | no covariate VS age, treatment, gender | 0.58 | 0.39 | 0.55 |

DRAFT

component. Advances in Data Analysis and Classification 14, 293-325 (2020)

[18] Peel, D., McLachlan, G.: Finite Mixture Models. John Wiley and Sons, Inc., Wiley Series in Probability and Statistics (2000)

[19] Pledger, S., Arnold, R.: Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics and Data Analysis 71, 241-261 (2014)

[20] Ranalli, M., Rocci, R.: Mixture models for ordinal data: a pairwise likelihood approach. Statistics and Computing 26, 529-547 (2016)

[21] Rocci, R., Vichi, M.: Two-mode multi-partitioning. Computational Statistics and Data Analysis 52, 1984-2003 (2008)

[22] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6(2), 461-464 (1978)

[23] Touloumis, A.: R package multgee: A generalized estimating equations solver for multinomial responses. Journal of Statistical Software 64(8), 1-14 (2015).

[24] Vichi, M.: Double k-means clustering for simultaneous classification of objects and variables. In: S. Borra, R. Rocci, M. Vichi, M. Schader (eds.) Advances in Classification and Data Analysis, pp. 43-52. Springer Berlin Heidelberg (2001)

[25] Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236-244 (1963)

# Partial membership models for soft clustering of multivariate count data

Emiliano Seri[a], Thomas Brendan Murphy[b], and Roberto Rocci[a]

[a]Department of Statistics, Sapienza University of Rome, Piazzale Aldo Moro 5, 000185, Rome, Italy;
emiliano.seri@uniroma1.it, roberto.rocci@uniroma1.it
[b]School of Mathematics and Statistics, University College Dublin, Dublin, D04 V1W8, Ireland;
brendan.murphy@ucd.ie

**Abstract**

The standard mixture modelling framework has been widely used to study heterogeneous populations, by modelling them as being composed of a finite number of homogeneous sub-populations. However, the standard mixture model assumes that each data point belongs to one and only one mixture component, or cluster, but when data points have fractional membership in multiple clusters this assumption is unrealistic. It is in fact conceptually very different to represent an observation as partly belonging to multiple groups instead of belonging to one group with uncertainty. For this purpose, various soft clustering approaches, or individual-level mixture models, have been developed. In this context, (6) formulated the Bayesian partial membership model (BPM) as an alternative structure for individual-level mixtures, which also captures partial membership in the form of attribute-specific mixtures, but does not assume a factorization over attributes. Our work proposes using the BPM for soft clustering of count data. Learning and inference are carried out using Markov chain Monte Carlo methods. The method is applied on Capital Bike share data of Washington DC from 15 of June to 15 of July 2022.

***Keywords:*** Partial membership models, Model based clustering, Mixture models

DRAFT

## 1. Introduction

Model-based clustering has been widely used among researchers to study heterogeneous populations, by modelling them as being composed of a finite number of homogeneous sub-populations (4; 7). Within this framework the observations in a dataset are modelled as they are drawn from one of several probability distributions. A clustering solution is sought whereby observations are partitioned into distinct groups, so that observations which have non-negligible posterior probability of belonging to more than one component are seen as having uncertain group membership, and are perhaps indicative of a poorly fit model. However, the standard mixture model assumes that each data point belongs to one and only one mixture component, or cluster, but, as explained in (11), when data points have fractional membership in multiple clusters this assumption is unrealistic; the idea of Mixed and Partial membership models accommodate partial membership. Following (6) example, let's consider an individual with a mixed ethnic background, say, partly Asian and partly European. It seems sensible to represent that individual as partly belonging to two different classes or sets. Being certain that a person is partly Asian and partly European, is very different than being uncertain about a person's ethnic background. The original idea for a mixed membership type of modeling goes back to at least the 1970s when the Grade of Membership (GoM) model was developed by mathematician Max Woodbury to allow for "fuzzy" classifications in medical

diagnosis problems (12). It was not until the early 2000s, with the widespread use of Bayesian methods and a better explanation of the duality between the discrete and continuous nature of latent structure in the GoM model, that a new Bayesian approach to the GoM model had been developed. Independently, within a short time of each other, three mixed membership models were developed to solve problems in three very different areas: (1) with Latent Dirichlet Allocation (LDA), (3) with Grade of Membership model (GoM) and (8) with Admixture model. Mixed membership models unifies the LDA, GoM, and Admixture models in a common framework and provides ways to construct other individual-level mixture models by varying assumptions on the population, sampling unit and latent variable levels, and the sampling scheme. In (6), Partial membership models are defined, which, albeit being part of the same framework, they overcome some of the drawbacks of mixed membership models.

## 2. Partial membership model

Consider a data set $\boldsymbol{X} = \{\boldsymbol{x}_{ij} : i = 1, 2, \ldots, N, j = 1, 2, \ldots, J\}$. In a finite mixture model, the density of a data point $\boldsymbol{x_i}$ given $\boldsymbol{\Theta}$, which contains the parameters for each of the $K$ mixture components is

$$P(\boldsymbol{x_i}|\boldsymbol{\Theta}) = \sum_{\tau_i} P(\boldsymbol{\tau}_i) \prod_{k=1}^{K} P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^{\tau_{ik}}. \tag{1}$$

Where $\boldsymbol{\tau}_i$ are the weights (or partial membership) which represents how much each data point belongs to each component, so $\tau_{ik} \in \{0, 1\}$ and $\sum_k \tau_{ik} = 1$. We relax the constrain $\tau_{ik} \in \{0, 1\}$ to take any continuous value in the range $[0, 1]$. So we change $\tau_{ik}$ from being binary to being in the simplex. The complete data likelihood become

$$P(\boldsymbol{x}_i|\boldsymbol{\Theta}) = \frac{1}{c} \int_{\boldsymbol{\tau}_i} P(\boldsymbol{\tau}_i) \prod_{k=1}^{K} P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^{\tau_{ik}} d\boldsymbol{\tau}_i.$$

We integrate over all values of $\boldsymbol{\tau}_i$ instead of summing, and since the product over clusters $K$ (in Equation 1) no longer normalizes, we put in a normalization constant $c$, which is a function of $\boldsymbol{\tau}_i$ and $\boldsymbol{\Theta}$. In this work we specify the case when the form of the distribution for each cluster $P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)$ are Poisson

$$P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{J} P_k(\boldsymbol{x}_{ij}|\lambda_{kj}) = \prod_{j} \frac{\lambda_{kj}^{\boldsymbol{x}_{ij}} e^{-\lambda_{kj}}}{\boldsymbol{x}_{ij}!}.$$

Consider a model with $K$ clusters and let $\boldsymbol{\delta}$ be a $K$-dimensional vector of positive hyperparameters ($\boldsymbol{\delta} \sim unif(a, b)$). We start by drawing mixture weights from a Dirichlet distribution

$$\boldsymbol{\tau}_i \sim Dir(\boldsymbol{\delta}).$$

For each data point $i$, we draw a partial membership vector $\boldsymbol{\tau}_i$. We assumed that each cluster $k$ is characterized by a Poisson distribution with natural parameters $\lambda_{kj}$ and that

$$\lambda_{kj} \sim conj(\alpha, \beta).$$

A prior and likelihood are said to be *conjugate* when the resulting posterior distribution is the same type of distribution as the prior. Gamma distribution is a conjugate prior for the Poisson because they share the same functional form

$$P(\lambda_{kj}) \propto \lambda_{kj}^{\alpha-1} e^{-\beta\lambda_{kj}}.$$

Where $\alpha$ and $\beta$ are hyperparameters of the prior. Given all these latent variables[1], each data point is drawn from

$$x_{ij} \sim Pois(\exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})).$$

[1]In the Bayesian framework the term latent variables could be used instead of parameters, to state that the model uses random variables that remain unobserved during inference.

DRAFT

Which is the shorthand for

$$P(\boldsymbol{x}_{ij}|\boldsymbol{\tau}_i, \lambda_{kj}) \sim Pois(\prod_{k=1}^{K} \lambda_{kj}^{\tau_{ik}}) = Pois(\exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})).$$

The generative process for $\boldsymbol{X}$ in the partial membership model, compared to those in mixed membership and mixture models is:

| **Mixture model** | **Mixed membership model** | **Partial membership model** |
|---|---|---|
| for($i$ in $1:N$) | for($i$ in $1:N$) | for($i$ in $1:N$) |
| $\boldsymbol{\tau}_i = \boldsymbol{\delta}$ | $\boldsymbol{\tau}_i \sim Dirichlet(\boldsymbol{\delta})$ | $\boldsymbol{\tau}_i \sim Dirichlet(\boldsymbol{\delta})$ |
| $\boldsymbol{Z}_i \sim Multinomial(\boldsymbol{\tau}_i)$ | for($j$ in $1:J$) | for($j$ in $1:J$) |
| for($j$ in $1:J$) | $\boldsymbol{Z}_{ij} \sim Multinomial(\boldsymbol{\tau}_i)$ | $\mu_{ij} = \exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})$ |
| $\boldsymbol{X}_{ij} \sim Poisson(\lambda_{Z_i,j})$ | $\boldsymbol{X}_{ij} \sim Poisson(\lambda_{Z_{ij},j})$ | $\boldsymbol{X}_{ij} \sim Poisson(\mu_{ij})$ |

Partial membership model does not assume a factorization over attributes. More generally, mixed membership (MM) models, assume that each data attribute (for instance in a text analysis example the data attributes could be the words) of the data point (e.g. document) is drawn independently from a mixture distribution given the membership vector for the data point, $x_{nj} \sim \sum_k \tau_{nk} P(\boldsymbol{x}|\lambda_{kj})$. MM models only makes sense when the objects (e.g. documents) being modelled constitute bags of exchangeable sub-objects (e.g. words). Partial membership models make no such assumption.

The complete-data posterior takes the form

$$P(\boldsymbol{\tau}, \boldsymbol{\lambda}|\boldsymbol{x}, \alpha, \beta, a, b) \propto P(\boldsymbol{\lambda}|\alpha, \beta) P(\boldsymbol{\tau}|\boldsymbol{\delta}) \prod_{k=1}^{K} \prod_{j=1}^{J} P_k(\boldsymbol{x}_j|\lambda_{kj})^{\tau_k}.$$

Learning in the BPM consists of inferring all unknown variables given $\boldsymbol{X}$, for which we employ Monte Carlo Markov chain (MCMC). Another advantage of BPM over MM models, is that in the latter there is a discrete latent variable for every sub-object, corresponding to which mixture component that sub-object was drawn from. This large number of discrete latent variables makes MCMC sampling in MM potentially much more expensive than in BPM models.

DRAFT

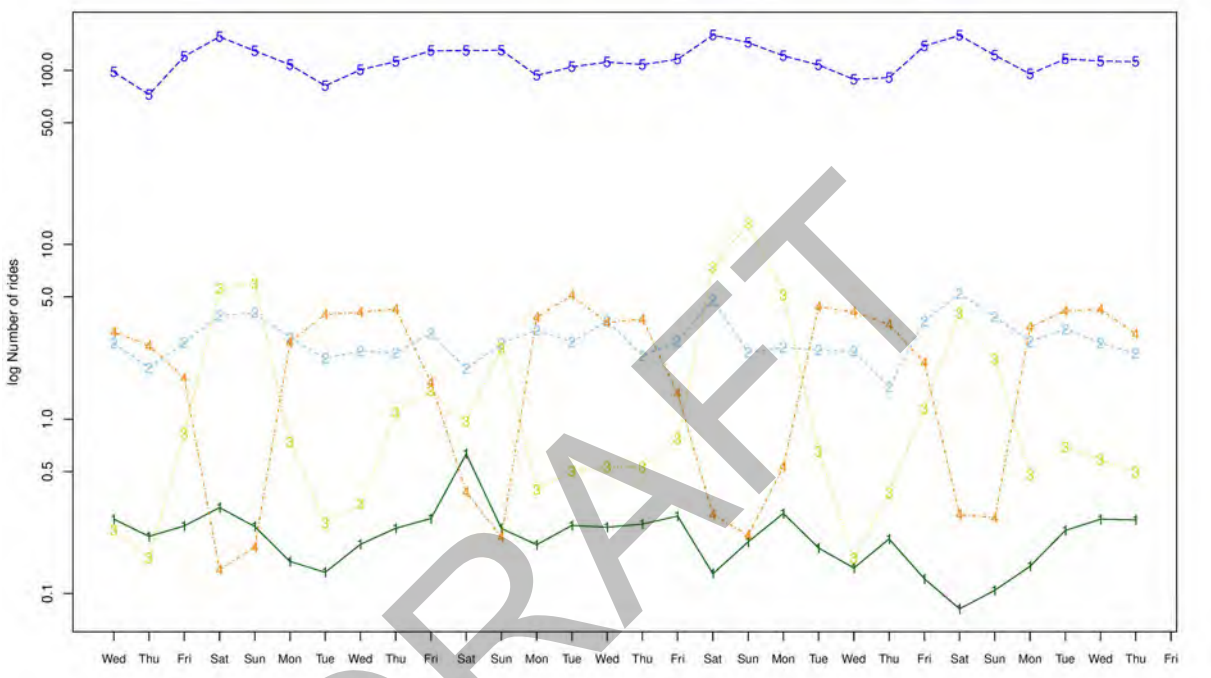
## 2.1 Model selection

According to (10), a statistical model is said to be *regular* if the map taking parameters to probability distributions is one-to-one and if its Fisher information matrix is positive definite. If a model is not regular, then it is said to be *singular*. If a statistical model contains a hierarchical structure, hidden variables, or a grammatical rule, then the model is generally singular. In singular statistical models, the maximum likelihood estimator does not satisfy asymptotic normality. Consequently, AIC is not equal to the average generalization error (5), and the Bayes information criterion (BIC) is not equal to the Bayes marginal likelihood (9), even asymptotically. In singular models, the maximum likelihood estimator often diverges, or even if it does not diverge, makes the generalization error very large. Therefore, the maximum likelihood method is not appropriate for singular models. On the other hand, Bayes estimation was proven to make the generalization error smaller if the statistical model contains singularities. WAIC (10), (Widely Applicable Information Criterion) could be used for estimating the predictive loss of Bayesian models, using a sample from the full-data posterior, and it is applicable to non-regular models, including non-identifiable models and non-realizable models.

# 3. Application: Washington DC bikes data

We applied partial membership model on the data of the bike sharing company of Washington DC. The data are collected daily, from 15 of June to 15 of July 2022, and record each single ride: date and

time of start of trip, date and time of end of trip, name, ID, longitude and latitude of starting station, name, ID, longitude and latitude of ending station. We calculated how many times bikes are collected from each station and we modelled these counts using a partial membership model, with the intent that a more deep exploration of the interactions between the bikes stations usage, could allow to improve the allocation of the bikes. All the analysis has been carried out using NIMBLE, a system for programming statistical algorithms for general model structures within R (2). The model with the lowest WAIC is the one with 5 profiles (or components). For a better visualization, in Figure 3. are represented the natural log of the profiles means, while Figure 3. shows a pie chart on each bike stations, that represent the profiles memberships of each.




DRAFT

Figure 1: Log of the expected number of rides per day from 15 of June to 15 of July 2022, conditional on profile membership, with 5 profiles.

It could be seen that profile 5 groups the busiest stations, which are mainly located in the center of the city. Profile 1 the less used ones, which are mainly in the outlying areas, profile 2 is an average usage stations cluster and looking at the map, it seems to connect the centre to the peripheral areas, profile 3 groups the stations mostly used during the weekends, with an high peak of usage during the holiday of Monday 4 of July, which is bank holiday in the States. The stations with an high membership to this profile, are often located near the river or green ares, or also in the outlying areas. Profile 4 is the group of the stations mostly used on working days.

## 4. Conclusions and future developments

Partial membership models provide the analyst with tools of greater flexibility than current model based clustering or standard distance-based clustering methods. We specified the model for count data and applied to bike sharing data of Washington DC. We think this model can be of great use in many applications, such as social sciences, genetics, natural sciences and textual analysis, and can overcome some of the limitations of mixed membership models. As a future developments, it would be interesting to compare the results on applications of partial membership with mixed membership type of models,

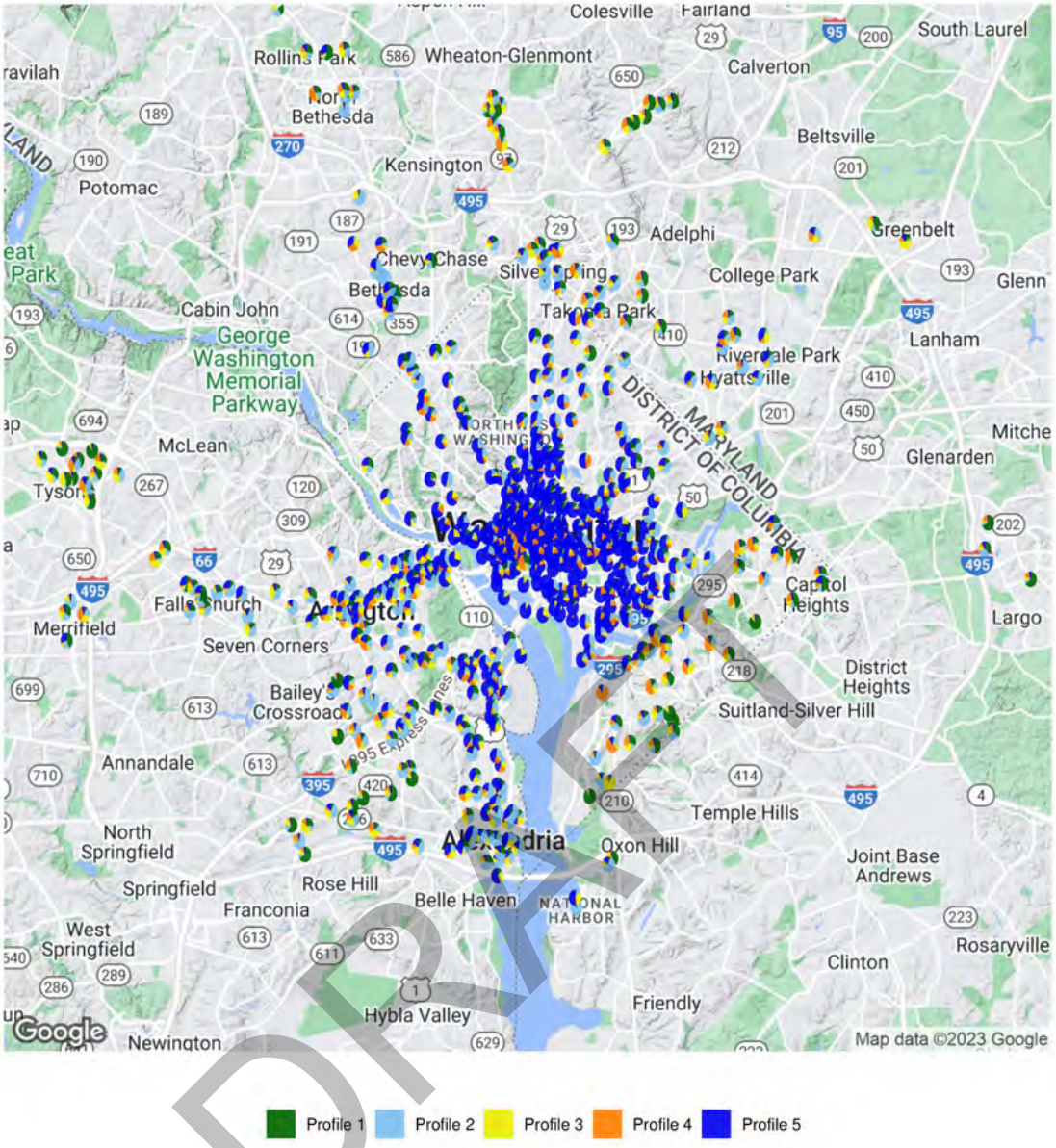



Figure 2: Bike stations' pie charts of profiles membership.

to be able to state which one suits better when dealing with count data. It would be also interesting to explore solutions to assess the over-dispersion issue.

## References


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017. doi: 10.1080/10618600.2016.1172487.

[3] E. A. Erosheva. Bayesian estimation of the grade of membership model. *Bayesian statistics*, 7: 501–510, 2003.

[4] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation.


DRAFT

*Journal of the American statistical Association*, 97(458):611–631, 2002. doi: 10.2307/3085676.

[5] K. Hagiwara. On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, 14(8):1979–2002, 2002. doi: 10.1162/089976602760128090.

[6] K. A. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine learning*, pages 392–399, 2008. doi: 10.1145/1390156.1390206.

[7] D. Peel and G. MacLahlan. *Finite mixture models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 2000. doi: 10.1002/0471721182.

[8] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000. doi: 10.1086/302959.

[9] S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001. doi: 10.1162/089976601300014402.

[10] S. Watanabe and M. Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

[11] A. White and T. B. Murphy. Exponential family mixed membership models for soft clustering of multivariate data. *Advances in Data Analysis and Classification*, 10(4):521–540, 2016. doi: 10.1007/s11634-016-0267-5.

[12] M. A. Woodbury, J. Clive, and A. Garson Jr. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research*, 11(3):277–298, 1978. doi: 10.1016/0010-4809(78)90012-5.

DRAFT

# Regression for mixture models for extremes

Viviana Carcaiso[a], Ilaria Prosdocimi[b], and Isadora Antoniano-Villalobos[b]

[a]Department of Statistical Sciences, University of Padova, Italy;
viviana.carcaiso@phd.unipd.it
[b]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy; ilaria.prosdocimi@unive.it, isadora.antoniano@unive.it

**Abstract**

In many practical problems the assumption that the data are generated by a single process does not hold. Mixture models can deal with this issue and represent a useful resource also in the context of extreme values. A typical scenario with extreme events is that there are two processes underlying the data: one that occurs with a much higher frequency and another which takes place more rarely but can lead to stronger magnitudes. However, simulation studies and applications to real data show that the rare type may not always correspond to the most extreme events and that the tails of the type-specific data-generating processes are not always identified by this information. Therefore, we aim at creating a new regression model that exploits the type of event as a covariate, together with other variables of interest, such as spatial characteristics.

***Keywords:*** extreme value theory, finite mixtures, environmental problems

DRAFT

## 1. Introduction

The typical assumption that the data are representative of a unique population of interest is often too restrictive in real world problems. This constitutes an issue that is interesting to tackle in the extreme value framework (1), where the focus is not as usual on the behaviour of the central part of the distribution of the process of interest, but on the tail behaviour, that corresponds to the rarely observed events. Extreme value theory aims indeed at quantifying the stochastic behaviour of a process at extremely large (or small) values on the basis of asymptotic results and provides approaches that are specifically designed for the analysis of this kind of data. In this context the observations, which consist in maximum values of the measured process in a block, are usually assumed to be i.i.d. draws from an appropriate long-tailed distribution, typically the Generalized Extreme Value (GEV) distribution, which is a limiting distribution supported by theoretical results, but other distributions have been proposed and exploited.

Examples of rare events where the data are actually generated by at least two distinct processes are rainfalls generated by typhoon and non-typhoon weather systems, floods originating from a mixture of rainfall and snowmelt or, in a non-environmental context, stock prices bursts linked to the impact of different economic cycles. In order to take into account the specific nature of the data and exploit appropriate methods to deal with variables whose distribution is determined by multiple components, it is possible to exploit finite mixture models (2; 3). These models are used to describe the data as being drawn from a density which is modelled as a convex combination of components, each with a specified parametric form. Hence, they typically involve more parameters than single population models, which results in a more complex estimation in favor of a gain in flexibility.

The research is driven by the application to hydrological area, where due to natural hazards the estimation of the frequency of the extreme events is crucial. In the literature there are numerous studies

on the application of mixture distributions for analysing hydrological extremes. The first contribution in this direction is the one of Rossi et al. (5), who defined the two-component extreme value (TCEV) distribution, which assumes the individual floods to arise from a mixture of two Exponential distributions and the number of events in a year to be generated from two Poisson distributions. Hence they specified a four-parameter distribution, that can be considered a generalization of the Gumbel distribution, which is a special case of the GEV distribution obtained by setting the shape parameter equal to 0. In particular, the TCEV distribution has CDF which is exactly the product of two Gumbel CDF's. Among others, Kjeldsen et al. (4) proposed instead a Gumbel mixture distribution for modelling extreme events from two different phenomena. Considering the independent random variables $X_1$ and $X_2$ that correspond to the annual maxima of two different processes, the CDF of the annual maximum $X$ is expressed in terms of conditional distributions as

$$F_X(x) = G_1(x)(1-\omega) + G_2(x)\omega,$$

where $\omega$ is the probability that the annual maximum value is a result of the process of type 2 and $G_1$ and $G_2$ are conditional distributions of the events of type 1 and type 2, respectively, that are two-parameter Gumbel distributions. This model, unlike the TCEV one, allows to deal with scenarios in which events of type 2 do not occur every year. It also exploits more information than the TCEV one, by taking into account the label that identifies the type of event associated to the annual maximum, but there is, however, one additional parameter to estimate. Furthermore, knowing a priori the population originating each event is often difficult in practice. Even when the data are labelled with the type of event, it is not always the case that these labels actually identify the various subgroups in the tail population. Indeed, the events that happen infrequently are not always the most extreme. In this research we aim at using finite mixture models incorporating the labels as a covariate of the model rather than a deterministic identifier.

DRAFT

## 2. Models for multi-component extremes

We are interested in finite mixture models for the analysis of series of maximum events originated from multiple populations, focusing on a flexible description of the tail of the distribution. We begin from models which assume that the labels are a known deterministic identifier, and extend the idea to allow for noisy labels.

### 2.1 Finite mixture models for fixed categories

Without loss of generality, we consider data from two different processes. In the TCEV model (5) the CDF of the annual maximum is written as the product of two Gumbel CDF's: one with location parameter $\mu = \theta_1 \log \lambda_1$ and scale parameter $\sigma = \theta_1$, and the other with location parameter $\mu = \theta_2 \log \lambda_2$ and scale parameter $\sigma = \theta_2$. Table 1 shows the average proportion of type 2 events in the 10 most extreme values in 500 samples of 1000 simulations from a TCEV model with different choices of ratios of parameters, accounting for the fact that by definition of the model $\lambda_1 > \lambda_2$. It is not straightforward to understand how the combination of parameters controls the prevalence in the tail, and it is difficult also

| | $\lambda_1/\lambda_2 = 1.5$ | $\lambda_1/\lambda_2 = 2$ | $\lambda_1/\lambda_2 = 5$ |
|---|---|---|---|
| $\theta_2/\theta_1 = 0.5$ | 0.0018 (0.2213) | 0.0012 (0.1565) | 0.0000 (0.0357) |
| $\theta_2/\theta_1 = 1$ | 0.4042 (0.3969) | 0.3244 (0.3321) | 0.1732 (0.1664) |
| $\theta_2/\theta_1 = 1.5$ | 0.9090 (0.5135) | 0.8794 (0.4591) | 0.7624 (0.2999) |
| $\theta_2/\theta_1 = 2$ | 0.9924 (0.5902) | 0.9892 (0.5450) | 0.9722 (0.4055) |

Table 1: Average proportion of events of type 2 in the 10 most extreme ones obtained from 500 series of 1000 simulations of the TCEV model with the corresponding ratios of location and scale parameters. Average proportion of type 2 events in the whole sample in parenthesis. For reference $\lambda_2 = 2$ always and $\theta_1 = 1$ everywhere except in the first row, where it is 2.

to acknowledge the role in the parameters in determining the total proportion of type 2 events (displayed in parenthesis in the table). Indeed, for some choices of ratios, type 2 events are not the rarest ones, but are still the strongest in terms of magnitude.

The Mixture Gumbel model (4) defines the CDF of the annual maximum as

$$F_X(x) = (1-\omega)\exp\left\{-\exp\left[-\frac{x-\mu_1}{\sigma_1}\right]\right\} + \omega\exp\left\{-\exp\left[-\frac{x-\mu_2}{\sigma_2}\right]\right\}, \tag{1}$$

with $\omega < 0.5$ since, as in (5), events of type 1 are expected to occur more frequently. The parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ are the location and scale parameters of the two Gumbel distributions of events from process 1 and 2, respectively. Figure 1 shows how the distribution of 1000 simulated data from this model (with $\omega = 0.2$) changes with different ratios between the location parameters and the scale parameters. The location parameter $\mu_1$ is kept equal to 10, while the scale parameter $\sigma_1$ is put equal to 3, except in the case $\mu_2/\mu_1 = 1.5$ and $\sigma_2/\sigma_1 = 0.5$, when it is 4, and in the case $\mu_2/\mu_1 = 2$ and $\sigma_2/\sigma_1 = 0.5$, with $\sigma_1 = 5$. Every histogram displays the distribution of the simulated mixture model with corresponding choices of the parameters, and the proportion of density in each bin that comes from each process is identified by using two different colours, light for process 1 and dark for process 2.

It is possible to recognise that the conditional distribution dominating the tail of the mixture can change depending on the parameter combinations: with some choices of the parameters the rare events (type 2) do not prevail in the right tail and frequent events (type 1) also can correspond to the most extreme values. This is also noticeable in Table 2, which displays the proportion of type 2 events in the 10 most extreme values among 1000 simulations (top 1%) from the Mixture Gumbel model with the same ratios of scale and location parameters as Figure 1. As expected type 2 events are the majority when both $\mu_2/\mu_1$ and $\sigma_2/\sigma_1$ are large. These results indicate that the model is flexible and can represent multiple scenarios, which is appealing since real life data may also behave in this way. Therefore, it is preferred over the TCEV one as a starting point of our model.

| | $\mu_2/\mu_1 = 1$ | $\mu_2/\mu_1 = 1.5$ | $\mu_2/\mu_1 = 2$ |
|---|---|---|---|
| $\sigma_2/\sigma_1 = 0.5$ | 0.0062 | 0.0350 | 0.1030 |
| $\sigma_2/\sigma_1 = 1$ | 0.2800 | 0.5408 | 0.8482 |
| $\sigma_2/\sigma_1 = 1.5$ | 0.7254 | 0.9276 | 0.9632 |
| $\sigma_2/\sigma_1 = 2$ | 0.9296 | 0.9742 | 0.9902 |

Table 2: Average proportion of events of type 2 in the 10 most extreme ones obtained from 500 series of 1000 simulations of the Mixture Gumbel model with the corresponding ratios of location and scale parameters. The mixing parameter $\omega$ is equal to 0.2.

DRAFT

## 2.2 Finite mixture models for uncertain categories

Although the division of the data point made using the labels may make sense from a physical point of view, it is not necessarily appropriate for describing the tails of the distribution, hence we want to use them to inform the inference but not completely condition it, using also other variables to enhance the model. Therefore, rather than having pre-fixed groups using the labels, we want to allow the data to be informative and to let the allocation be driven by them. Without loss of generality, we assume that the data come from two populations.

Considering the data $x = (x_1, \ldots, x_n)$ and the latent allocation variables $z = (z_1, \ldots, z_n)$, with $z_i$ that identifies the mixture component $x_i$ belongs to, and the denoting by $\ell = (\ell_1, \ldots, \ell_n)$ an observed vector of binary labels such that $\ell_i$ indicates the type of event generating the data point $x_i$, for $i = 1 \ldots, n$, we define the model

$$\begin{aligned} x_i \mid z_i = j &\overset{ind}{\sim} \text{Gumbel}\,(x_i \mid \theta_j)\,; \quad \theta_j = (\mu_j, \sigma_j) \\ z_i \mid \ell_i &\overset{ind}{\sim} \text{Bernoulli}\,(z_i \mid \omega_i)\,, \\ \omega_i &= \frac{\exp(\beta_0 + \beta_1 \ell_i)}{1 + \exp(\beta_0 + \beta_1 \ell_i)}, \quad i = 1, \ldots, n. \end{aligned} \tag{2}$$

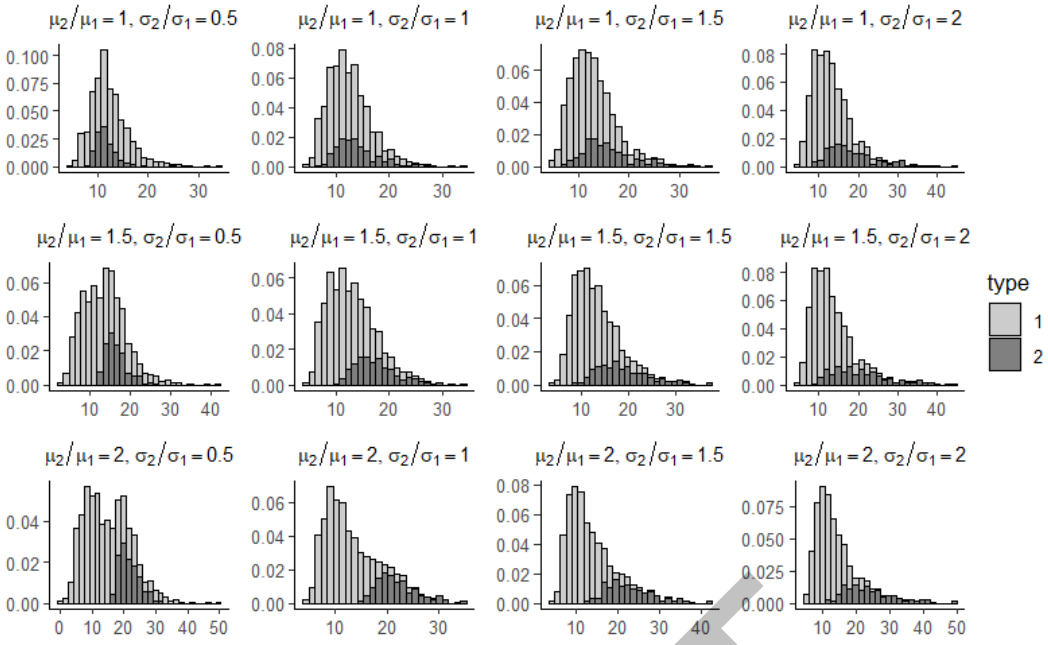



Figure 1: Histograms of the distribution of 1000 simulated data from the Mixture Gumbel model with different ratios of the location parameters and the scale parameters. The mixing parameter $\omega$ is always equal to 0.2. For each bin the area is coloured according to the proportion that is due to process 1 (light) and process 2 (dark).

Thus $\omega_i$ is derived from a logistic regression as a function of the labels, but this could be generalised to include additional covariates informing the data allocation in the tails. This hierarchical structure allows a more flexible modelling of extreme events that do not originate from a single population and, since the labelling is data-driven, it does not require the labels to be known.

## 3. Preliminary results

We carry out a simulation study to assess what happens when the labels are not a good representation of the mixture allocation for the tails, using both Model 1 and Model 2. Indeed, we explore how the results change and how robust they are by considering also a scenario where the labels are not the actual identifiers of the type of process, but a small percentage of them is swapped and therefore wrong. We produce 500 series of $n = 1000$ annual maximum events generated from a Mixture Gumbel model with parameters $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5$ and $\omega = 0.2$, and we estimate the parameters by numerical maximum likelihood. For Model 2 we allocate to type 2 the data points with the estimated $\omega_i$ greater or equal to 0.5, and the other ones to type 1. For comparison with the Mixture Gumbel model, an estimate of a single $\omega$ is then computed as the number of units assigned to type 2 over the total. In Model 1 the parameter $\omega$ is simply estimated as the ratio between the number of observed type 2 events and the total number of events.

Tables 3 and 4 show the maximum likelihood estimates related to the two different likelihoods when the labels are the actual ones and when 10% of them are wrong, respectively. It is possible to notice that if the true labels are provided (so they actually identify the type of process) both models can accurately estimate the parameters, whereas when the information on the type of process is wrong, even by a small percentage, we get a more robust model by using the binary regression idea (Model 2). Similarly, the return level plots in 2 show that when the labels are not a good representation of the mixture allocation the model that solely relies on them for estimation (Model 1) is not able to correctly capture high quantiles while Model 2 does.

DRAFT

| | Model with fixed $\omega$ | | | | Model with varying $\omega_i$s | | | |
|---|---|---|---|---|---|---|---|---|
| | 1st Qu. | Median | Mean | 3rd Qu. | 1st Qu. | Median | Mean | 3rd Qu. |
| $\mu_1$ | 9.985 | 10.06 | 10.06 | 10.13 | 9.957 | 10.03 | 10.03 | 10.11 |
| $\sigma_1$ | 2.998 | 3.048 | 3.052 | 3.105 | 2.958 | 3.016 | 3.016 | 3.079 |
| $\mu_2$ | 20.12 | 20.37 | 20.39 | 20.63 | 20.12 | 20.38 | 20.40 | 20.65 |
| $\sigma_2$ | 4.708 | 4.887 | 4.899 | 5.078 | 4.695 | 4.866 | 4.878 | 5.061 |
| $\omega$ | 0.193 | 0.200 | 0.200 | 0.208 | 0.193 | 0.200 | 0.200 | 0.208 |

Table 3: Summary of 500 maximum likelihood estimates of the parameters of a Mixture Gumbel model with $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5, \omega = 0.2$. On the left the estimates assuming Model 1 and on the right the ones from Model 2.

| | Model with fixed $\omega$ | | | | Model with varying $\omega_i$s | | | |
|---|---|---|---|---|---|---|---|---|
| | 1st Qu. | Median | Mean | 3rd Qu. | 1st Qu. | Median | Mean | 3rd Qu. |
| $\mu_1$ | 10.01 | 10.09 | 10.09 | 10.17 | 9.949 | 10.04 | 10.04 | 10.12 |
| $\sigma_1$ | 3.027 | 3.088 | 3.091 | 3.152 | 2.958 | 3.023 | 3.025 | 3.102 |
| $\mu_2$ | 16.50 | 16.78 | 16.77 | 17.03 | 19.90 | 20.36 | 20.21 | 20.68 |
| $\sigma_2$ | 6.477 | 6.649 | 6.655 | 6.828 | 4.695 | 4.914 | 4.979 | 5.185 |
| $\omega$ | 0.252 | 0.260 | 0.260 | 0.268 | 0.252 | 0.260 | 0.260 | 0.268 |

Table 4: Summary of 500 maximum likelihood estimates of the parameters of a Mixture Gumbel model with $\mu_1 = 10, \mu_2 = 20, \sigma_1 = 3, \sigma_2 = 5, \omega = 0.2$ and 10% of the labels swapped. On the left the estimates assuming Model 1 and on the right the ones from Model 2.

DRAFT

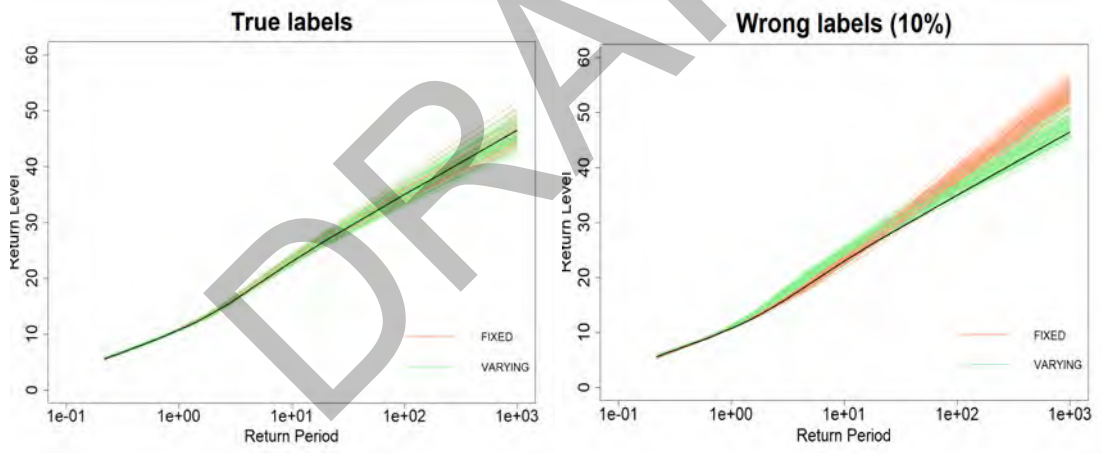



Figure 2: Return level plots corresponding to 100 different estimates of the parameters of the Mixture Gumbel model using Model 1 (pink) and Model 2 (green), when the labels are true on the left and when 10% are wrong on the right. In black the return level curve obtained with the true values of the parameters.

## 4. Discussion

When dealing with real world problems, it is reasonable to assume that extremes events originate from a number of different types of data-generating processes. However it is not easy to find data on extreme events that have information about the type of phenomenon which generated them. A further issue that is not very evident in the literature concerns the fact that even when this information is made available by domain experts, is not necessarily one that allows to discriminate between the multiple groups in the tail of the population. We define a model in which the allocation of the data points is not solely based on the knowledge on the type of process, i.e. the labels, but it exploits them to inform the

inference, leading to results that are more robust to noise in the labels.

We are interested in enhancing the model by using Bayesian methods to exploit the ability of setting priors which encode the understanding we have of the problem, and to borrow information between the groups to estimate the model parameters. The Bayesian setting also allows to integrate in the estimation the uncertainty in the labels. Furthermore, we aim at exploring the results of the simulation study with real data applications, which are usually not characterised by such a high number of observations, with a focus on the problem of unknown categories.

## References


[1] S. Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

[2] S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert. *Handbook of mixture analysis*. CRC press, 2019.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[4] T. R. Kjeldsen, H. Ahn, I. Prosdocimi, and J.-H. Heo. Mixture gumbel models for extreme series including infrequent phenomena. *Hydrological Sciences Journal*, 63(13-14):1927–1940, 2018.

[5] F. Rossi, M. Fiorentino, and P. Versace. Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20(7):847–856, 1984.


DRAFT

# Robust matrix-variate mixtures of regressions

Salvatore D. Tomarchio[a] and Michael P. B. Gallaugher[b]

[a]University of Catania, Department of Economics and Business, Catania, Italy;
daniele.tomarchio@unict.it
[b]Baylor University, Department of Statistical Science, Waco, TX, USA;
michael_gallaugher@baylor.edu

**Abstract**

Finite mixtures of regressions are a model-based clustering approach commonly adopted in many regression-type analyses. Unfortunately, real data often present atypical points that make unreliable the classical normality assumption of the mixture components. Thus, to robustify this approach, we introduce finite mixtures of matrix-variate contaminated normal regressions (FM-MVCNR). Furthermore, once the model is estimated and the observations are assigned to the groups, a finer intra-group classification in typical and atypical points can be directly obtained. From the analyses conducted on simulated data, we show the negative consequences of the wrong normality assumption in the presence of heavy-tailed clusters or noisy matrices, and how they are properly addressed by our FM-MVCNR.

***Keywords:*** matrix-variate, robustness, mixture-models, regression

DRAFT

## 1. Introduction

The analysis of matrix-variate data is attracting a growing interest in the statistical literature, particularly within the model-based clustering framework (for recent contributions, see e.g. (1; 6; 10; 13; 15)). A matrix-variate dataset is characterized by three modes, namely $p$ rows, $r$ columns, and $n$ layers. Depending on the entities indexed in the three modes, different examples are obtained (for a survey, see (17)).

When the data at hand are composed of a $p \times r$ response matrix $\mathcal{Y}$ and a $q \times r$ covariate matrix $\mathcal{X}$, and when there is a latent source of heterogeneity, finite mixtures of matrix-variate regression (FM-MVR) models constitute a reference framework of analysis (7; 14; 2; 9). For a continuous response matrix $\mathcal{Y}$, attention is generally focused on finite mixtures of matrix-variate normal regressions (FM-MVNR) because of their computational and theoretical convenience. Unfortunately, real data are often contaminated by atypical points that affect the estimation of the model parameters and data classification (5). Thus, their detection, and the development of robust methods insensitive to their presence, is a crucial task.

Based on the above considerations, in this work, we introduce finite mixtures of matrix-variate contaminated normal regressions (FM-MVCNR). The matrix-variate contaminated normal (MVCN), used for modeling the conditional distribution of the responses, is a heavy-tailed generalization of the matrix-variate normal (MVN) distribution. More in detail, it is a two-component matrix-variate normal mixture in which one of the components, with a large prior probability, represents the typical observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the atypical observations. Because of its heavier-than-normal tails, it provides greater flexibility in modeling data with heavy-tailed groups. This distribution has been successfully considered for the definition of unconditional matrix-variate mixture models in (16).

As it will be better explained in Sect. 2, once the FM-MVCNR is fitted to the data, each observation can be first assigned to one of the groups and then classified as either typical or atypical using maximum *a posteriori* probabilities. Thus, we have a model for simultaneous clustering and detection of atypical points in a matrix-variate regression context. This aspect is of particular importance for matrix-variate data given that visualization techniques - and, therefore, the visual detection of atypical points - are a problematic exercise.

The paper is organized as follows. In Sect. 2, we introduce the FM-MVCNR by providing information about its characteristics, parameter estimation, and atypical points detection. In Sect. 3, we conduct a simulated analysis where we show the negative consequences of the wrong normality assumption in the presence of heavy-tailed clusters or noisy matrices, and how they are properly addressed by our FM-MVCNR. Finally, in Sect. 4. we summarize our manuscript.

## 2. Methodology

Let $\mathcal{Y}$ be a continuous random matrix of dimension $p \times r$ containing $p$ responses measured over $r$ occasions. Let us also consider a random matrix $\mathcal{X}$ of dimension $q \times r$ containing $q$ covariates evaluated over $r$ occasions. Assume there exist $K$ subgroups in the data. Then, the probability density function (pdf) of a matrix-variate FM-MVR is

$$h(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta}_{\mathbf{Y}|k}), \tag{1}$$

where $\pi_k > 0$ is the mixing proportion, with $\sum_{k=1}^{K} \pi_k = 1$, $f(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta}_{\mathbf{Y}|k})$ is the conditional distribution of the responses, and $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{\Theta}_{\mathbf{Y}|k}\}_{k=1}^{K}$. Additionally, for each $k$, the location parameter in $\boldsymbol{\Theta}_{\mathbf{Y}|k}$ is a linear function of $\mathbf{X}$ depending on some other parameters.

In this manuscript, $f(\mathbf{Y}|\mathbf{X};\boldsymbol{\Theta}_{\mathbf{Y}|k})$ in (1) has the functional form of the matrix-variate contaminated normal distribution, whose pdf for a generic $p \times r$ random matrix $\mathcal{H}$ is

$$\alpha\phi(\mathbf{H};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}) + (1-\alpha)\phi(\mathbf{H};\mathbf{M},\eta\boldsymbol{\Sigma},\boldsymbol{\Psi}), \tag{2}$$

where $\phi(\mathbf{H};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})$ is the density of the MVN distribution having pdf

$$\frac{1}{(2\pi)^{\frac{pr}{2}}|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}} \exp\left[-\frac{\delta(\mathbf{H};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})}{2}\right], \tag{3}$$

with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ and $r \times r$ covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, respectively, and squared Mahalanobis distance $\delta(\mathbf{H};\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}) = \text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{H}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{H}-\mathbf{M})'\right]$. In (2), $\alpha$ is the proportion of typical points in the group $k$, and $\eta$ is an inflation parameter accounting for the degree of contamination in the group $k$. We assume a linear relationship $\mathbf{M}(\mathbf{X}^*;\mathbf{B}_k) = \mathbf{B}_k\mathbf{X}^*$, where $\mathbf{B}_k$ is a $p \times (1+q)$ matrix of regression coefficients and $\mathbf{X}^*$ is a $(1+q) \times r$ matrix containing a first row of ones (to incorporate the intercept in the model) and the covariates $\mathbf{X}$. Note that (2) approaches (3) as $\alpha \to 1^-$ and $\eta \to 1^+$.

We recall that an MVCN distribution can be represented as a scale mixture of MVN distributions with mixing random variable $W$ being Bernoulli distributed (16). This characteristic is particularly useful for maximum likelihood parameter estimation, as briefly discussed in Section 2.1.

### 2.1 Parameter estimation

To estimate the parameters of our model, we implement the expectation conditional-maximization (ECM) algorithm (8). In detail, let $\mathbf{S} = \{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^{N}$ be a sample of $N$ independent observations. Within an ECM framework, $\mathbf{S}$ is viewed as incomplete: we have a first source of incompleteness that arises from the fact that we do not know the component membership of each observation, and a second source of incompleteness that arises from the aforementioned representation of the MVCN distribution in terms of a scale

DRAFT

mixture of MVN distributions. To address the first source, we use an indicator vector $\mathbf{z}_i = (z_{ik}, \ldots, z_{iK})$, where $z_{ik} = 1$ if observation $i$ is in the group $k$, and $z_{ik} = 0$ otherwise. To consider the second source, for each observation $(\mathbf{Y}_i, \mathbf{X}_i)$ in the group $k$, we have $W_{ik\mathbf{Y}} \sim$ Bernoulli $(\alpha_{\mathbf{Y}|k})$. Based on these sources of incompleteness, we can write the complete-data log-likelihood as

$$
\begin{aligned}
l_c(\boldsymbol{\Theta}) = & \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln(\pi_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[w_{ik\mathbf{Y}} \ln(\alpha_{\mathbf{Y}|k}) + (1 - w_{ik\mathbf{Y}}) \ln(1 - \alpha_{\mathbf{Y}|k})\right] \\
& + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[-\frac{r}{2}\ln|\boldsymbol{\Sigma}_{\mathbf{Y}|k}| - \frac{p}{2}\ln|\boldsymbol{\Psi}_{\mathbf{Y}|k}| - \frac{pr}{2}(1 - w_{ik\mathbf{Y}})\ln(\eta_{\mathbf{Y}|k})\right. \\
& \left. -\frac{1}{2}(w_{ik\mathbf{Y}} + \frac{1 - w_{ik\mathbf{Y}}}{\eta_{\mathbf{Y}|k}})\delta_k\left(\mathbf{Y}_i; \mathbf{B}_k\mathbf{X}_i^*, \boldsymbol{\Sigma}_{\mathbf{Y}|k}, \boldsymbol{\Psi}_{\mathbf{Y}|k}\right)\right].
\end{aligned} \tag{4}
$$

The ECM algorithm of our model consists of an E-Step and two CM-Steps. At the E-step, we calculate the conditional expectation of (4), given the observed data and the current estimate of the parameters. Then, we maximize the conditional expectation of (4) with respect to two separate sets of parameters, one for each CM-step. Here, we limit to report that closed-form expressions are available for all the parameters of the model, in the fashion of (16).

## 2.2 Atypical points detection

An interesting feature of the FM-MVCNR is the capability of detecting atypical observations. To label data points, we consider an *a posteriori* procedure. First of all, each observation $(\mathbf{Y}_i, \mathbf{X}_i)$ is assigned to one of the $K$ groups through the maximum *a posteriori* probabilities (MAP) operator

$$
\text{MAP}(\widehat{z}_{ik}) = \begin{cases} 1 & \text{if } \max_h\{\widehat{z}_{ih}\} \text{ occurs in group } h = k, \\ 0 & \text{if otherwise,} \end{cases}
$$

where $\widehat{z}_{ik}$ is the estimated posterior probability that a point $(\mathbf{Y}_i, \mathbf{X}_i)$ belongs to the $k$th component of the model. Then, let $\widehat{w}_{ik}$ be the estimated value of $W_{ik}$ at the convergence of the ECM algorithm. The commonly adopted decision rule when contaminated distributions are used (see, e.g. 4; 12; 16), consists of considering a point typical if $\widehat{w}_{ik} > 0.5$. Thus, once the observation has been classified in one of the $K$ groups, this approach reveals richer information about the role of that observation in that group.

DRAFT

# 3. Synthetic data analyses

Here, we simulate different scenarios that may arise when analyzing real datasets. Specifically, we generate data from the following three data generation processes (DGPs): (a) FM-MVNR, (b) FM-MVCNR, and (c) FM-MVNR where 5% of points have been modified to incorporate noisy values. Regarding scenario (c), for each selected point, the values of $\mathbf{Y}$ are substituted by random numbers generated from a uniform distribution over the interval $[a_{\mathbf{Y}}, b_{\mathbf{Y}}]$, being $a_{\mathbf{Y}}$ and $b_{\mathbf{Y}}$ the minimum and the maximum value of $\mathbf{Y}$ in the sample.

We set $p = 2$, $q = 3$, $r = 5$, $K = 2$, and the following values for the parameters

$$
\pi_1 = \pi_2 = 0.50, \quad \alpha_{\mathbf{Y}} = (0.90, 0.80), \quad \eta_{\mathbf{Y}} = (15.00, 20.00),
$$

$$
\mathbf{B}_1 = \begin{bmatrix} 2.00 & 1.00 & 1.00 & -1.00 \\ 3.00 & 1.00 & -1.00 & 1.00 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} -10.00 & 1.00 & 1.00 & -1.00 \\ -8.00 & 1.00 & -1.00 & 1.00 \end{bmatrix},
$$

$$
\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{bmatrix}, \quad \boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = \begin{bmatrix} 1.00 & 0.70 & 0.49 & 0.34 & 0.24 \\ 0.70 & 1.00 & 0.70 & 0.49 & 0.34 \\ 0.49 & 0.70 & 1.00 & 0.70 & 0.49 \\ 0.34 & 0.49 & 0.70 & 1.00 & 0.70 \\ 0.24 & 0.34 & 0.49 & 0.70 & 1.00 \end{bmatrix}.
$$

For each DGP, we simulate 100 samples of size $N$, where $N \in (100, 500)$, yielding a total of 600 generated datasets. According to the above simulation scheme, we have scenarios characterized by no atypical points (a), heavy-tailed clusters (b), and noisy points (c). On each generated dataset we fit the FM-MVNR and FM-MVCNR with $K = 2$, and their performances are discussed in Sect. 3.1 in terms of parameter recovery, classification, and fitting performance.

## 3.1 Results

First of all, we examine the parameter recovery of the considered models. To this purpose, we compute the average mean squared error (MSE) of the estimated regression coefficients over the 100 datasets simulated by each DGP. Results are reported in Table 1.

Table 1: Average MSEs of the estimated regression coefficients under the three scenarios.

| Model | | $N = 100$ | $N = 500$ |
|---|---|---|---|
| *Scenario (a)* | | | |
| FM-MVNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.008 & 0.004 & 0.005 & 0.005 \\ 0.012 & 0.005 & 0.004 & 0.006 \end{bmatrix}$ | $\begin{bmatrix} 0.002 & 0.001 & 0.001 & 0.001 \\ 0.002 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.012 & 0.005 & 0.006 & 0.006 \\ 0.011 & 0.005 & 0.004 & 0.006 \end{bmatrix}$ | $\begin{bmatrix} 0.003 & 0.001 & 0.001 & 0.001 \\ 0.003 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |
| FM-MVCNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.008 & 0.004 & 0.005 & 0.005 \\ 0.012 & 0.005 & 0.004 & 0.006 \end{bmatrix}$ | $\begin{bmatrix} 0.002 & 0.001 & 0.001 & 0.001 \\ 0.002 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.012 & 0.005 & 0.006 & 0.006 \\ 0.011 & 0.005 & 0.004 & 0.006 \end{bmatrix}$ | $\begin{bmatrix} 0.003 & 0.001 & 0.001 & 0.001 \\ 0.003 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |
| *Scenario (b)* | | | |
| FM-MVNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.406 & 0.003 & 0.003 & 0.002 \\ 0.324 & 0.002 & 0.002 & 0.002 \end{bmatrix}$ | $\begin{bmatrix} 0.175 & 0.000 & 0.000 & 0.001 \\ 0.163 & 0.000 & 0.001 & 0.001 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.402 & 0.002 & 0.003 & 0.003 \\ 0.303 & 0.002 & 0.003 & 0.002 \end{bmatrix}$ | $\begin{bmatrix} 0.140 & 0.000 & 0.001 & 0.000 \\ 0.129 & 0.000 & 0.000 & 0.000 \end{bmatrix}$ |
| FM-MVCNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.014 & 0.001 & 0.001 & 0.001 \\ 0.016 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ | $\begin{bmatrix} 0.003 & 0.000 & 0.000 & 0.000 \\ 0.004 & 0.000 & 0.000 & 0.000 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.020 & 0.001 & 0.001 & 0.001 \\ 0.019 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ | $\begin{bmatrix} 0.004 & 0.000 & 0.000 & 0.000 \\ 0.004 & 0.000 & 0.000 & 0.000 \end{bmatrix}$ |
| *Scenario (c)* | | | |
| FM-MVNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.347 & 0.068 & 0.066 & 0.052 \\ 0.523 & 0.047 & 0.043 & 0.069 \end{bmatrix}$ | $\begin{bmatrix} 0.330 & 0.021 & 0.024 & 0.018 \\ 0.489 & 0.023 & 0.020 & 0.021 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.687 & 0.071 & 0.062 & 0.072 \\ 0.372 & 0.045 & 0.073 & 0.061 \end{bmatrix}$ | $\begin{bmatrix} 0.736 & 0.023 & 0.025 & 0.038 \\ 0.372 & 0.031 & 0.030 & 0.031 \end{bmatrix}$ |
| FM-MVCNR | $\mathbf{B}_1$ | $\begin{bmatrix} 0.016 & 0.007 & 0.006 & 0.005 \\ 0.014 & 0.006 & 0.004 & 0.005 \end{bmatrix}$ | $\begin{bmatrix} 0.003 & 0.001 & 0.001 & 0.001 \\ 0.003 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |
| | $\mathbf{B}_2$ | $\begin{bmatrix} 0.016 & 0.006 & 0.005 & 0.007 \\ 0.015 & 0.005 & 0.006 & 0.006 \end{bmatrix}$ | $\begin{bmatrix} 0.002 & 0.001 & 0.001 & 0.001 \\ 0.002 & 0.001 & 0.001 & 0.001 \end{bmatrix}$ |

DRAFT

By starting with scenario (a), i.e. when there are no atypical points, we see that both models provide the same results. This is because, in this situation, the FM-MVCNR tends to the FM-MVNR (refer to Sect. 2). Regardless of the considered model, the MSEs are negligible and improve with the increase of

$N$. Oppositely, when scenario (b) is considered, the FM-MVCNR performs better than the traditional FM-MVNR. Indeed, the latter model is not robust, with consequences on the parameter estimates, as we can see by the MSEs that are regularly higher than those of the FM-MVCNR. As in scenario (a), the MSEs improve with the increase of $N$. Lastly, under scenario (c), i.e. when there are noisy matrices, we immediately observe that the estimates of the FM-MVNR are even worse than those of scenario (b). On the contrary, because of its robustness, the FM-MVCNR performs comparably well, as shown by the negligible MSEs that become better as $N$ increases.

Regarding the classification and fitting performance, we compute the average adjusted Rand index (ARI; 3) and average Bayesian information criterion (BIC; 11) of the models over the 100 datasets simulated by each DGP. Results are illustrated in Table 2. As we can see, under scenario (a), both models always provide a perfect data classification. However, in scenarios (b) and (c), the FM-MVCNR, because of its robustness, shows better results than the FM-MVNR. Similarly, from the fitting point of view, we notice that the FM-MVNR produces better BICs under scenario (a). However, in scenarios (b) and (c), the FM-MVCNR displays far better BICs, highlighting its greater flexibility in modeling data having atypical points.

Table 2: Average ARI and BIC for the considered models over the simulated datasets.

| Model | ARI | | BIC | |
|---|---|---|---|---|
| | $N = 100$ | $N = 500$ | $N = 100$ | $N = 500$ |
| *Scenario (a)* | | | | |
| FM-MVNR | 1.000 | 1.000 | 2476.279 | 11741.900 |
| FM-MVCNR | 1.000 | 1.000 | 2494.699 | 11766.760 |
| *Scenario (b)* | | | | |
| FM-MVNR | 0.894 | 0.943 | 3888.669 | 19301.020 |
| FM-MVCNR | 0.967 | 0.980 | 3198.380 | 15227.990 |
| *Scenario (c)* | | | | |
| FM-MVNR | 0.805 | 0.810 | 3958.796 | 19873.750 |
| FM-MVCNR | 0.811 | 0.815 | 3027.217 | 14507.830 |

DRAFT

For evaluating the performances of the FM-MVCNR in detecting atypical points in scenario (c), we consider the true positive rate (TPR), measuring the proportion of atypical points that are correctly identified as atypical, and the false positive rate (FPR), corresponding to the proportion of typical points incorrectly classified as atypical. Their average values over the 100 datasets simulated for each $N$ are reported in Table 3. From the obtained results, we note that our model always recognizes the noisy matrices as atypical, and very rarely labels a typical observation as atypical.

Table 3: Average TPRs and FPRs under Scenario (c) over the simulated datasets.

| | TPR | | FPR | |
|---|---|---|---|---|
| Model | $N = 100$ | $N = 500$ | $N = 100$ | $N = 500$ |
| FM-MVCNR | 1.000 | 1.000 | 0.008 | 0.006 |

## 4. Conclusions

In this manuscript, robust finite mixtures of matrix-variate regressions have been introduced. Specifically, we considered the matrix-variate contaminated normal (MVCN) for modeling the conditional distribution of the responses in the mixture components. The results of our simulated analyses showed

that in the presence of heavy-tailed clusters or noisy matrices, the mean squared errors of the regression coefficients are regularly higher when the normal distribution is used for the mixture components. Conversely, the use of the MVCN distribution, because of its heavier tails, results in more robust parameter estimation, classification, and fitting compared to the traditional case based on the matrix-variate normal distribution. Lastly, our model performs very well in detecting atypical points in the data.

## References


[1] Gallaugher, M. P., McNicholas, P. D.: Finite mixtures of skewed matrix variate distributions. Pattern Recognit. **80**, 83–93 (2018)

[2] Gallaugher, M. P., Tomarchio, S. D., McNicholas, P. D., Punzo, A.: Model-based clustering via skewed matrix-variate cluster-weighted models. J. Stat. Comput. Simul. **92**(13), 2645–2666 (2022)

[3] Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

[4] Maruotti, A., Punzo, A.: Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. Comput. Stat. Data Anal. **113**, 475–496 (2017)

[5] Mazza, A., Punzo, A.: Mixtures of multivariate contaminated normal regression models. Stat. Pap **61**(2), 787–822 (2020)

[6] Melnykov, V., Zhu, X.: On model-based clustering of skewed matrix data. J. Multivar. Anal. **167**, 181–194 (2018)

[7] Melnykov, V., Zhu, X.: Studying crime trends in the USA over the years 2000-2012. Adv. Data Anal. Classif. **13**, 325–341 (2019)

[8] Meng, X. L., Rubin, D. B.: Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**(2), 267–278 (1993)

[9] Punzo, A., Tomarchio, S. D.: Parsimonious Finite Mixtures of Matrix-Variate Regressions. In Bekker, A., Ferreira, J. T., Arashi, M., Chen, D. (eds.) Innovations in Multivariate Statistical Modeling: Navigating Theoretical and Multidisciplinary Domains, 385-398. Springer, Cham (2022).

[10] Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)

[11] Schwarz, G.: Estimating the dimension of a model. Ann. Stat. 461–464 (1978)

[12] Tomarchio, S. D., Punzo, A.: Dichotomous unimodal compound models: application to the distribution of insurance losses. J. Appl. Stat. **47**(13-15), 2328–2353 (2020)

[13] Tomarchio, S. D., Punzo, A., Bagnato, L.: Two new matrix-variate distributions with application in model-based clustering. Comput. Stat. Data Anal. **152**, 107050 (2020)

[14] Tomarchio, S. D., McNicholas, P. D., Punzo, A.: Matrix normal cluster-weighted models. J. Classif. **38**(3), 556–575 (2021)

[15] Tomarchio, S. D.: Matrix-variate normal mean-variance Birnbaum-Saunders distributions and related mixture models. Comput. Stat. 1–28 (2022)

[16] Tomarchio, S. D., Gallaugher, M. P., Punzo, A., McNicholas, P. D.: Mixtures of matrix-variate contaminated normal distributions. J. Comput. Graph. Stat. **31**(2), 413–421 (2022)

[17] Viroli, C.: Model based clustering for three-way data structures. Bayesian Anal. **6**(4), 573–602 (2011)


DRAFT

# On the use of auxiliary information to define the sampling design for large-scale geospatial data

Chiara Bocci and Emilia Rocco

Department of Statistics, Computer Science, Applications "G. Parenti" - University of Florence
chiara.bocci@unifi.it, emilia.rocco@unifi.it

**Abstract**

In many fields of application it's common to be interested in spatially-related phenomena and in particular to deal with attributes that, being defined on continuous spatial domains, are observed on a fine grid. It is well known that in order to investigate these phenomena through a sample survey it is more accurate to spread the sample over space and to exploit the rich auxiliary information available from various types of large-scale observations. Such information is often used in the estimation stage of a survey but it may be interesting to use it in the design stage as well. Therefore, we propose a two-step sampling design which investigates the relation between the auxiliary and study variables in order to identify when and how it is useful to exploit this information, in addition to the units' spatial location, in the the second-step of the sampling selection process.

***Keywords:*** Balanced sampling, Cross-correlation, Local pivotal method, Spread sampling, Unequal probability sampling

DRAFT

## 1. Introduction

In many fields of application including forestry, geology, ecology, and similar, data contains geographical coordinates and such information can be used in the sampling design development process. Since nearby units interact with one another and tend to be influenced by the same set of natural and anthropogenic factors, geographical data generally show a spatial pattern and an uneven spatial distribution over the population. In such situations, it is well known that selecting the units spatially well spread over the study area allows to collect more information and consequently provides a better estimation of the population parameters. Moreover, technological advances have led to a growing availability of low-cost spatial data ready-to-use, frequently derived from large-scale observations (including GPS sensors data, remote sensing data, and similar), and this auxiliary information can be used in the sampling design development process in addition to the units' spatial location.

Many sampling methods have been suggested in literature to select a well-spread sample, for a comprehensive review we refer to (7), (8), (1) and (9). Some of these methods implement simultaneously the selection of well-spread samples and the use of auxiliary variables in the selection process. Among these are the Spatially Balanced Sampling through the Local Pivotal Method (LPM) (4) and the Double Balanced Sampling (5). The first method allows the use of auxiliary variables in two different ways: (i) through the drawing of samples spread in a multidimensional space defined by the auxiliary variables along with the spatial coordinates, therefore producing samples that are balanced on the auxiliary variables as well as on the space; or (ii) through the selection of unequal probability spatially balanced samples, using an auxiliary variable or a function of more auxiliary variables for the calculus of the inclusion probabilities. The second method selects samples that are simultaneously well spread and balanced on the auxiliary variables by combining the use of the cube method (2) and the local pivotal method.

In the design-based approach, the use of auxiliary variables in the design phase and/or in the estimation phase is based on some assumptions (usually not explicit) on the relationship between the response and the auxiliary variables and the resulting gain in the estimation efficiency depends on the validity of these assumptions. For spatially related data, specific considerations on this relation are opportune: nearby units tend to be similar with respect to each variable and in some situations the relationship between the study and auxiliary variables may be wholly or partially due to a spatial cross-covariability. In such cases, the use of the auxiliary variable in addition to the units' geographical location does not always produce gains in efficiency. The relationship between the target and auxiliary variables is obviously never known exactly. Moreover, for large-scale phenomena, it is often not even plausible to assume a unique relationship that holds everywhere. In order to identify a sampling design that could be globally applied by accounting for different areas characteristics in both the study and auxiliary variables, as well as for the differences in their relation, we propose a two-step informative design based on the sequential use of the LPM: (i) in the first step, a well-spread sample is selected and the information collected is used to investigate the relation between the auxiliary and study variable; (ii) then, on the basis of this analysis, a decision is made on whether or not to consider the auxiliary variables in drawing the second-step well-spread sample. The final sample is given by the union of the two samples.

## 2. Notation and sampling strategy

Usually, in a spatial setting, the population units are plots or cells of a grid overlapping an area of interest. A value, $y_i$, of a variable of interest is associated with each unit $i$ $(i = 1, ..., N)$ of the population. Moreover for each unit the spatial location $\mathbf{s}_i$, $\mathbf{s} \in \mathbf{R^2}$ is known. Here, in addition we assume to know the value $x_i$ of an auxiliary variable for each unit of the population.

To draw a spatial sample from such a population we consider as starting point the spatially balanced sampling through LPM (4). The basic idea of LPM is to avoid that units close in distance appear together in the sample. First an inclusion probability $0 < \pi_i \leq 1$ is assigned to each unit so that their sum over the population is equal to the fixed sample size $n$. These probabilities may be constant or variable. When we have access to an auxiliary variable (commonly defined as a size measure) whose values are thought to be approximately proportional to the unknown response values we can set $\pi_i$ proportional to the known $x_i$. That is, for $i = 1, ..., N$

$$\pi_i = n \frac{x_i}{\sum_U x_j} \tag{1}$$

The sample is then obtained in at most $N$ steps. At each step one unit $i$ is selected randomly from the available population and another unit $j$ is chosen among the remaining units in the population by minimizing a distance function among them. This can be a univariate or a multivariate function that measures the distance with respect to one or more auxiliary variables, among which we can include the spatial coordinates. When all the variables are continuous the Euclidean distance is commonly used. Moreover, when multiple auxiliary variables are used, they should be standardized or scaled in order to balance the contribution of each variable. After the selection of the unit $i$ and $j$ their inclusion probabilities are updated by using the following rule:

$$\begin{aligned}
&\text{if } \pi_i + \pi_j < 1 \text{ then } (\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \end{cases} \\
&\text{if } \pi_i + \pi_j \geq 1 \text{ then } (\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{1 - \pi_j}{2 - \pi_i - \pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{1 - \pi_i}{2 - \pi_i - \pi_j} \end{cases}
\end{aligned} \tag{2}$$

As a result, in each step at least one unit is removed from the frame, either because its probability becomes zero, and consequently it is definitely excluded from the sample, or because its probability becomes one and therefore is included in the sample. The procedure continues, updating at each step the probabilities of inclusion obtained in the previous step, until all units in the population are processed.

DRAFT

We propose the following two-step sampling design. First, an initial sample $S_0$ of size $n_0 \leq n$ is selected by applying two times the spatially balanced sampling through LPM with constant inclusion probabilities $\pi_i$s: at the first time for drawing a sample $S$ of $n$ units from the population and at the second time for drawing $n_0$ units from $S$. Then the information collected from $S_0$ is used to investigate the relationship between the auxiliary and the study variable. On the basis of the results of this exploratory analysis, the second-step sample $S_1$ of size $n_1 = n - n_0$ is either assumed equal to $(S - S_0) = (S \bigcap S_0^c)$ or is drawn from $(U - S_0) = (U \bigcap S_0^c)$ using a spatially balanced sampling design that exploits also the auxiliary variable. The final sample is $S_f = (S_0 \cup S_1)$ and to estimate the population mean we use the Horvitz-Thompson type estimator with the following inclusion probabilities:

$$\pi_i' = Pr(i \in S_0) + (1 - Pr(i \in S_0) Pr(i \in S_1) =$$

$$= \begin{cases} \frac{n}{N} & \text{if } S_1 \text{ is drawn with constant probabilities} \\ \frac{n_0}{N} + \left(1 - \frac{n_0}{N}\right) \frac{(n-n_0)x_i}{\sum_U X_j - \sum_{S_0} X_j} & \text{otherwise} \end{cases} \tag{3}$$

We consider two ways for exploiting the auxiliary variable to draw the second-step sample:

1. **SeqUneqLPM**: apply a spatially balanced sampling through LMP with unequal inclusion probabilities $\pi_i$s proportional to the auxiliary variable;
2. **SeqBivLPM**: apply a sampling, with equal inclusion probabilities, balanced through LMP in the space spanned by both the geographical coordinates and the auxiliary variable.

## 3. Simulation study

In this section we compare our two proposals with several one-step designs all based on the use of the units' spatial location and/or of the auxiliary variable in the design phase:

3. **SpatLPM**: The original formulation of spatially balanced sampling through LPM which produces samples that are well spread in the geographic space and is based on equal inclusion probabilities;
4. **AuxLPM**: Sampling, with equal inclusion probabilities, balanced through LMP in the space spanned by the auxiliary variable;
5. **BivLPM**: Sampling, with equal inclusion probabilities, balanced through LMP in the space spanned by both the geographical coordinates and the auxiliary variable;
6. **UneqLPM**: Spatially balanced sampling through LMP with unequal inclusion probabilities $\pi_i$s proportional to the auxiliary variable;
7. **UneqSampford** Sampford sampling, with unequal inclusion probabilities proportional to the auxiliary variable

An additional design that we could consider for comparison is the double balanced sampling of (5), however this design is highly computationally demanding when applied to a big dataset, and would be unfeasible in our experiments. Conversely, the LPM design has been optimized for large datasets using k-d trees (6), allowing to run our Monte Carlo experiments in a reasonable amount of time.

We investigate the performance of the different sampling designs through Monte Carlo experiments based on several synthetic datasets. In each of them the auxiliary ($X$) and response ($Y$) variables are drawn from a stationary bivariate spatial process $[X(\mathbf{s}), Y(\mathbf{s})]$ with $\mathbf{s} \in [0, 10]^2$ ($1000 \times 1000$ grid). Following Diggle and Ribeiro (3, Chapter 3), each bivariate spatial process in turn is obtained as:

$$X(\mathbf{s}) = a * Z_1(\mathbf{s}) + c * Z_2(\mathbf{s}) + k_1 \qquad Y(\mathbf{s}) = b * Z_1(\mathbf{s}) + d * Z_3(\mathbf{s}) + k_2 \tag{4}$$

- $Z_1(\mathbf{s})$, $Z_2(\mathbf{s})$ and $Z_3(\mathbf{s})$ are independent univariate stationary Gaussian processes with an Exponential variogram with scale 1 and sill $C + C_0 = 50$, where $C$ is the partial sill and $C_0$ is the nugget. For $Z_2(\mathbf{s})$ and $Z_3(\mathbf{s})$ we assume $C = 50$ and $C_0 = 0$ in all scenarios, while for $Z_1(\mathbf{s})$ their values vary ($C = 50, 30, 0$) to change the proportion of the co-variability that has spatial structure;

DRAFT

- $a$, $b$, $c$ and $d$ are constants which vary in order to obtain different correlation level and structure;
- $k_1$ and $k_2$ are adding constants to guarantee $X(\mathbf{s}) > 0$ and $Y(\mathbf{s}) > 0$.

Overall, we present our results for 21 synthetic datasets which differ in the spatial distribution of both the study and auxiliary variables, as well as in their relation. The complete list of settings used to generate the synthetic datasets is presented in Table 1. To give a better idea of the different relations between $X$, $Y$ and $\mathbf{s}$ that can be simulated in our data, Figure 1 shows variables $X(\mathbf{s})$ and $Y(\mathbf{s})$ generated under settings A1, A6 and B6: in scenario A1 we observe a weak correlation between $X$ and $Y$ (equal to $0.298$), with both variables strongly related with space; in both scenarios A6 and B6 the correlation between $X$ and $Y$ is stronger (more than $0.7$), but in B6 part of the co-variability (about $40\%$) is not spatially related.

Table 1: Observed correlation for the simulated populations, by scenarios.

| | | | | | $Corr(X,Y)$ | | |
|---|---|---|---|---|---|---|---|
| Scenario | a | b | c | d | Settings A | Settings B | Settings C |
| 1 | 0.6 | 0.6 | 1 | 1 | 0.298 | 0.310 | 0.298 |
| 2 | 1 | 0.385 | 0.385 | 1 | 0.361 | 0.400 | 0.345 |
| 3 | 0.82 | 0.82 | 1 | 1 | 0.424 | 0.437 | 0.429 |
| 4 | 1 | 1 | 1 | 1 | 0.515 | 0.526 | 0.522 |
| 5 | 1 | 1 | 0.82 | 0.82 | 0.607 | 0.617 | 0.615 |
| 6 | 1 | 1 | 0.6 | 0.6 | 0.738 | 0.744 | 0.747 |

Settings A: $Z_1$, $Z_2$, $Z_3$ with $C = 50$, $C_0 = 0$
Settings B: $Z_1$ with $C = 30$, $C_0 = 20$ and $Z_2$, $Z_3$ with $C = 50$, $C_0 = 0$
Settings C: $Z_1$ with $C = 0$, $C_0 = 50$ and $Z_2$, $Z_3$ with $C = 50$, $C_0 = 0$

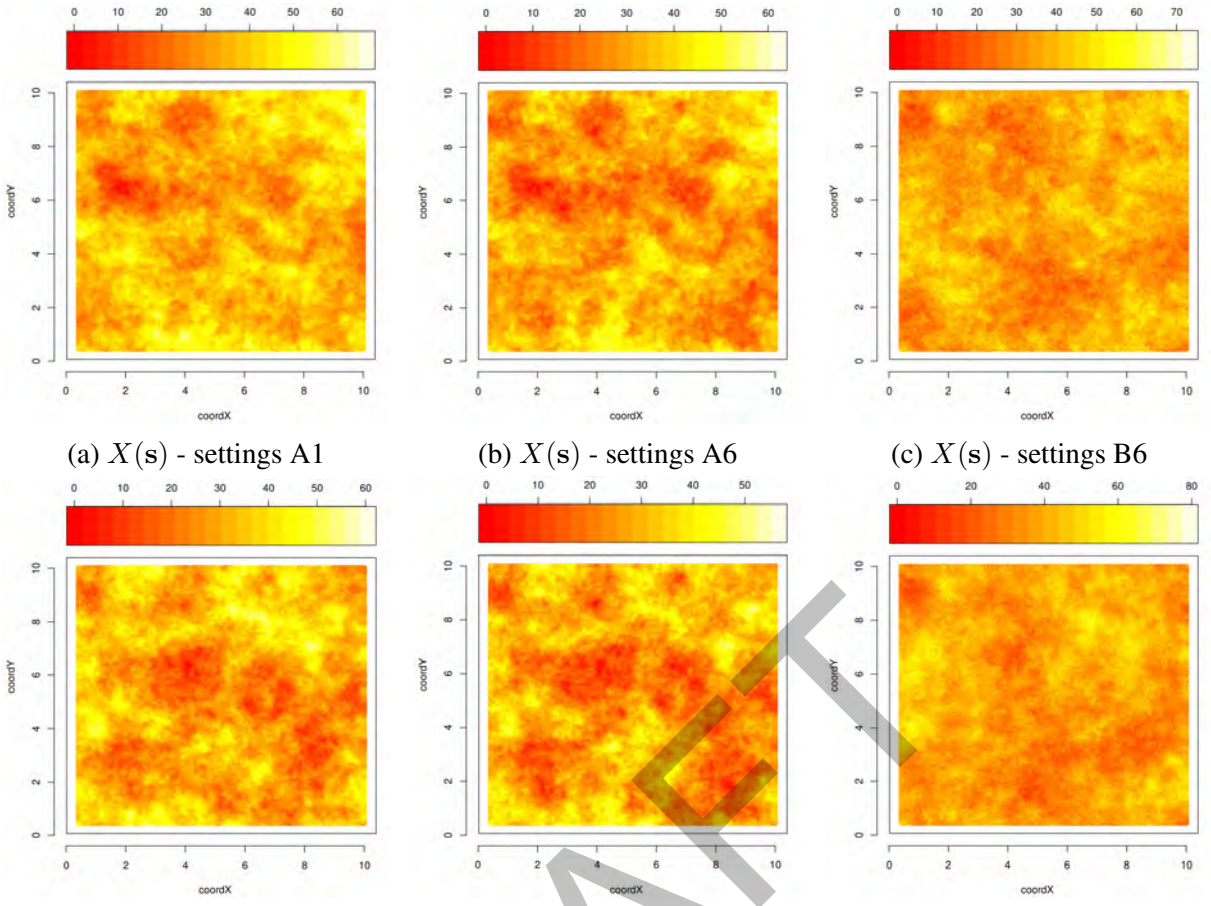
We choose to simulate scenarios with the different settings discussed above because when the analysis concerns a phenomenon measured at global scale it is common to observe different pattern between different areas of the globe and our aim is to find a strategy which could be globally applied by accounting for the various areas characteristics.

For each synthetic dataset we select 1000 replicate samples with $n = 1000$ and $n_0 = 200$. Figure 2 presents the empirical design effect (deff) of the mean estimator for each of the sampling designs, calculated in comparison with the simple random sampling (SRS). Results confirm that, as expected, when we analyse spatial-related phenomena spreading the sample over the area of interest is always convenient: the SpatLPM strategy is always better than the SRS, the UneqSampford and the AuxLPM. Nonetheless, the additional use of the auxiliary information can improve the efficiency of the estimates, in particular if it is used to calculate the inclusion probabilities in the unequal probability designs.

It is important to note that, in order to evaluate when it is more or less convenient to use also the auxiliary variable, it is not enough to consider the correlation between $X$ and $Y$: given the same level of correlation, estimates' efficiency depends on the proportion of co-variability that is related to space. If the co-variability is all defined by a spatial structure (that is, when $Z_1$ has $C = 50$), the SpatLPM design (with equal selection probabilities) is enough; on the other hand when part (or all) of the co-variability in not spatially related (that is, when $Z_1$ has $C = 30$ or $C = 0$), the additional auxiliary variable improves the estimates' efficiency, especially if used to define the unequal inclusion probabilities (UneqLPM).

Unfortunately, in many real situations we don't know in advance the nature and the strength of the relationship between the study and the auxiliary variable. The suggested two-step design aims at addressing this lack of information by using the first-step sample to evaluate it from the data. In particular, we estimate the parameters (nugget, sill and range) of the cross-semivariogram function of $X$ and $Y$, and consequently the nugget-to-sill ratio (NSR), to measure the proportion of co-variability that is not related to space. If this proportion is close to zero the auxiliary variable is useful only for high levels of correlation (as in Scenario A), on the contrary when the NSR increases (as in Scenarios B and C) we observe a gain in efficiency with any correlation above 0.4. Obviously, when the auxiliary variable is relevant, the two-step approach has a cost in term of efficiency in comparison to its correspondent one-step designs (that is the UneqLPM and the BivLPM). But this cost is unavoidable if the relation between

DRAFT




(a) $X(\mathbf{s})$ - settings A1 (b) $X(\mathbf{s})$ - settings A6 (c) $X(\mathbf{s})$ - settings B6

(d) $Y(\mathbf{s})$ - settings A1 (e) $Y(\mathbf{s})$ - settings A6 (f) $Y(\mathbf{s})$ - settings B6

Figure 1: Variables $X(\mathbf{s})$ and $Y(\mathbf{s})$ simulated under settings A1, A6 and B6.

DRAFT

the study and the auxiliary variable is unknown a priori and, more important, is lower that the efficiency loss that we would get if using the "wrong" design. Finally, it's worth noting that the SeqUneqLPM always outperform the SeqBivLPM, confirming that also in the two-step design is preferable to use the auxiliary variable to define the unequal inclusion probabilities.

## References

[1] Benedetti, R., Piersimoni, F., Postiglione, P.: Spatially balanced sampling: A review and a reappraisal. Int. Stat. Rev. **85**,439–454 (2017)

[2] Deville, J.C., Tillé, Y.: Efficient balanced sampling: The cube method. Biometrika **91**, 893–912 (2004)

[3] Diggle, P.J., Ribeiro, P.J.: Model-based Geostatistics. Springer, New York (2007)

[4] Grafström, A., Lundström, N.L.P., Schelin, L.: Spatially balanced sampling through the pivotal method. Biometrics **68**, 514–520 (1986)

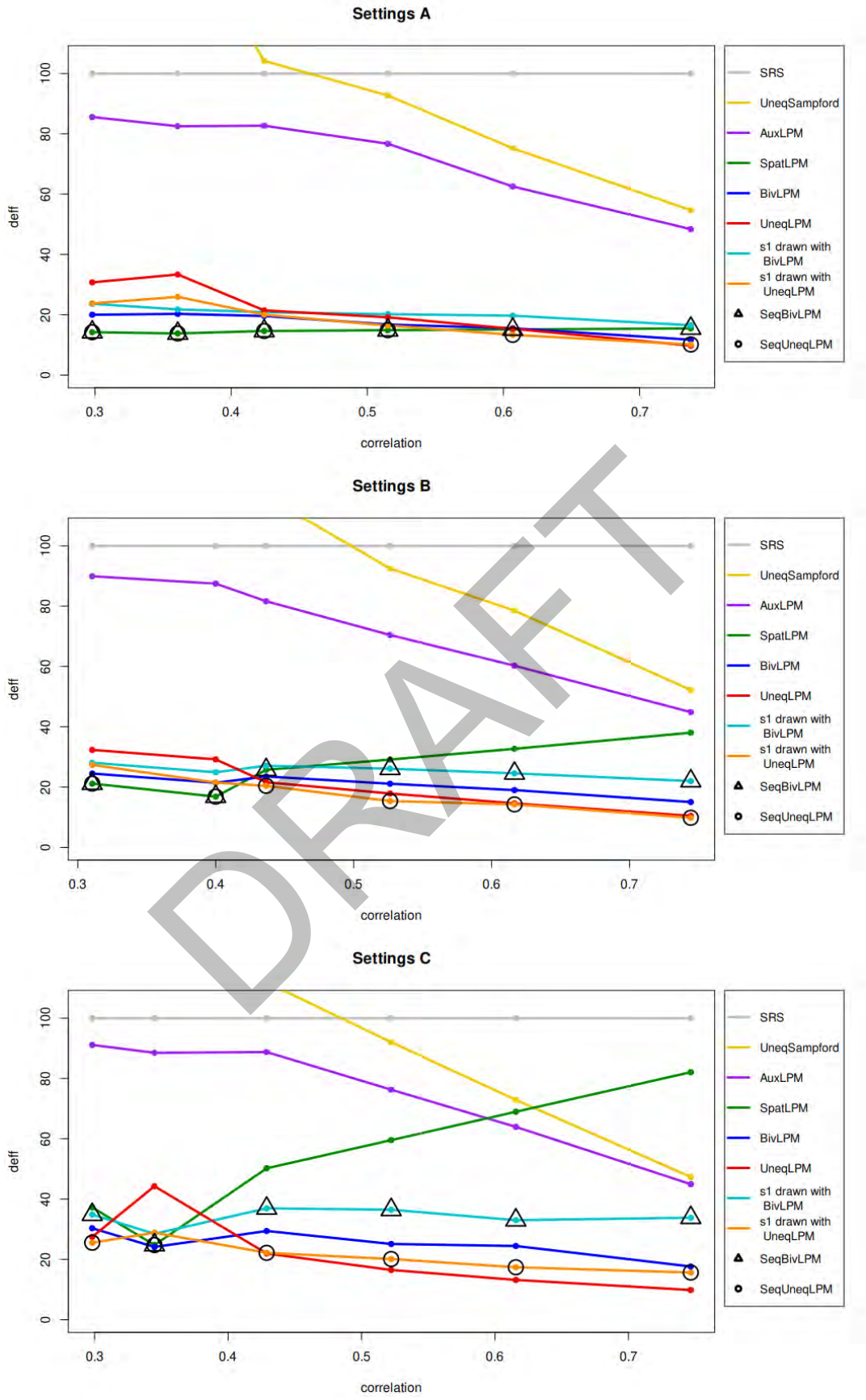
[5] Grafström, A., Tillé, Y.: Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics **24**, 5120–131 (2012)

[6] Lisic, L., Cruze, N.: Local Pivotal Methods for Large Surveys. In: Proceedings ICES V, Geneva Switzerland (2016)

[7] Tillé, Y.: Sampling and Estimation from Finite Populations. Wiley, New York (2020)

[8] Tillé, Y., Wilhelm, M.: Probability sampling designs: Balancing and principles for choice of design. Stat. Sci. **32**, 176–189 (2017)

[9] Wang, J.F., Stein, A., Gao, B.B., Ge, Y.: A review of spatial sampling. Spat. Stat. **2**, 1–14 (2012)




DRAFT

Figure 2: Empirical design effect (deff) by correlation levels. Data simulated under scenarios A, B and C; 1000 replications, size of the initial sample $n_0 = 200$ and full sample size = 1000.

# Robustness and Balance of Sampling or Experimental Designs and Mixture of Designs

Yves Tillé[a] and Ejub Talovic[a]

[a]University of Neuchatel, Bellevaux 51, 2000 Neuchâtel, Switzerland; yves.tille@unine.ch, ejub.talovic@unine.ch

**Abstract**

We present simulations of different major sampling designs to compare their balance and robustness. We also present an existing algorithm to generate mixture of sampling designs in the context of design of experiments and we also present our own algorithm which is simpler and has desirable properties for interpretability and theoretical results.

***Keywords:*** balanced sampling, variance operator, design of experiments, Algorithm

DRAFT

## 1. Introduction

The notion of robustness for a design of experiments has been introduced by (16) and (12). This notion is also valid for the sampling design. When we use balanced designs to decrease the variance due to an auxiliary variable, the variance may increase due to an effect which we call lack of robustness. This robustness can be written as the largest eigenvalue of the variance operator of a sampling or experimental design. If this eigenvalue is large, then it might induce a large variance in the Horvitz-Thompson estimator of the total. A consideration of a trade-off between the efficiency and the robustness of a design has been proposed by (12). They also propose an algorithm called "Gram-Schmidt Walk design" (GS) in which there is a tuning parameter that allows to oscillate between a design that would be totally robust and a balanced design that would integrate the auxiliary information as well as possible. Depending on the risk aversion, one can choose the appropriate value of the tuning parameter.

In this article, we review different sampling and experimental designs. We evaluate the robustness and balance of each of these designs using simulations. We propose a very simple procedure that allows to mix two designs that also uses a tuning parameter. This method offers a much more flexible alternative to the GS design of (12). In addition, this method allows any design to be mixed with any other, which opens the way to multiple combinations of designs. It can be used to handle the balance and robustness trade-off. We can also consider using this method for experimental or sampling design with more than one variable of interest. We next present a set of simulations that show that our mixing method gives similar results to the GS design while being simpler and much easier to interpret.

## 2. Notations and Main designs

Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size $N$. A sample without replacement is a subset, $s \in U$, of the population. A sampling design $p(s)$ assigns to each sample a probability such that $p(s) \geq 0$ and

$$\sum_{s \subset U} p(s) = 1.$$

A random sample $S$ takes as value $s$ with probability $Pr(S = s) = p(s)$.

Let $a_k$ denote the indicator random variable which takes as value 1 if unit $k$ is in $S$ and 0 otherwise. If the sampling is with replacement $a_k$ is the number of times unit $k$ is selected in the sample. The random sample can also be defined by the column vector of the indicator variables $\mathbf{a}^\top = (a_1, \ldots, a_N)$. The inclusion probability is the probability that a unit is selected in the sample, i.e. $\pi_k = \Pr(k \in S) = \mathrm{E}_a(a_k), k \in U$, where $\mathrm{E}_a(.)$ is the expectation under the sampling design. We also define the vector $\boldsymbol{\pi}^\top = \mathrm{E}_a(\mathbf{a})^\top = (\pi_1, \ldots, \pi_N)$. It is assumed that the sum of the inclusion probabilities is an integer number denoted by $n$.

The joint inclusion probabilities are the probabilities that two units are jointly selected in the sample $\pi_{k\ell} = \mathrm{E}_a(a_k a_\ell)$, with $\pi_{kk} = \pi_k$, for all $k, \ell \in U$. The matrix of joint inclusion probabilities is thus $\boldsymbol{\Pi} = \mathrm{E}_a(\mathbf{a}\ \mathbf{a}^\top)$. The covariances between the indicator variables is defined by $\Delta_{k\ell} = \mathrm{cov}_a(a_k a_\ell) = \mathrm{E}_a(a_k a_\ell) - \mathrm{E}_a(a_k)\mathrm{E}_a(a_\ell) = \pi_{k\ell} - \pi_k \pi_\ell$, with $\Delta_{kk} = \pi_k(1 - \pi_k)$, for all $k, \ell \in U$. The variance-covariance matrix is thus $\boldsymbol{\Delta} = \mathrm{var}_a(\mathbf{a}) = \boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}^\top$, where $\mathrm{var}_a(.)$ is the expectation under the sampling design.

*Simple Random Sampling Without Replacement* (SRSWOR) is the design with a fixed sample size $n$ for which all samples of size $n$ have the same probability $n!(N - n)!/N!$ of being selected. For this design,

$$\pi_k = \frac{n}{N}, \Delta_{kk} = \frac{n(N-n)}{N^2}, \pi_{k\ell} = \frac{n(n-1)}{N(N-1)},$$

and

$$\Delta_{k\ell} = -\frac{n(N-n)}{(N-1)N^2}, k \neq \ell \in U.$$

*Stratification* consists of partitioning the population in $H$ strata $U_1, \ldots, U_H$ of sizes $N_1, \ldots, N_H$. Next, in each stratum $h$, a sample of size $n_h$ is selected independently from the other strata with SRSWOR. For this design $\pi_k = n_h/N_h, k \in U_h$, and

$$\Delta_{k\ell} = \begin{cases} \dfrac{n_h(N_h - nh)}{N_h^2} & \text{if } k = \ell \in U_h \\ -\dfrac{n_h(N_h - n_h)}{(N_h - 1)N_h^2} & \text{if } k \neq \ell \in U_h \\ 0 & \text{if } k \in U_h, \ell \in U_i, h \neq i. \end{cases}$$

*Maximum entropy design* also called *Conditional Poisson Sampling* (CPS) is the design with fixed sample size $n$ that maximizes the entropy

$$-\sum_{s \subset U | \# s = n} p(s) \log p(s),$$

subject fixed inclusion probabilities $\pi_k$. The implementation of this design is quite complex and has only recently been resolved (4; 5; 6; 23). It is possible to compute the matrix of inclusion probabilities that satisfy the Yates-Grundy condition, for instance by using the R sampling package (23; 25). The Yates-Grundy condition allow us to guarantee a relatively low bound on the largest eigenvalue of the variance matrix of the sampling design. Therefore, sampling designs with this property are robust. If the inclusion probabilities are equal, the maximum entropy design reduces to SRSWOR.

The *Sampford design* (21) is described in the book of (1) as follows "Select the first unit with probability proportional to measure of size $\pi_k/n$. At each subsequent draw, select with probability of selection proportional to $\pi_k(1 - \pi_k)$ with replacement. If any unit is selected twice, reject the whole sample selected and start again." This design satisfy the Yates-Grundy condition. If the inclusion probabilities are equal, it also reduces to SRSWOR. At each subsequent draw, select with probability of selection proportional to $\pi_k(1 - \pi_k)$ with replacement. If any unit is selected twice, reject the whole sample selected and start again." This design satisfies the Yates-Grundy condition. If the inclusion probabilities are equal, it also reduces to SRSWOR.

*Unequal probabilities systematic sampling* (18; 19) that is defined as follows. First, compute the cumulated inclusion probabilities $V_k = \sum_{\ell=1}^{k} \pi_k$ with $V_0 = 0$ and $V_N = n$. Next generate a uniform

DRAFT

continuous random variable $u$ and select in the sample the units $k$ such that $V_{k-1} \leq u + j < V_k$ for $j = 0, \ldots, n-1$. This sample has a fixed sample size $n$. Systematic sampling has no more than $N$ samples with non-zero probability (20). This design has thus a very small entropy. The joint inclusion probabilities can be computed for instance by using the R sampling package (25). *Equal probability systematic sampling* is the special case where the inclusion probabilities are equal. Moreover, if $N/n$ is an integer, then only $N/n$ samples have a non-null probability of being selected.

*The pivotal method* is one of the special cases of the splitting method proposed by (8). The method was republished by (22). At each step of the method, two units whose inclusion probabilities are not integer are modified randomly:

$$(\widetilde{\pi}_i, \widetilde{\pi}_j) = \begin{cases} (\min(1, \pi_i + \pi_j), \max(\pi_i + \pi_j - 1, 0)) & \text{with probability } q \\ (\max(\pi_i + \pi_j - 1, 0), \min(1, \pi_i + \pi_j)) & \text{with probability } 1 - q, \end{cases}$$

with

$$q = \frac{\min(1, \pi_i + \pi_j) - \pi_j}{2\min(1, \pi_i + \pi_j) - \pi_i - \pi_j}.$$

By choosing at each step a couple of close units in a space, (11) have built a method to obtain well spread samples. The method satisfies the Yates-Grundy condition (17; 3). If the initial probabilities are all equal to 1/2, the design reduces to a *matched pairs design*(Matched). The Matched design is also a stratified design where $N_h = 2$ and $n_h = 1$, for $h = 1, \ldots, H$. In a Matched design, there are $H = N/2$ pairs on units and

$$\Delta_{k\ell} = \begin{cases} \frac{1}{4} & \text{if } k = \ell \in U_h \\ -\frac{1}{4} & \text{if } k \neq \ell \in U_h \\ 0 & \text{if } k \in U_h, \ell \in U_i, h \neq i. \end{cases} \tag{1}$$

Let matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ denote the $N \times p$ matrix of $p$ covariates. *Balanced sampling* are methods for selecting random samples $\mathbf{a}$ that approximatively satisfies the balancing equations:

$$\sum_{k \in U} \frac{a_k \mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k. \tag{2}$$

Fixed sample size is a special case of balanced sampled when matrix $\mathbf{X}$ contains a variable that is equal or proportional to the inclusion probabilities. *The cube method* (Cube) (9) outputs samples that are approximately balanced and also respect the inclusion probabilities of each unit of the population. When all inclusion probabilities are equal to $\pi = n/N$, then the sample mean is approximately equal to the population mean

$$\frac{1}{n} \sum_{k \in U} a_k \mathbf{x}_k \approx \frac{1}{N} \sum_{k \in U} \mathbf{x}_k.$$

*Cube matched pairs design* (Cube-Matched) (24) combines the matched pairs design with balanced sampling by the Cube method. Pairs of similar units are constructed and only one unit is selected in each pair. Moreover, the design is balanced on covariates. This method is a simple application of the stratified-balanced sampling where each stratum contains only two units whose one is selected in the sample (2; 13; 15).

*Simple random sampling with replacement* consists of selecting $n$ times from the population a unit with equal probabilities $1/N$ and with replacement. The variable $a_k$ contains the number of times the unit $k$ is selected in the sample and has a binomial distribution with parameter $1/N$ and exponent $n$. Thus, $\mathrm{E}_a(a_k) = n/N$ and $\Sigma_{kk} = \mathrm{var}_a(a_k) = n(N-1)/N^2 = n/N - n/N^2$. Vector $\mathbf{a}$ has a multinomial distribution and $\Sigma_{k\ell} = \mathrm{cov}_a(a_k, a_\ell) = -n/N^2, k \neq \ell$.

*Multinomial sampling* consists of selecting $n$ times from the population a unit with equal probabilities $p_k = \pi_k/n$ and with replacement. The variable $a_k$ contains the number of times the unit $k$ is selected in the sample and has a binomial distribution with parameter $p_k = \pi_k/n$ and exponent $n$. Thus,

DRAFT

$\mathrm{E}_a(a_k) = np_k = \pi_k$ and $\Sigma_{kk} = \mathrm{var}_a(a_k) = np_k(1-p_k) = \pi_k - \pi_k^2/n$. Vector $\mathbf{a}$ has a multinomial distribution and $\mathrm{var}_a(\mathbf{a}) = \boldsymbol{\Sigma}$, where $\Sigma_{k\ell} = \mathrm{cov}_a(a_k, a_\ell) = -np_kp_\ell = -\pi_k\pi_\ell/n, k \neq \ell$. When $p_k = 1/N$, the multinomial sampling reduces to the SRSWR. The properties of the main sampling designs are described in Table 1. The proofs of these properties can be found in (23).

Table 1: Summary of the properties of some sampling methods

| | fixed size | without replacement | equal probabilities | Yates-Grundy Condition | Maximum Entropy |
|---|---|---|---|---|---|
| Poisson | | ✓ | | ✓ | ✓ |
| Bernoulli | | ✓ | ✓ | ✓ | ✓ |
| SRSWOR | ✓ | ✓ | ✓ | ✓ | ✓ |
| Stratification | ✓ | ✓ | ✓ | ✓ | ✓ |
| CPS | ✓ | ✓ | | ✓ | ✓ |
| Sampford | ✓ | ✓ | | ✓ | |
| Unequal Systematic | ✓ | ✓ | | | |
| Equal Systematic | ✓ | ✓ | ✓ | | |
| Pivotal | ✓ | ✓ | | ✓ | |
| Cube | ✓ | ✓ | | | |
| Multinomial | ✓ | | | ✓ | ✓ |
| SRSWR | ✓ | | ✓ | ✓ | ✓ |

DRAFT

## 3. Mixing designs

The GS design (12) handles the trade-off between balance and robustness. The basic step of the GS design is very similar to the Cube method of (9). In the Cube method, at each step, the vector of inclusion probabilities is randomly changed in a direction that respects the balancing equations. However, (12) uses the GS algorithm to balance an augmented covariate vector, which is a scaled concatenation of the covariates of the units and a unit-unique indicator variable. A tuning parameter $\phi$ allows to give more importance either to the robustness or to the balancing. It allows one to make a mixture of designs: a robust design like the Bernoulli design and a balanced design. When $\phi \in \{0, 1\}$, the GS algorithm will correspond either to a Bernoulli design or balanced design like the cube method.

We propose another method which allows to mix two designs that is presented in Algorithm 1. This method is much simpler than the GS design. It requires only three lines of code to be implemented. Moreover, it allows to mix any design with any other design. The inclusion probabilities of both sampling are supposed to be the same.

**Algorithm 1** Simple method for mixing designs

- Select a sample $\mathbf{a}_A$ with a design $p_A(.)$, inclusion probabilities $\boldsymbol{\pi}$ and a variance-covariance matrix $\boldsymbol{\Delta}_A$.
- Next, we compute $\boldsymbol{\pi}_B = \sqrt{\phi}\,\mathbf{a}_A + (1-\sqrt{\phi})\boldsymbol{\pi}$.
- We then select a sample $\mathbf{a}$ with a design $p_B(.)$, inclusion probabilities $\boldsymbol{\pi}_B$ and a variance-covariance matrix $\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)$.

**Result 1.** *With Algorithm 1:*

*(i)* $\mathrm{E}_a(\mathbf{a}) = \boldsymbol{\pi}$,
*(ii)* $\boldsymbol{\Delta} = \mathrm{var}_a(\mathbf{a}) = \mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\} + \phi\boldsymbol{\Delta}_A$,
*(iii)* $\frac{1}{1-\phi}\mathrm{diag}\{\mathrm{E}_a(\boldsymbol{\Delta}_B)\} = \mathrm{diag}(\boldsymbol{\Delta}) = \mathrm{diag}(\boldsymbol{\Delta}_A)$.

*Proof.*

*(i)* The final inclusion probabilities are

$$\mathrm{E}_a(\mathbf{a}) = \mathrm{E}_a\ \mathrm{E}_a(\mathbf{a}|\mathbf{a}_A) = \mathrm{E}_a\{\sqrt{\phi}\ \mathbf{a}_A + (1-\sqrt{\phi})\boldsymbol{\pi}\} = \boldsymbol{\pi}.$$

*(ii)* The final variance matrix is:

$$\begin{aligned}\boldsymbol{\Delta} &= \mathrm{var}_a(\mathbf{a}) = \mathrm{E}_a\ \mathrm{var}_a(\mathbf{a}|\mathbf{a}_A) + \mathrm{var}_a\ \mathrm{E}_a(\mathbf{a}|\mathbf{a}_A)\\ &= \mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\} + \mathrm{var}_a\{\sqrt{\phi}\ \mathbf{a}_A + (1-\sqrt{\phi})\boldsymbol{\pi}\}\\ &= \mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\} + \phi\boldsymbol{\Delta}_A. \qquad (3)\end{aligned}$$

*(iii)* Vectors $\mathbf{a}$ and $\mathbf{a}_A$ have the same expectation. Thus $\mathrm{diag}(\boldsymbol{\Delta}) = \mathrm{diag}(\boldsymbol{\Delta}_A)$. From Equality (3), we have

$$\mathrm{diag}(\boldsymbol{\Delta}) = \mathrm{diag}\{\mathrm{E}_a(\boldsymbol{\Delta}_B)\} + \phi\ \mathrm{diag}(\boldsymbol{\Delta}_A),$$

and thus

$$\frac{1}{1-\phi}\mathrm{diag}\{\mathrm{E}_a(\boldsymbol{\Delta}_B)\} = \mathrm{diag}(\boldsymbol{\Delta}).$$

□

Ideally, we would have that $\mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\} \approx (1-\phi)\boldsymbol{\Delta}_B(\boldsymbol{\pi})$ for better interpretability of $\boldsymbol{\Delta}$ and theoretical results. It would mean that $\boldsymbol{\Delta} \approx \phi\boldsymbol{\Delta}_A + (1-\phi)\boldsymbol{\Delta}_B$. However, it is very difficult to calculate analytically the term $\mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\}$ and probably impossible with a mixture that uses a balanced sampling design. We carried out simulations of a mixture of design where $\boldsymbol{\Delta}_A$ and $\boldsymbol{\Delta}_B$ correspond respectively to a SRSWOR and balanced sampling design-based on the cube method. For $\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)$, the cube method has been parametrized in a way to ensure that the inclusion probabilities are $\boldsymbol{\pi}_B$ while also having the propriety that $\mathbf{X}^\top\mathbf{D}^{-1}\mathbf{a} \approx \mathbf{X}^\top\mathbf{D}^{-1}\boldsymbol{\pi}_B$. The auxiliary variables are generated from a normal distribution, $\pi = 1/2$ and $N = 250$. The simulations show that $\mathrm{E}_a\{\boldsymbol{\Delta}_B(\boldsymbol{\pi}_B)\}$ is very close to $(1-\phi)\boldsymbol{\Delta}_B(\boldsymbol{\pi})$.

Moreover if $\lambda_1, \lambda_{A1}, \lambda_{B1}$ are respectively the largest eigenvalues of $\boldsymbol{\Delta}$, $\boldsymbol{\Delta}_A$ and $\mathrm{E}_a(\boldsymbol{\Delta}_B)/(1-\phi)$, then

$$\lambda_1 \leq \phi\lambda_{A1} + (1-\phi)\lambda_{B1}.$$

DRAFT

## 4. Simulations

In this section, we compare the different sampling or experimental designs. In order to evaluate the robustness of the designs and the quality of the approximations, we performed simulations of the different designs on a clinical trial data set. It contains information about 40 patients split into two groups before and after their corneal astigmatism surgery (14). We are only interested in the variables measured before the operations. We retain five covariates: sex, eye, age, axis and topographic astigmatism. Two of those covariates are categorical. The number of patients is $N = 40$ and the inclusion probabilities are all equal to $\pi = 1/2$ and therefore $n = 20$. For the Matched and Cube-Matched designs, the pairs have been constructed by discrete linear programming by minimizing the sum on the pairs of the squares of the Mahalanobis distances between the paired units. The estimation of the largest eigenvalue by simulation is unstable, especially when the largest eigenvalues in a design are the same or similar. There are slight differences between the maximal eigenvalue estimated through simulation and the exact theoretical results can be noticed. Therefore, we performed a very large number of simulations (SIM=10,000,000).

We also ran the GS design of (12) using the Julia implementation (7). We used a version of the algorithm that outputs experimental designs with fixed size for each group. Unfortunately, we found that this algorithm does not provide comparable results depending on whether the variables have the same means or variances. This is due to the way the covariates are normalized in the GS algorithm. Therefore,

we centred and reduced all variables and made them orthogonal using a singular value decomposition. Without this change of variables, the results obtained by the GS method are inconsistent. This algorithm also creates a mixture of design which depends on a tuning parameter.

Figure 1 contains the scatterplot of the squares of the Mahalanobis distances by the largest eigenvalues estimated through simulations of the different designs. On the left side, one sees that the most robust designs are Bernoulli and SRSWOR but they are not balanced. The Cube-Matched method is on the right the least robust but most balanced method. Bernoulli design is not shown on figure because it is not balanced enough. Systematic sampling design is not shown on the figure because it is neither robust nor balanced. Some sampling designs, that we introduced, are not shown in the figure as some of them are reduced to SRSROW or the Matched design when the inclusion probabilities equal are all equal to $0.5$.

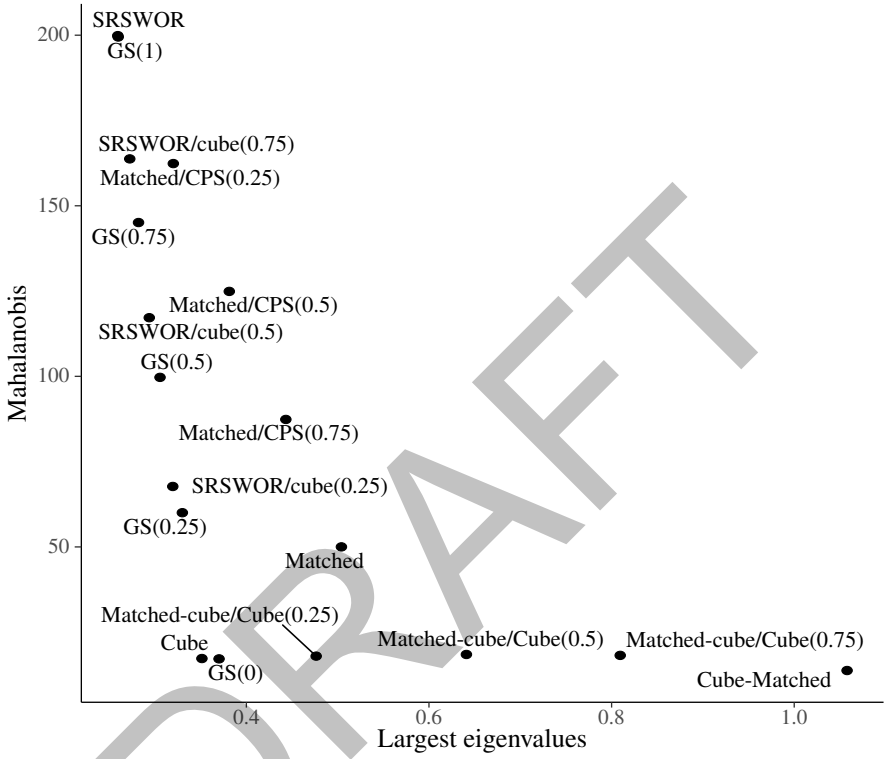



Figure 1: Scatterplot of the squares of the Mahalanobis distances by the largest eigenvalues.

DRAFT

The Cube method is a very good compromise between the two extremes as it is well balanced and relatively robust. The Matched method is not to be recommended as it is less balanced and less robust than the Cube method. The Matched method can also have difficulties finding pairs which are similar enough. However, one advantage of matching is that the samples or group generated from this method are relatively well-spread while samples generated from a balanced design like the Cube method could output samples which are not well-spread. Generating a mixture of design between the Cube method and the Matched method could give a good compromise in certain cases. See (10) for the advantages of well-spread samples. We also notice that the designs resulting from mixtures with our method lie between the two designs from which they originate in the graph in function of their weights $\phi$. There are interesting designs by mixing SRSWOR and Cube or by mixing Cube and Cube-Matched. The GS design with $\phi = 1$ corresponds to SRSWOR, with $\phi = 0$ it corresponds to the cube method. When $\phi = 0, 25, 0.5$ and $0.75$, they are close to the mixtures between SRSWOR and the Cube method we proposed.

## References


[1] BREWER, K. R. W., HANIF, M.: *Sampling with Unequal Probabilities*. Springer, New York, (1983).

[2] CHAUVET, G.: Stratified balanced sampling. *Survey Methodology 35* (2009), 115–119.

[3] CHAUVET, G.: On a characterization of ordered pivotal sampling. *Bernoulli 18*, 4 (2012), 1320–1340.

[4] CHEN, X.-H.: Poisson-binomial distribution, conditional Bernoulli distribution and maximum entropy. Tech. rep., Department of Statistics, Harvard University, (1993).

[5] CHEN, X.-H., DEMPSTER, A. P., LIU, J. S.: Weighted finite population sampling to maximize entropy. *Biometrika 81* (1994), 457–469.

[6] CHEN, X.-H., LIU, J. S.: Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica 7* (1997), 875–892.

[7] CHRIS, H., DANIEL, S.: Github, (2020) Available at https://github.com/crharshaw/GSWDesign.jl

[8] DEVILLE, J.-C., TILLÉ, Y.: Unequal probability sampling without replacement through a splitting method. *Biometrika 85* (1998), 89–101.

[9] DEVILLE, J.-C., TILLÉ, Y.: Efficient balanced sampling: The cube method. *Biometrika 91* (2004), 893–912.

[10] GRAFSTRÖM, A., LUNDSTRÖM, N. L. P.: Why well spread probability samples are balanced? *Open Journal of Statistics 3*, 1 (2013), 36–41.

[11] GRAFSTRÖM, A., LUNDSTRÖM, N. L. P., SCHELIN, L.: Spatially balanced sampling through the pivotal method. *Biometrics 68*, 2 (2012), 514–520.

[12] HARSHAW, C., SLAVJE, F., SPIELMAN, D. A., ZHANG, P.: Balancing covariates in randomized experiments with the gram–schmidt walk design. *Universtiy of Yale, unpublished* (2022).

[13] HASLER, C., TILLÉ, Y.: Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis 74* (2014), 81–94.

[14] HAYDER, F.: Corneal astigmatism. Zenodo.

[15] JAUSLIN, R., EUSTACHE, E., TILLÉ, Y.: Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science 4* (2021), 783–795.

[16] KAPELNER, A., KRIEGER, A. M., SKLAR, M., AZRIEL, D.: Optimal rerandomization via a criterion that provides insurance against failed experiments, (2019) Available at https://arxiv.org/abs/1905.03337

[17] KRAMER, J. B., CUTLER, J., RADCLIFFE, A.: Negative dependence and Srinivasan's sampling process. *Combinatorics, Probability and Computing 20*, 3 (2011), 347–361.

[18] MADOW, L. H., MADOW, W. G.: On the theory of systematic sampling. *Annals of Mathematical Statistics 15* (1944), 1–24.

[19] MADOW, W. G. On the theory of systematic sampling, II. *Annals of Mathematical Statistics 20* (1949), 333–354.

[20] PEA, J., QUALITÉ, L., TILLÉ, Y.: Systematic sampling is a minimal support design. *Computational Statistics & Data Analysis 51* (2007), 5591–5602.

[21] SAMPFORD, M. R.: On sampling without replacement with unequal probabilities of selection. *Biometrika 54* (1967), 499–513.

[22] SRINIVASAN, A.: Distributions on level-sets with applications to approximation algorithms. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science* (2001), IEEE, pp. 588–597.

[23] TILLÉ, Y.: *Sampling Algorithms*. Springer, New York, (2006).

[24] TILLÉ, Y.: Some solutions inspired by survey sampling theory to build effective clinical trials. *International Statistical Review 90*, 3 (2022), 481–498.

[25] TILLÉ, Y., MATEI, A.: *sampling: Survey Sampling*, (2021). R package version 2.9.


DRAFT

# Robustness Bounds for Sampling and Experimental Designs

Ejub Talovic[a] and Yves Tillé [a]

[a]University of Neuchatel, Bellevaux 51, 2000 Neuchâtel, Switzerland; ejub.talovic@unine.ch, yves.tille@unine.ch

**Abstract**

For both experimental and sampling designs, the efficiency of designs has been extensively studied. There are many ways to incorporate auxiliary information into designs. However, when we want to decrease the variance due to an auxiliary variable by the use of balanced designs, we risk seeing it increase due to another variable. For the most common designs, we study the robustness in the sense of the largest increase in variance compared to simple random sampling or a Bernoulli design. This robustness can be written as the largest eigenvalue of the variance operator. We determine lower and upper bounds and approximations of this eigenvalue for different designs.

***Keywords:*** balanced sampling, design of experiments, variance operator

DRAFT

## 1. Introduction

Both sampling designs and experimental designs have a common feature: in both cases, random variables are generated in order to select units either to constitute a sample or to create a test group and a control group. For survey sampling, the recommendations relate primarily to issues of design efficiency under certain hypotheses relative to the variables of interest to be estimated. We can thus resort to designs of fixed size, stratified, balanced with equal or unequal probabilities (see amongst other 3; 14; 15). For experimental designs, randomization between the test group and the control group is also recommended. There are also stratification techniques, blocking, matching, rerandomization, balanced sampling to make the two groups as equivalent as possible (see amongst other 13; 11; 12; 17; 10; 8; 16). Unfortunately, everything has a price. It can be shown that if a design is more efficient for one variable, it is necessarily less efficient for another variable. Thus, one cannot rely solely on a single efficiency criterion to determine the appropriate design.

The notion of robustness for a design of experiments has been introduced by (9) and (6). This notion is also valid for the sampling design. When we use balanced or efficient designs to decrease the variance due to an auxiliary variable, the variance may increase due to an effect which we call lack of robustness. This robustness can be written as the largest eigenvalue of the variance operator of a sampling or experimental design. If this eigenvalue is large, then it might induce a large variance in the Horvitz-Thompson estimator of the total. We can thus evaluate the maximum price to pay for the application of a design that integrates auxiliary information.

We review the different sampling and experimental designs. We evaluate the robustness of each of these designs. To do so, we calculate upper and lower bounds for the largest eigenvalues of the variance matrices. We also give approximations for these eigenvalues. These calculations show the determining role of the Yates-Grundy conditions which allow us to calculate a relatively low upper bound. We also study the potential effect of the choice of the design on the variance of the Horvitz-Thompson estimator.

## 2. Sampling designs and designs of experiments

Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size $N$. A sample without replacement is a subset, $s \in U$, of the population. A sampling design $p(s)$ assigns to each sample a probability such that $p(s) \geq 0$ and

$$\sum_{s \subset U} p(s) = 1.$$

A random sample $S$ takes as value $s$ with probability $Pr(S = s) = p(s)$.

Let $a_k$ denote the indicator random variable which takes as value 1 if unit $k$ is in $S$ and 0 otherwise. If the sampling is with replacement $a_k$ is the number of times unit $k$ is selected in the sample. The random sample can also be defined by the column vector of the indicator variables $\mathbf{a}^\top = (a_1, \ldots, a_N)$. The first-order inclusion probability is the probability that a unit is selected in the sample, i.e. $\pi_k = \Pr(k \in S) = \mathrm{E}_a(a_k), k \in U$, where $\mathrm{E}_a(.)$ is the expectation under the sampling design. In this article, when the order is not defined, it means we refer to the first-order inclusion probabilities. We also define the vector $\boldsymbol{\pi}^\top = \mathrm{E}_a(\mathbf{a})^\top = (\pi_1, \ldots, \pi_N)$. It is assumed that the sum of the inclusion probabilities is an integer number denoted by $n$.

The joint inclusion probabilities are the probabilities that two units are jointly selected in the sample $\pi_{k\ell} = \mathrm{E}_a(a_k a_\ell)$, with $\pi_{kk} = \pi_k$, for all $k, \ell \in U$. The matrix of joint inclusion probabilities is thus $\boldsymbol{\Pi} = \mathrm{E}_a(\mathbf{a}\ \mathbf{a}^\top)$. The covariances between the indicator variables is defined by $\Delta_{k\ell} = \mathrm{cov}_a(a_k a_\ell) = \mathrm{E}_a(a_k a_\ell) - \mathrm{E}_a(a_k)\mathrm{E}_a(a_\ell) = \pi_{k\ell} - \pi_k \pi_\ell$, with $\Delta_{kk} = \pi_k(1 - \pi_k)$, for all $k, \ell \in U$. The variance-covariance matrix is thus $\boldsymbol{\Delta} = \mathrm{var}_a(\mathbf{a}) = \boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}^\top$, where $\mathrm{var}_a(.)$ is the expectation under the sampling design.

The aim is to estimate the total of a variable of interest $y_k$ given by

$$t_y = \sum_{k \in U} y_k.$$

If $\pi_k > 0, k \in U$, then the Horvitz-Thompson estimator (7) given by

$$\hat{t}_y = \sum_{k \in S} \frac{y_k}{\pi_k}$$

is an unbiased estimator of $t_y$. The variance of this estimator is

$$\mathrm{var}_a(\hat{t}_y) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell} = \mathbf{y}^\top \mathbf{D}^{-1} \boldsymbol{\Delta} \mathbf{D}^{-1} \mathbf{y},$$

where $\mathbf{D} = \mathrm{diag}(\boldsymbol{\pi})$. If the sampling design has a fixed sample size, (18) showed that the variance can also be written:

$$\mathrm{var}_a(\hat{t}_y) = -\frac{1}{2} \sum_{k \in U} \sum_{\ell \in U} \left( \frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}.$$

If $\pi_{k\ell} > 0$, for $k, \ell \in U$, two estimators of variance can be constructed. The Horvitz-Thompson estimator of the variance is given by

$$\widehat{\mathrm{var}}_a(\hat{t}_y) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

The Yates-Grundy estimator is only available for sampling designs with fixed sample size and is given by

$$\widehat{\mathrm{var}}_a(\hat{t}_y) = -\frac{1}{2} \sum_{k \in S} \sum_{\ell \in S} \left( \frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{1}$$

**Definition 1.** *A design satisfies the Yates-Grundy condition if* $\Delta_{k\ell} \leq 0$ *for all* $k \neq \ell \in U$.

DRAFT

Somes authors suggest using designs that satisfy this condition, since the Yates-Grundy estimator of the variance cannot then be negative (18). This condition is also called negative correlation. The Yates-Grundy condition is an important hypothesis to construct central limit theorems in finite population (2; 1; 5). In this paper, we will show that the Yates-Grundy condition also contribute to robustify a sampling or experimental design.

In randomized experiments, the problem is somewhat different. Suppose we have a population $U$ of size $N$ and that we want to test a treatment $T$ on the units $k \in U$. If unit $k$ receives the treatment, its response variable becomes $y_k^T$. If unit $k$ does not receive the treatment, it belongs to the control group and is response is denoted by $y_k^C$. This is a randomized experiment with two groups. The goal is to estimate

$$\tau = \frac{1}{N}\sum_{k\in U} y_k^T - \frac{1}{N}\sum_{k\in U} y_k^C.$$

Here we adopt a purely design-based approach as in (6). The values of $y_k^T$ and $y_k^C$ cannot be observed at the same time for unit $k$, because no unit can belong to both the control and treatment groups. We select a sample with inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ and apply the treatment to the chosen units.

Then, $\tau$ can be estimated using the difference of the two the Horvitz-Thompson estimators:

$$\begin{aligned}\widehat{\tau} &= \frac{1}{N}\sum_{k\in U}\frac{y_k^T a_k}{\pi_k} - \frac{1}{N}\sum_{k\in U}\frac{y_k^C(1-a_k)}{1-\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\frac{y_k^T(1-\pi_k)+y_k^C\pi_k}{(1-\pi_k)\pi_k}a_k - \frac{1}{N}\sum_{k\in U}\frac{y_k^C}{(1-\pi_k)}.\end{aligned}$$

If we define $\mathbf{z} = (z_1, \ldots, z_N)^\top$, where

$$z_k = \frac{y_k^T(1-\pi_k)+y_k^C\pi_k}{(1-\pi_k)\pi_k},$$

the variance becomes

$$\mathrm{var}_a(\widehat{\tau}) = \frac{\mathbf{z}^\top \mathrm{var}_a(\mathbf{a})\mathbf{z}}{N^2} = \frac{\mathbf{z}^\top \boldsymbol{\Delta}\mathbf{z}}{N^2}. \tag{2}$$

In both experimental and survey designs, the variance operator $\boldsymbol{\Delta}$ plays a fundamental role in the precision of the estimates.

The calculation of the variance given in (2) may seem similar to the calculation of the variance in a sample design. However, an important difference is that $z_k$ is not known at the sample level. Estimating the variance is therefore a more difficult problem than for a sample design. In sampling designs, as in experimental designs, the $\boldsymbol{\Delta}$ matrix plays a determining role. If the largest eigenvalue of $\boldsymbol{\Delta}$ is large, then there is a risk that the variance of the Horvitz-Thompson estimator becomes large.

## 3. Upper bounds of the variance

The $\boldsymbol{\Delta}$ matrix is the heart of the precision problem for both the sampling and experimental designs. The analysis of this matrix is therefore the crucial question for the determination of the design that best fits the objectives of this randomization. The natural way to analyse the $\boldsymbol{\Delta}$ operator is to perform a diagonalization. We can look for the vector $\mathbf{u}$ which maximizes

$$\frac{\mathbf{u}^\top \boldsymbol{\Delta}\mathbf{u}}{\mathbf{u}^\top\mathbf{u}}.$$

The solution is that $\mathbf{u}$ is the eigenvector of $\boldsymbol{\Delta}$ associated to the largest eigenvalue $\lambda_1$. This eigenvalue is somehow associated with the greatest risk of the design, as it is a factor of the variance for the variable that will be most poorly estimated.

DRAFT

A first result shows that there is no design that is uniformly better than the others for given inclusion probabilities. Consider two designs $p_1(s)$ and $p_2(s)$ with the same inclusion probabilities $\pi_k$ and whose variance operators are respectively $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$. A design $p_1(s)$ would be uniformly better than a design $p_2(s)$ if $\mathbf{u}^\top \mathbf{\Delta}_1 \mathbf{u} \leq \mathbf{u}^\top \mathbf{\Delta}_2 \mathbf{u}$ for all $\mathbf{u} \in \mathbb{R}^N$ and that there is at least one vector $\mathbf{y}$ such that $\mathbf{y}^\top \mathbf{\Delta}_1 \mathbf{y} < \mathbf{y}^\top \mathbf{\Delta}_2 \mathbf{y}$. We refer to the following result:

**Result 1.** *For any design $p_1(s)$, there is no uniformly better design than $p_2(s)$ with the same first-order inclusion probability.*

*Proof.* (by contradiction) Suppose there is a design $p_1(s)$ uniformly better than a design $p_2(s)$ with the same first-order inclusion probabilities. In this case, $\mathbf{u}^\top(\mathbf{\Delta}_2 - \mathbf{\Delta}_1)\mathbf{u} \geq 0$, for all $\mathbf{u} \in \mathbb{R}^N$, which implies that the array $(\mathbf{\Delta}_2 - \mathbf{\Delta}_1)$ is semi-definite positive. Since the trace of $(\mathbf{\Delta}_2 - \mathbf{\Delta}_1)$ is zero, all eigenvalues of $(\mathbf{\Delta}_2 - \mathbf{\Delta}_1)$ are zero, which is in contradiction with the fact that there is at least one vector $\mathbf{y}$ such as $\mathbf{y}^\top \mathbf{\Delta}_1 \mathbf{y} < \mathbf{y}^\top \mathbf{\Delta}_2 \mathbf{y}$. □

We can see the problem like this. Since the inclusion probabilities determine the diagonal of $\mathbf{\Delta}$, two designs that have the same inclusion probabilities have the same trace for $\mathbf{\Delta}$ and therefore the same sum of eigenvalues. Each eigenvalue gives a variance factor in the direction of an eigenvector. So, if the variance is decreased in one direction, it is necessarily increased in another direction.

This result shows that there is no design which is better than all the others. Therefore, we cannot use a precision criterion linked to the variance to find the optimal design. If additional auxiliary information is not available, it is reasonable to try to equalize the eigenvalues and thus to take the most random design possible, that is to say, the design that maximizes the entropy. Additionally, we often measure several variables of interest so it is preferable that the dispersion is more or less the same in all directions.

Let us show a few basic results. Let us start with sampling designs with fixed sample size and with equal inclusion probabilities.

**Result 2.** *Let $p(s)$ be a sampling design of fixed sample size $n$ from a population $U$ of size $N$ and with equal inclusion probabilities $\pi_k = \pi, k \in U$, then $\mathbf{\Pi}$ and $\mathbf{\Delta}$ have the same eigenvectors.*

*Proof.* Since $\mathbf{1}$ is the eigenvector of $\mathbf{\Delta}$ associated with the null eigenvalues, all the other eigenvectors are orthogonal to $\mathbf{1}$ and are thus centred. On one hand, we have $\mathbf{\Delta}\mathbf{1} = \mathbf{\Pi}\mathbf{1} - \boldsymbol{\pi}\boldsymbol{\pi}^\top \mathbf{1} = \mathbf{\Pi}\mathbf{1} - \boldsymbol{\pi} n = \mathbf{0}$. Thus, $\mathbf{\Pi}\mathbf{1} = \boldsymbol{\pi} n$ and $\mathbf{1}$ is an eigenvector of $\mathbf{\Pi}$ associated with the eigenvalue $n\pi$. On the other hand, if $\mathbf{u}$ is a centred eigenvector of $\mathbf{\Delta}$, $\mathbf{\Delta}\mathbf{u} = \lambda\mathbf{u}$, thus $\mathbf{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}^\top \mathbf{u} = \lambda\mathbf{u}$. Since $\boldsymbol{\pi}^\top \mathbf{u} = 0$, vector $\mathbf{u}$ is also an eigenvector of $\mathbf{\Pi}$ with the same eigenvalue as for $\mathbf{\Delta}$. □

The maximal eigenvalue of fixed sample size and with equal probabilities sampling designs is

$$\lambda_1 \leq \min\left(\frac{n^2}{N}, \frac{(N-n)^2}{N}\right).$$

This bound can be reached with systematic sampling. When the inclusion probabilities are unequal, the problem becomes a little more difficult.

**Result 3.** *For any sampling design on a finite population $U$ of size $N$, the largest eigenvalue of $\mathbf{\Delta}$ is smaller than or equal to $\max_{k \in U}(\sum_{\ell \in U} \pi_{k\ell})$. If the sampling design has a fixed sample size $n$, then the bound can be express as $n \max_{k \in U}(\pi_k)$.*

*Proof.* Let us denote the eigenvalues of $\mathbf{\Pi}$ by $\lambda_1, \ldots, \lambda_n$. Using the Perron-Frobenius inequality, we know that

$$\lambda_1 \leq \max_{k \in U}\left(\sum_{\ell \in U} \pi_{k\ell}\right).$$

The matrix $\boldsymbol{\pi}\boldsymbol{\pi}^\top$ is a symmetric matrix of rank 1, whose only non-zero eigenvalue $\lambda_1'$ is equal to $\sum_{k \in U} \pi_k^2$ with the eigenvector $\boldsymbol{\pi}$. The largest eigenvalue of $-\boldsymbol{\pi}\boldsymbol{\pi}^\top$ is thus 0. Each eigenvalue of $\mathbf{\Delta}$

DRAFT

is non-negative, because it is a covariance matrix. Both $\boldsymbol{\Pi}$ and $\boldsymbol{\pi}\boldsymbol{\pi}^\top$ are Hermitian matrices. Therefore, from Weyl's inequality, we deduce that the largest eigenvalue of $\boldsymbol{\Delta}$ is smaller than or equal to $\max_{k\in U}(\sum_{\ell\in U}\pi_{k\ell})$, i.e.,

$$\lambda_1 \leq \max_{k\in U}\left(\sum_{\ell\in U}\pi_{k\ell}\right).$$

If the sampling design has a fixed sample size $n$, then

$$\max_{k\in U}\left(\sum_{\ell\in U}\pi_{k\ell}\right) = \max_{k\in U}\left\{\mathrm{E}_a\left(a_k\sum_{\ell\in U}a_\ell\right)\right\} = n\max_{k\in U}(\pi_k).$$

□

**Result 4.** *For any sampling design on a finite population $U$ of size $N$, the largest eigenvalue of $\boldsymbol{\Delta}$ is smaller than or equal to* $\max_{k\in U}\left(\sum_{\ell\in U}|\Delta_{k\ell}|\right) = \max_{k\in U}\left(\sum_{\ell\in U}|\pi_{k\ell}-\pi_k\pi_\ell|\right)$.

*Proof.* The proof is a simple corollary of the Gershgorin circle theorem. However, the result can also be derived from Theorem 4.1 of (4). □

The bound of Result 4 will often be smaller than the one from Result 3 but it is not always the case. The sampling design that selects one unit out of a population of 3 units with equal probability is a simple example where $n\max_{k\in U}(\pi_k) = 1/3 < \max_{k\in U}\left(\sum_{\ell\in U}|\Delta_{k\ell}|\right) = 4/9$.

If the Yates-Grundy conditions are satisfied, a simpler bound than the previous ones can be deduced.

**Corollary 1.** *For any sampling design that satisfies the Yates-Grundy conditions and a fixed sample size $n$, the largest eigenvalue of $\boldsymbol{\Delta}$ is smaller than or equal to* $2\max_{k\in U}\{\pi_k(1-\pi_k)\}$.

We know that $\Delta_{kk} = \pi_k(1-\pi_k)$ for all $i \in 1,\ldots,N$. Every off-diagonal entry of the matrix $\boldsymbol{\Delta}$ is negative because the selections of the units are all negatively correlated between each other due to the Yates-Grundy condition. Another property of $\boldsymbol{\Delta}$ is that the sum of every row is equal to 0 due to fixed sample size. We obtain that $\pi_k(1-\pi_k) = \sum_{\ell\in U|\ell\neq k}\Delta_{k\ell}$ for all $k \in U$. Thus,

$$\max_{k\in U}\left(\sum_{\ell\in U}|\Delta_{k\ell}|\right) = 2\max_{k\in U}\left\{\pi_k(1-\pi_k) + \sum_{\ell\in U|\ell\neq k}\Delta_{k\ell}\right\} = 2\max_{k\in U}\left\{\pi_k(1-\pi_k)\right\}.$$

The result follows from Result 4.

**Corollary 2.** *For any sampling design with a fixed sample size $n$ on a finite population $U$ of size $N$, the largest eigenvalue of $\boldsymbol{\Delta}$ is smaller than or equal to* $2\max_{k\in U}\left(\sum_{\ell\in U|\Delta_{k\ell}>0}\Delta_{k\ell}\right)$.

Corollary 2 shows that Yates-Grundy conditions are desirable for fixed samples size designs as they give a lower theoretical maximal eigenvalue of $\boldsymbol{\Delta}$. It also indicates that sampling designs with fixed sample size that have a lot of units $a_i$ that are strongly positively correlated can be very weakly robust. Systematic sampling, which gives very high or maximal possible eigenvalues, corresponds to such sampling design. We have also calculated results for the minimal maximal eigenvalue of $\boldsymbol{\Delta}$.

## 4. Balance and robustness trade-off

Let matrix $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top$ denote the $N \times p$ matrix of $p$ covariates. *Balanced sampling* are methods for selecting random samples $\mathbf{a}$ that approximatively satisfies the balancing equations:

$$\sum_{k\in U}\frac{a_k\mathbf{x}_k}{\pi_k} \approx \sum_{k\in U}\mathbf{x}_k. \tag{3}$$

DRAFT

Fixed sample size is a special case of balanced sampled when matrix $\mathbf{X}$ contains a variable that is equal or proportional to the inclusion probabilities. Define $\mathbf{D} = \text{diag}(\boldsymbol{\pi})$. For sampling designs that select balanced samples, we deduce from Equation (3) that $\text{var}_a(\mathbf{X}^\top \mathbf{D}^{-1}\mathbf{a}) = \mathbf{X}^\top \mathbf{D}^{-1}\boldsymbol{\Delta}\mathbf{D}^{-1}\mathbf{X} \approx 0$. In other words, balanced sampling designs will have a covariance matrix $\boldsymbol{\Delta}$ with $p$ eigenvalues that are equal or close to 0. If the inclusion probabilities are all equal, the kernel of $\boldsymbol{\Delta}$ can be generated by the $p$ covariates from $\mathbf{X}$. If we consider that the inclusion probabilities are fixed, then the trace $\boldsymbol{\Delta}$ is fixed. The trace is equal to the sum of the eigenvalues. Therefore, the use of a balanced sampling design induces that some eigenvalues are equal or close to 0, which will be pushed on the other non-zero eigenvalues of $\boldsymbol{\Delta}$.

In balanced designs, it seems difficult to bound the eigenvalues more tightly than in Result 4. Indeed, balanced designs can respect the fixed sample size, but do not respect the Yates-Grundy conditions. Ideally, one would like balancing to cause an approximately identical increase in all directions for variables that are orthogonal to the balancing variables. On the basis of simulations, this seems to be the case. The Yates-Grundy conditions are not satisfied and it is therefore difficult to bound the largest eigenvalue. For algorithms that output balanced samples, the matrix $\boldsymbol{\Delta}$ can usually only be estimated using simulations, which adds another layer of difficulty to bound the largest eigenvalue.

Expression (2) shows that if the objective is to minimize the worst-case design-based variance, then the largest eigenvalue of $\boldsymbol{\Delta}$ should be as small as possible. If we assume that the covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ have an effect on the response, then the advantages of balanced sampling appear. Suppose that $y_k^C = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k, y_k^T = \mathbf{x}_k\boldsymbol{\beta} + \tau + \varepsilon_k$ and that all inclusion probabilities are equal to $\pi$ for any $k \in \{1, \ldots, N\}$. We assume that the sample size is fixed in order to make the term related to $\tau$ in the following expression disappear. Using Expression (2), the design-based variance becomes

$$\text{var}_a(\widehat{\tau}) = \left\{\frac{1}{N\pi(1-\pi)}\right\}^2 \left\{\boldsymbol{\beta}^\top \text{var}_a(\mathbf{X}^\top\mathbf{a})\boldsymbol{\beta} + \boldsymbol{\varepsilon}^\top\boldsymbol{\Delta}\boldsymbol{\varepsilon} + 2\boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{\Delta}\boldsymbol{\varepsilon}\right\}, \tag{4}$$

where $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \ldots, \varepsilon_N)$.

In order to minimise the variance under the design of $\widehat{\tau}$ in the worst-case scenario, both the maximal eigenvalue of $\boldsymbol{\Delta}$ and $\text{var}_a(\mathbf{X}^\top\mathbf{a})$ should be considered. Minimizing both terms is not possible so a trade-off has to be considered.

DRAFT

## References


[1] BERTAIL, P., CHAUTRU, E., CLÉMENÇON, S.: Empirical processes in survey sampling with (conditional) Poisson designs. *Scandinavian Journal of Statistics 44*, 1 (2017), 97–111.

[2] BRÄNDÉN, P., JONASSON, J.: Negative dependence in sampling. *Scandinavian Journal of Statistics 39*, 4 (2012), 830–838.

[3] BREWER, K. R. W., HANIF, M.: *Sampling with Unequal Probabilities*. Springer, New York, (1983).

[4] DOL, W., STEERNEMAN, T., WANSBEEK, T.: Matrix algebra and sampling theory: The case of the Horvitz-Thompson estimator. *Linear algebra and its applications 237* (1996), 225–238.

[5] GERBER, M., CHOPIN, N., WHITELEY, N.: Negative association, ordering and convergence of resampling methods. *The Annals of Statistics 47*, 4 (2019), 2236–2260.

[6] HARSHAW, C., SLAVJE, F., SPIELMAN, D. A., ZHANG, P.: Balancing covariates in randomized experiments with the gram–schmidt walk design. *Universtiy of Yale, unpublished* (2022).

[7] HORVITZ, D. G., THOMPSON, D. J.: A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47* (1952), 663–685.

[8] JOHANSSON, P., SCHULTZBERG, M., RUBIN, D. B.: On optimal re-randomization designs, (2019). Working paper, Department of Statistics, Uppsala University.

[9] KAPELNER, A., KRIEGER, A. M., SKLAR, M., AZRIEL, D.: Optimal rerandomization via a criterion that provides insurance against failed experiments, (2019) Available at https://arxiv.org/abs/1905.03337

[10] LI, X., DING, P., RUBIN, D. B.: Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences 115*, 37 (2018), 9157–9162.

[11] MORGAN, K. L., RUBIN, D. B.: Rerandomization to improve covariate balance in experiments. *The Annals of Statistics 40*, 2 (2012), 1263–1282.
[12] MORGAN, K. L., RUBIN, D. B.: Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association 110*, 512 (2015), 1412–1421.
[13] SIMON, R.: Restricted randomization designs in clinical trials. *Biometrics 35* (1979), 503–512.
[14] TILLÉ, Y.: *Sampling Algorithms*. Springer, New York, (2006).
[15] TILLÉ, Y.: *Sampling and Estimation From Finite Populations*. Wiley, Hoboken, (2020).
[16] TILLÉ, Y.: Some solutions inspired by survey sampling theory to build effective clinical trials. *International Statistical Review 90*, 3 (2022), 481–498.
[17] XU, Z., KALBFLEISCH, J. D.: Propensity score matching in randomized clinical trials. *Biometrics 66*, 3 (2010), 813–823.
[18] YATES, F., GRUNDY, P. M.: Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B15* (1953), 235–261.

DRAFT

# Statistical Matching: Hotdeck or Propensity Score?

Elena Dalla Chiara[a], Marcello D'Orazio[b], and Federico Perali[c]

[a] University of Verona, Center of Economic Documentation (CIDE), Verona, Italy; elena.dallachiara@univr.it

[b] Italian National Institute of Statistics (Istat), Department for Statistical Production, Rome, Italy; marcello.dorazio@istat.it

[c] University of Verona, Department of Economics, Verona, Italy; federico.perali@univr.it

**Abstract**

This empirical study compares hotdeck and propensity score statistical matching (SM) procedures with the aim of providing useful guidance to researchers interested in the implementation of SM methods. Both methods are statistically robust, but the implementation of the propensity score method needs much more fine tuning.

**Key words:** Data fusion, matching methods, nearest neighbor

DRAFT

## 1. Introduction

In developed countries, socio-economic surveys with a national coverage normally focus on specific topics such as income, consumption, time use and quality of life to reach a high level of data detail and accuracy. When the research or policy objective is to obtain a reliable representation of living standards, costs of living, labor choices it is necessary to combine different surveys implementing a statistical matching procedure.

*Statistical matching* (SM) refers to a wide set of methods that can be divided into three classes: nonparametric, parametric, and mixed. The hotdeck SM methods, and in particular the nearest neighbor distance hotdeck SM method (D'Orazio et al. 2006), are very popular approaches within the non-parametric class. Parametric SM approaches assume a specific statistical model whose parameters are estimated exploiting all the available data. On the other hand, mixed SM approaches combine elements of both parametric and nonparametric approaches. Under the umbrella of the mixed SM approaches we can include also the methods related to the propensity score matching approach, traditionally applied also to estimate causal effects using observational data (Dalla Chiara et al. 2019, Kum and Masterson 2010, Rosenbaum and Rubin 1983 and 1985).

To the best of our knowledge, the statistical matching literature offers statistical comparisons of different methods within the same framework (Caliendo and Kopeinig 2008, Dehejia and Wahba 2002, Gu and Rosenbaum 1993, Rosenbaum and Rubin 1985), but comparisons of the statistical performance of competing nonparametric and semi-parametric methods are seldom done. This empirical study goes in this latter direction and compares hotdeck and propensity score SM procedures within the same experimental setting aimed at the integration of the US Annual Social and Economic Supplement, providing accurate information about income, with the US Consumer Expenditure Survey, containing information about both expenditures and incomes. The final objective consists in the creation of a robust integrated data set.

## 2. Methodology

Statistical matching (also known as *data fusion*) indicates a large set of statistical methods to integrate two data sources, identified as A and B, referred to the same target population (typically data from independent probabilistic sample survey) with the objective of investigating the relationship between variables, Y and Z, not jointly observed in a single data source, i.e. the Y variable is observed only in A while Z is available solely in B (D'Orazio et al. 2006). SM exploits information shared by both the sources, commonly identified as *common variables* $(X_1, X_2, \dots, X_p)$ assumed to have the same definition. Most SM methods integrate data at the micro-level to create a "synthetic" dataset including all the variables needed for planned subsequent analyses. The dataset, often denoted as *matched* or *fused*, is 'synthetic' because it is not the result of a direct observation.

In SM applications, only a subset of the common variables, said *matching variables*, is considered; parsimony is the guiding principle in selecting them (D'Orazio et al. 2017). Unfortunately, this way of working implicitly introduces an underlying "model" for the relationship between Y and Z; in other words, it is assumed that Y and Z are independent conditional on the matching variables (cf. D'Orazio et al. 2006). This assumption is seldom valid in SM applications; generally, it approximately holds when the set of matching variables includes at least a proxy variable of one of the target variables; i.e. a variable having a very strong correlation/association with Y or Z (see e.g. Donatiello et al. 2016).

### 2.1 Hotdeck

SM *hotdeck* methods consist in imputing the missing target variable in the A data source considered as reference (*recipient*) with values of Z observed on selected donors from B (donor dataset). The *nearest neighbor distance* (NND) is the most popular and consists in imputing for each observation in *A* the value of *Z* observed on the closest donor in *B*, whereas closeness is measured in terms of a distance (e.g. Manhattan, Euclidean, Gower, etc.) calculated on the chosen matching variables. It is common to divide units in both *A* and *B* in suitable donation classes selected from matching variables such that donation is allowed only between units belonging to the same class.

The *rank distance* hotdeck is a special case of the NND. It calculates the distance on a single continuous matching variable, but it replaces the observed values with the corresponding percentage points of the estimated empirical cumulative distribution. This device permits to overcome measurement errors that introduce bias in the observation of the chosen variable in one of the surveys that would make impossible a direct comparison of the observed values. This method is particularly suited to handle cases of a single matching variable that is a strong proxy of one of the target ones. For instance, in SM of socio-economic surveys aimed at studying the relationship between income and consumption, it may happen that the income variable is observed in both surveys but in one case it is heavily affected by negative bias because observed values are systematically below the true income. Introduction of donation classes often requires estimating the empirical cumulative distribution for each donation class, but the results are reliable if donation classes are not too small.

Finally, *random* hotdeck is designed to handle a set of matching variables that are all categorical. Essentially, for each recipient unit having given characteristics, the donor is chosen at random among the units showing the same characteristics in the donor dataset. This method corresponds to estimating the distribution of Z conditional on the matching variables in the donor dataset and then drawing an observation from it (D'Orazio 2015).

### 2.2 Propensity Score

The propensity score (PS) when applied in SM applications can be seen as a mixed approach that (1) fits an explicit statistical model (logistic model) and then (2) imputes Z in the recipient dataset applying matching algorithm that resemble hotdeck to the predictions (*scores*) of the model fitted in the first step. PS and hotdeck SM approaches differ in terms of: (i) the procedure adopted for the selection of the used common variables, that in the PS method (first step) is not necessarily parsimonious but requires the balancing of the logistic model that predicts the scores, and (ii) in the second step for the techniques used in associating the donor's observations to the nearest neighbor in the recipient survey.

The *balancing score* $b(X)$ is a function of a subset of the observed covariates $X$ such that the conditional distribution of $X$ given $b(X)$ is independent of assignment in the treatment. Therefore, the propensity score, one of the possible balancing scores, is the probability of treatment assignment conditional

DRAFT

on a set of measured covariates (Rosenbaum and Rubin, 1983). This means that for each value of $b(X)$, the distribution of the observed covariates is the same in the treated and untreated group. The propensity score is based on two assumptions that imply the strong ignorability in treatment assignment (Rosenbaum and Rubin 1983): the conditional independence and the common support. In the context of statistical matching the terms treated, and control refer to the recipient data set and the donor data set respectively.

The PS is estimated using a logistic model on selected set of covariates common to both surveys and its estimated score can be considered a synthetic indicator of the shared variables used in this function. The accuracy of the model specification is evaluated inspecting the distribution of both the propensity score, that also allows to investigate the region of common support, and the common covariates. Balance consists in exploring if the distribution of common covariates is similar in the recipient and donor database. After conditioning on propensity score, there should be no systematic differences in observed covariates between the two data sets (Austin 2011). The standardized difference (SD) is the most common balance diagnostic measure; it is computed as the ratio of the mean difference between recipient and donor sample and the root square of the pooled variance. A SD less than 0.1 denotes a negligible difference (Normand et al. 2001). Austin (2009) and Imai et al. (2008) also proposed to investigate higher moments of the distribution in particular the variance ratio between the two groups. A ratio close to 1 means group balance.

Practically, in the second step of the SM procedure the estimated PS is used to associate donor's observations to the closest observation in the recipient sample implementing a matching method. There is not an algorithm valid for all circumstances, but the choice should be evaluated case-by-case on the data structure and every choice entails a trade-off between bias and variance (Caliendo and Kopeinig 2008, Dehejia and Wahba 2002).

In our work we adopt the nearest neighbor (NN) method to facilitate comparability with the hotdeck procedure, that also uses a NN algorithm, implementing a NN with replacement using four donor observations to match one unit in the recipient data set. In HD applications the term "NN distance" refers to the same NN algorithm as in the PS, but the distance in the HD context is a choice among options like Euclidean or Manhattan. On the other hand, the implementation of the PS method does not require a choice among distances because the implied distance is the propensity score.

## 3. Data description

In the SM application, we use the 2019 wave of the Annual Social and Economic Supplement (ASEC) and the Consumer Expenditure Survey (CE), both administrated by the US Census Bureau. The surveys are nationally representative and share the same sampling design, a multistage stratified sample. The primary sample units (PSUs) are small clusters of counties grouped together into geographic entities called "core-based statistical areas" (CBSAs).

The common variables were aligned in terms of definition and classification, and we also compared their distribution to understand whether differences in the estimated distributions may point to problems that hinder comparability.

The CE collects data on expenditure, income, and demographics. The CE consists of two separate surveys: the Interview Survey and the Diary Survey. In this study we use the Interview Survey that includes 9,453 households (HHs).

The ASEC is an annual cross-sectional survey that provides the basic monthly demographic and labor force data collected with the Current Population Survey (CPS), plus additional data on work experience, income, noncash benefits, health insurance coverage, and migration. The ASEC survey is larger than the CE survey (47,475 HHs after some cleaning) but since we decided to use it as recipient database (recipient should be smaller than the donor in the application of hotdeck SM techniques) then we selected a subsample of 9,000 units with probability proportional to the sample weight. This approach ensures having approximatively a representative subsample.

## 4. Results

In this section we present some results of the SM approaches, where the ASEC subsample plays the role of recipient (treated), while the CE assumes the role of donor (control). The objective is to impute in ASEC the consumption variables observed in CE.

DRAFT

### 4.1 Hotdeck

Following the parsimony principle, the selected set of matching variables consists of relatively few strong predictors of the target variables: the HH size, the number of HH members with an employment and the HH annual income. This latter variable is also available in CE survey and shows an estimated distribution close to that observed in ASEC, which however remains the most reliable survey for investigating HH income. Including HH income in the set of matching variables is fundamental in SM as it makes holding the conditional independence assumption, i.e. the independence between income and consumption conditional on the matching variables.

In the application of the NND hotdeck the units in both data sources are divided in groups obtained crossing HH size with the number of employed HH members, then for each recipient unit in a given donation class it is imputed the consumption observed on the donor unit in CE having the closest distance in terms of HH income. This way of working considers the income observed in the two data sources as perfectly comparable and therefore exploitable for calculating the distances.

If it is assumed that HH income observed in CE cannot be directly compared with the corresponding variable available in ASEC, it is still possible to indirectly include HH income in the matching step by replacing its values with corresponding percentage points of the estimated empirical cumulative distribution. This approach corresponds to the rank distance hotdeck. The empirical cumulative distribution of the HH income is estimated separately for each category of the HH size. Consequently, a recipient HH with a given size "receives" the expenditure variables observed in the closest HH being the closest in terms of estimated empirical cumulative distribution of the income conditional to the HH same size.

A slight modification of rank distance hotdeck consists in picking the donor at random, that is a donor among those being at a distance less or equal to a given threshold in terms of distance calculated on the estimated empirical cumulative distribution of the HH income conditional to the HH size. In our application the chosen threshold is 0.05.

DRAFT

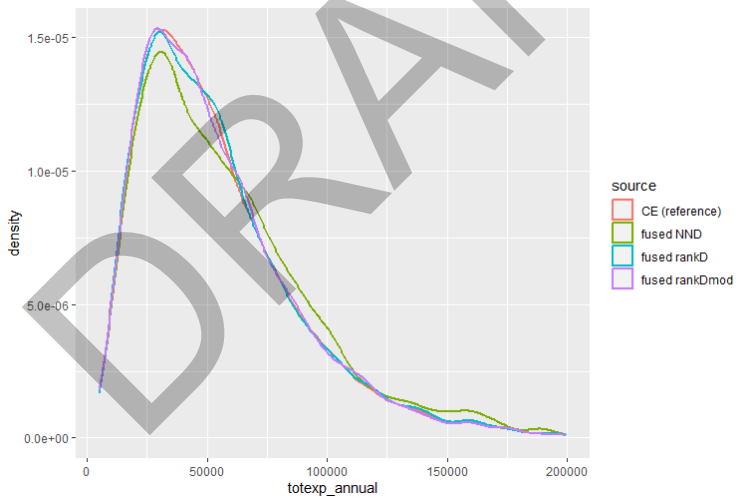



Figure 1 Estimated distribution of HH annual consumption expenditures in the fused dataset obtained after the applied hotdeck SM procedures and in the CE survey (reference).

The common way of assessing the results of the SM applications consists in comparing the estimated distribution of the imputed variable Z (HH consumption expenditures in our case, "totexp_annual") in the recipient dataset after the SM procedure with the reference distribution estimated on the donor. Figure 1 shows that the NND distance using directly the income ("fused NND" in the plot) performs poorly if compared to the other methods that contrarily use the percentage points of the conditional percentage points of the estimated empirical cumulative distribution of the income. In general, the slight modification of rank distance hotdeck ("fused rankDmod" in the plot) produces a curve quite close to the reference one estimated on CE ("CE (reference)" in the plot). This result is confirmed when looking at the estimated average values of HH consumption expenditures in different subsets of HHs formed considering the HH size, the number of perceivers in the HH, etc. (details not reported for sake of summary).

## 4.2 Propensity Score

The outcomes of PS matching are affected by the choice of the most appropriate propensity score model specification and by the matching algorithm, including the choice of the matching ratio and the replacement or not of the donor units. These issues affect the marginal distribution of the imputed variable.

We implemented two PS models. The first includes the same predictor variables used in the hotdeck methodology. We term this parsimonious specification as "short" model (Goldberger 1991) noting that in the hotdeck nonparametric method only short specifications are preferable due to the characteristics of these methods and more in general to the "curse of dimensionality" problem. This specification is used with the purpose to compare the result of the two methodologies to control for the confounding effect stemming from a different model specification. Donors with the same propensity score are more likely to have a different pattern of covariates when the propensity score model is less accurate than with a more detailed model. The propensity score balances only with respect to the common covariates, not preventing residual confounding by unmeasured factors. In a statistical imputation framework, this aspect has a strong implication on the precision of variables imputation.

For this reason our main selected model specification includes: HH size, number of employed members in the HH, HH income quintile, region of residence, education of the HH head, mortgage, age of the HH head and family type.

The quality assessment of the matching procedure first investigates the distribution of the PS in the original and matched data. The estimated PS has a suitable overlap in terms of support of distribution with a good control match for each observation in the recipient group, including the observations on the tail of the distribution, in both models, but in the main model "unmatched control units" covers almost all the same range of "matched units" in treated and control group. We can conclude that a parsimonious specification in the PS implementation leads to an underspecified logistic model and therefore we report the results only for the fully specified model.

The balance is not difficult to obtain because the donor and recipient databases have a good intrinsic balance so that in our data the balance is not influenced by the matching algorithm chosen as shown in Figure 2 (left panel).

DRAFT

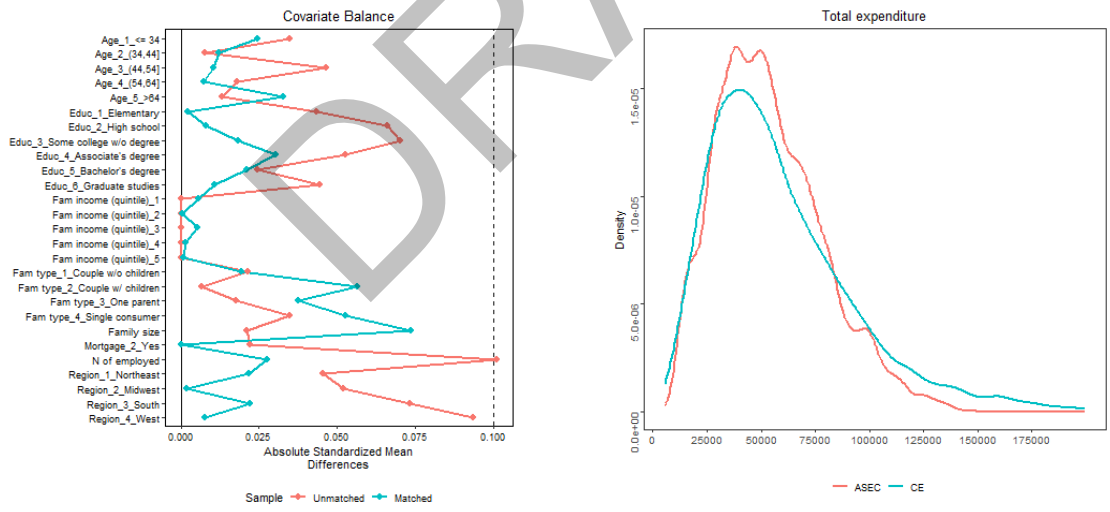



Figure 2 Covariate balance before and after matching (left panel) and Distribution of total expenditure in fused data set (ASEC) and CE data set (right panel)

To examine the matching quality, we inspect the distribution of the extra information (Z) that we transfer from the donor sample to obtain the fused ASEC data set. We control both if in the recipient sample these variables preserve the same distribution as the donor sample, and then we look at the marginal distribution of the imputed variable by the single value of the covariates used to estimate the PS in both databases. For illustrative purposes in Figure 2 (right panel) we report the marginal distribution only for total expenditure that is well preserved. To compare the marginal distribution in the recipient and donor data sets we compute both the ratio of mean and the ratio of median calculated as the ratio between the quantity in the recipient and in the donor sample. There is not a specific threshold

supporting that the information in the two samples can be considered similar, but the closer the ratio is to 1, the greatest the correspondence. The marginal distribution of the common covariates is well preserved in the fused database compared to the donor data set. Major discrepancies are observed in the lower and in the highest income category, in the lower level of education and in families with one parent.

## 5. Conclusions

This empirical exercise compares hotdeck and propensity score statistical matching (SM) procedures. We hope that our results will provide useful guidance to researchers interested in the implementation of SM methods who can now be reassured because both methods are statistically robust and perform SM satisfactorily. We do warn though that the implementation of the propensity score method needs much more care in the execution of the matching steps. Failing to do so may expose the procedure to significant biases.

## References


[1] Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine (2009) doi: 10.1002/sim.3697

[2] Austin, P.C.: An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research (2011) doi: 10.1080/00273171.2011.568786

[3] Caliendo, M., Kopeing, S.: Some practical guidance for the implementation of the propensity score matching. Journal of Economic Surveys (2008) doi: 10.1111/j.1467-6419.2007.00527.x

[4] Dalla Chiara, E., Menon, M., Perali, F.: An Integrated Database to Measure Living Standards. Journal of Official Statistics (2019) doi: 10.2478/jos-2019-0023

[5] Dehejia, R.H, Wahba, S.: Propensity score-matching methods for nonexperimental causal studies. The Review of Economics and Statistics (2002) doi: 10.1162/003465302317331982

[6] Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M, Spaziani, M.: The role of the conditional independence assumption in statistically matching income and consumption. Statistical Journal of the IAOS (2016) doi: 10.3233/SJI-161000

[7] D'Orazio, M.: Integration and imputation of survey data in R: the StatMatch package. Romanian Statistical Review 2/2015, 57--68 (2015)

[8] D'Orazio, M., Di Zio M., Scanu, M.: Statistical Matching, Theory and Practice. Wiley, Chichester (2006)

[9] D'Orazio, M, Di Zio, M., Scanu, M.: The Use of Uncertainty to Choose Matching Variables in Statistical Matching. International Journal of Approximate Reasoning (2017) doi: 10.1016/j.ijar.2017.08.015

[10] Goldberger, A.S.: A Course in Econometrics. Harvard University Press, Cambridge (1991)

[11] Gu, X.S, Rosenbaum, P.R.: Comparison of Multivariate Methods: Structure, Distances, Algorithms. Journal of Computational and Graphical Statistics (1993) doi: 10.2307/1390693

[12] Imai, K., King, G., Stuart, E. A.: Misunderstandings between experimentalists and observationalists about causal inference". Journal of the Royal Statistical Society, Serie A (Statistics in Society) (2008) doi 10.1111/j.1467-985X.2007.00527.x

[13] Kum, H., Masterson, T.N.: Statistical Matching using Propensity Score: Theory and Application to the Analysis of the Distribution of Income and Wealth. Journal of Economic and Social Measurement (2010) doi: 10.3233/JEM-2010-0332

[14] Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., McNail, B. J.: Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. Journal of Clinical Epidemiology (2001) doi 10.1016/s0895-4356(00)00321-8

[15] Rosenbaum, P.R., Rubin, D.B.: The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika (1983) doi: 10.1093/biomet/70.1.41

[16] Rosenbaum, P.R., Rubin, D.B.: Construction a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistician (1985) doi: 10.2307/2683903


DRAFT

# A Hierarchical Spatio-Temporal Model for Time-Frequency Data: An application in bioacoustic analysis

Hiu Ching Yip[a], Gianluca Mastrantonio[a], Enrico Bibbona[a], Daria Valente[b], and Marco Gamba[b]

[a]Politecnico di Torino, Italy; hiu.yip@polito.it, gianluca.mastrantonio@polito.it, enrico.bibbona@polito.it

[b]Università di Torino, Italy; daria.valente@unito.it, marco.gamba@unito.it

DRAFT

**Abstract**

A hierarchical spatio-temporal model that infers the latent spectral shape from a set of bio-acoustic signals by means of the Nearest neighbour Gaussian process is proposed. The model aims to account for the effects of the relative relationship between time and the spectral shape of the recorded vocalizations and that of time discretization. The goal is to obtain a representative model of the inherent acoustic structure of the species.

***Keywords:*** Bio-acoustic, time-frequency, spatio-temporal model, nearest neighbour Gaussian process, spectral shape

## 1. Motivation & Data

In comparative bio-acoustic studies, one area of interest is to understand the vocalizations of non-human primates in order to provide insights into the evolutionary mechanism of the communication systems of our closest relatives. Since bioacoustic data are almost always represented in the form of a spectrogram, bioacoustic analysis is therefore a form of time-frequency analysis that requires computational methods to process and learn from the bio-acoustic signals in large quantities. The most commonplace practice is to apply feature engineering methods in order to select and compare a set of basis-features. Such methods often treat the time-frequency bins of spectrograms as independent features and are known to entail perceptual bias due to the reliance on biologists to manually select relevant features. The identification and interpretation of meaningful features are usually costly to acquire and difficult to generalize for cross-species comparison. Furthermore, feature engineering methods in bioacoustic analysis, whether supervised or unsupervised, almost always ignore the effects of time by assuming that all the time-frequency bins from various recorded bio-acoustic signals are independent of each other. The aim of this project is to propose a spatio-temporal model that accounts for the effects of time in bio-acoustic analysis.

The available dataset for this work is a set of vocal signals of lemurs that were recorded in Madagascar. The format of the dataset is similar to those in (2). Each recorded analogue digital signals is

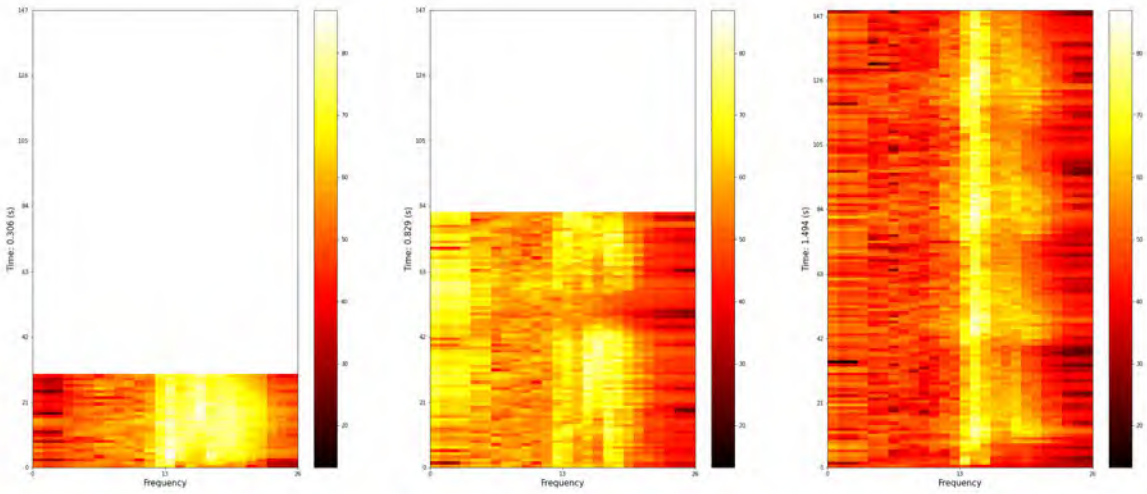



Figure 1: The spectrograms of three recorded signals labelled by the same species label and the same call-type from the dataset. The $x$-axis and $y$-axis represent the discretized frequency and time domain, respectively. Each discrete sound point of the discretized recording is measured in decibel scale on the third dimension of the spectrogram. Each axis is labelled by the number of discrete time and frequency coordinates. The number of frequency coordinates on the $x$-axis is the same for all recorded signals, while the number of time coordinates on the $y$-axis differ. This discernible disparity between the numbers of time coordinates on the $y$-axes is a result of the difference between the durations of the three signals. The unique duration of each signal is indicated on the label of the $y$-axis, which are $0.306$, $0.829$ and $1.494$ seconds respectively (from left to right). The corresponding numbers of time coordinates are $30$, $82$, and $149$ respectively, while all have $26$ frequency coordinates.

DRAFT

discretized using a constant time-step of $0.01$ seconds during the initial stage of signal pre-processing. Figure 1 is a spectrogram representation of 3 discretized recorded signals of different durations from the dataset. Each signal is categorized by one species label and one call-type label. The species label is the acronym of the scientific name of the lemur that emitted the recorded signal, whereas the call-type label of each recording is assigned according to the behaviour of the animal during the emission of the signal.

## 2. The Model

Assume that the recordings that are characterized by the same species but different behavioural call types are independent from each other. The model specification is then restricted to the recorded signals that are labelled by a single call type from a single species. Let $N$ be the total number of signals that are classified by the same combination of species and call type with $i = 1, \cdots, N$. As per Figure 1, each recorded signal is represented by a spectrogram with one axis representing the time domain and another representing the frequency domain. Assume that each $i$-th recorded signal is a realization of a two-dimensional process $\mathcal{Y}_i(t, h) \in \mathbb{R}$ where $t \in \mathbb{R}_{\geq 0}$, $h \in \mathbb{R}$ over an observed regular grid. Let $n_{t,i}$ denote the number of time coordinates on the discrete time axis $\mathcal{T}_i^* \subset [0, l_i]$ where $l_i$ is the duration of the signal and let $n_h$ denote the number of log-frequency bins on the frequency axis $\mathcal{H}$, respectively. The regular grid of the $i$-th recorded signal is then composed of $n_i = n_{t,i} \times n_h$ time-frequency coordinates in total. In order to compare the recorded signals of different durations, the time domain of each recorded

signal $\mathcal{T}_i^* \subset [0, l_i]$ is rescaled into a new time domain, denoted by $\mathcal{T}_i \subset [0, 1]$, such that it is always one in duration. By contrast, the number of frequency coordinates $n_h$ is constant for all signals and so, it follows that the log-frequency bands that are denoted by the frequency coordinates on $\mathcal{H}$ are also the same for all recorded signals. Hence, let the realization of the process $\mathcal{Y}_i(t, h)$ on a regular grid be denoted by $\boldsymbol{y}_i = \{y_{i,t,h}\}_{t \in \mathcal{T}_i, h \in \mathcal{H}}$ where

$$\mathcal{T}_i = \left\{ \frac{k-1}{n_{t,i}-1} | k = 1, \ldots, n_{t,i} \right\}, \quad \mathcal{H} = \{0.23k + \log 63 | k = 1, \ldots, n_h\} \tag{1}$$

Clearly, each observed spectrogram represented by the $n_{t,i} \times n_h$ regular grid $\mathcal{T}_i \times \mathcal{H}$ is unique in its own right as a consequence of the varying numbers of time coordinates $n_{t,i}$. As explained earlier on, the aim of the model is to infer the latent spectral shape of vocalizations from a dataset of $N$ recorded signals that share the same species and call type labels. It is reasonable to assume that all $N$ recorded acoustic signals have the same inherent spectral shape which can be described by the same latent Gaussian process. This ideal representation of the inherent spectral shape of the vocalizations from a single species of a single call type is henceforth termed the "mother call". Let $\mathcal{W}_i(t, h)$ be the latent process that describes the mother call. Assume that if $(i, t, h) \neq (i', t', h')$, then $\mathcal{Y}_i(t, h)$ and $\mathcal{Y}_{i'}(t', h')$ are conditionally independent given the mother call. The model is:

$$\begin{aligned} \mathcal{Y}_i(t, h) &= \mu_i + \mathcal{W}(t, h) + \epsilon_i(t, h) \\ \mathcal{W}(t, h) &\sim \text{GP}(0, C(\cdot, \cdot | \boldsymbol{\theta})) \\ \epsilon_i(t, h) &\sim \text{GP}(0, \tau_i^2) \end{aligned} \tag{2}$$

where $\mu_i \in \mathbb{R}$ is the mean sound intensity of the $i$-th recorded signal, $\mathcal{W}(t, h)$ is the latent process over the domain of the mother call, $\epsilon_i(t, h) \sim \text{GP}(0, \tau_i^2)$ is the $i.i.d.$ random noise, and

$$C((t, h), (t', h') | \boldsymbol{\theta}) = Cov(\mathcal{W}(t, h), \mathcal{W}(t', h')) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^+ \tag{3}$$

is the cross-covariance function for the latent process $\mathcal{W}(\cdot, \cdot)$ that is dependent on the vector of parameters $\boldsymbol{\theta}$. The covariance function $C(\cdot, \cdot | \boldsymbol{\theta})$ for the latent process $\mathcal{W}(\cdot, \cdot)$ is defined by

$$C((t, h), (t', h') | \boldsymbol{\theta}) = \sigma^2 \text{exp}(-\phi_t |t - t'| + \phi_h |h - h'|) \tag{4}$$

such that $\boldsymbol{\theta} = (\sigma, \phi_t, \phi_h)$. The parameters that need to be inferred are the time-frequency decay $\phi_t, \phi_h$ and the variance $\sigma$, respectively. Note that the equivalent formulation is the generative model $\mathcal{Y}_i(t, h) \mid \mathcal{W}(t, h) \sim \text{GP}\big(\mu_i + \mathcal{W}(t, h), \tau_i^2\big)$. Dependence between the entire set of recorded signals is thus introduced through the latent process $\mathcal{W}(t, h)$ which describes the mother call. That is, if the model in equation (2) is marginalized over the latent process, then the observed processes are dependent on each other. Subsequently, the acoustic structures of each recorded signal is composed of the natural change in the spectral shape across the observed time-frequency grid with independent error $\tau_i^2$, which need to be accounted for by an additional component on the diagonal of the covariance matrix for the observed processes. This is the nugget effect that arises from the covariance of the variables in each observed process. Let $\mathbb{I}(\cdot)$ be an indicator function, then the covariance function for the observed processes is:

$$Cov\big(\mathcal{Y}_i(t, h), \mathcal{Y}_{i'}(t', h')\big) = C((t, h), (t', h') | \boldsymbol{\theta}) + \tau_i^2 \mathbb{I}((i, t, h) = (i', t', h')) \tag{5}$$

Since direct implementation of the generative model in equation (2) is computationally infeasible due to the fact that it requires the mother call to be sampled for every single recorded sound point in the dataset, the marginalized model is implemented instead. To simplify notations, re-write each single recorded sound point $y_{i,t,h} = y_{i,j}$ where $y_{i,j}$ is the realization of $\mathcal{Y}_i(t, h)$. Define $\boldsymbol{y}_i = \{y_{i,j} | j = 1, \cdots, n_i\}$ as the vector of realizations from the $i$-th recorded signal such that the elements are sorted in the ascending order of time and the increasing value of log-frequencies within each time. Define $\mathbf{1}_i$ as the $n_i \times 1$ vector of ones. Write $\boldsymbol{\Sigma}_i$ as the exact covariance matrix of $\boldsymbol{y}_i$ and write $\boldsymbol{\Sigma}_{i,i'}$ as the exact

DRAFT

cross-covariance matrix of $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i'}$ that are given by the covariance function in equation (5). The joint density of all realizations can be expressed by:

$$\begin{matrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\\ \vdots\\ \boldsymbol{y}_N\end{matrix} \sim \mathrm{N}\left(\begin{pmatrix}\mu_1\mathbf{1}_1\\ \mu_2\mathbf{1}_2\\ \vdots\\ \mu_3\mathbf{1}_N\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{1,2} & \cdots & \boldsymbol{\Sigma}_{1,N}\\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_2 & \cdots & \boldsymbol{\Sigma}_{2,N}\\ \vdots & \cdots & \ddots & \vdots\\ \boldsymbol{\Sigma}_{N,1} & \boldsymbol{\Sigma}_{N,2} & \cdots & \boldsymbol{\Sigma}_N\end{pmatrix}\right) \tag{6}$$

In a more compact form, write $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1,\cdots,N}$ as the collection of all realizations and re-write formula (6) into $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{y}})$.

## 3. Implementation

The available dataset of lemur signals for this paper has a total number of time-frequency coordinates of $n = \sum_i^N n_{t,i} n_h$. With a computational cost of $\mathcal{O}(n^3)$, the inversion of the exact covariance matrix in equation (6) is too computationally expensive; accordingly, one of the methods for efficiently and accurately approximating Gaussian processes, namely the Nearest Neighbours Gaussian Process (NNGP) method, is adopted in this work. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the density of the realizations $\boldsymbol{y}$ that depends on the parameters $\boldsymbol{\theta}$ which can be decomposed by conditioning as follow:

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= f(\boldsymbol{y}_1|\boldsymbol{\theta}) \prod_{i=2}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta}) \\ &= f(y_{1,1}|\boldsymbol{\theta}) \prod_{j=2}^{n_1} f(y_{1,j}|y_{1,j-1}, y_{1,j-2}, \cdots, y_{1,1}, \boldsymbol{\theta}) \times \\ &\prod_{i=2}^{N} f(y_{i,1}|\boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \cdots, \boldsymbol{y}_1, \boldsymbol{\theta}) \prod_{j=2}^{n_i} f(y_{i,j}|y_{i,j-1}, y_{i,j-2}, \cdots, y_{i,1}, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \cdots, \boldsymbol{y}_1, \boldsymbol{\theta}) \end{aligned} \tag{7}$$

The idea of the NNGP method is that for a Gaussian process which is stationary, if the covariance function is monotonic with respect to the distances between the spatio-temporal coordinates, then only the immediate neighbourhoods rather than the entire conditional sets are necessary to approximate the likelihoods of the realizations of the process. Define $\mathcal{N}_{i,j}$ as a subset of variables in the conditional set of $y_{i,j}$ which is the immediate neighbourhood called the neighbour set. The elements of the neighbour set are called neighbours. Since the covariance function in equation (4) is monotonically decreasing with respect to the form of distance that it depends on, the neighbours in the neighbour set should have minimal non-zero distances and the formation of the neighbour set should be characterized by a distance function that measures the absolute time-frequency lags on the time-frequency domain. The above density of $\boldsymbol{y}$ can then be approximated into:

$$f(\boldsymbol{y}|\boldsymbol{\theta},) \approx \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{i,j}|\mathcal{N}_{i,j}, \boldsymbol{\theta}) \tag{8}$$

with $\mathcal{N}_{1,1}$ being an non-empty neighbour set. The distance function that gives rise to the neighbour set is defined by :

$$d((t,h),(t',h')) = |t-t'| + |h-h'| \tag{9}$$

It is more reasonable to select neighbours for $y_{i,j}$ from both the same $i$-th signal and the previous $i-1$-th signal instead of just the same recorded signal alone on the grounds that the realizations $\boldsymbol{y}_i$ and $\boldsymbol{y}_{i-1}$ are not independent in spite of the product form of the approximated density in the above equation (8). Spatio-temporal dependence between the two different recordings has to be re-introduced

DRAFT

into the approximated density through the neighbour sets. For the $j$-th observed sound variable of the $i$-th recorded signal, denote the neighbour set with elements selected only from the $i$-th signal by $\mathcal{N}_{i,j}^{i}$ and similarly, denote the neighbour set with elements selected only from the previous $i-1$-th signal by $\mathcal{N}_{i,j}^{i-1}$. The definition of the neighbour set for the realized variable $y_{i,j}$ is:

$$\mathcal{N}_{i,j} = \mathcal{N}_{i,j}^{i} \cup \mathcal{N}_{i,j}^{i-1} \tag{10}$$

Following (1), setting the size of the neighbour set in between 10 to 20 should enable an accurate approximation of the original process. Finally, the main objective of this work is to obtain the representative acoustic structure of the vocalizations that belong to the same species and call-type, the mother call, which is the finite realizations of $\mathcal{W}(t,h)$ over a specified grid. By the NNGP method, the approximated precision matrix $\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}$ admits a Cholesky decomposition and enables standard posterior sampling. Let $\mathbf{\Sigma}_{w,\boldsymbol{y}}$ be the $1 \times n$ cross-covariance vector of $\mathcal{W}(t,h)$ and $\boldsymbol{y}$. The realization of the latent process $\mathcal{W}(t,h)$ at any location can then be obtained by:

$$\mathcal{W}(t,h)|\boldsymbol{y},\boldsymbol{\theta} \sim \mathrm{N}\big(\mathbf{\Sigma}_{w,\boldsymbol{y}}\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}),\ \sigma^2 - \mathbf{\Sigma}_{w,\boldsymbol{y}}\mathbf{\Sigma}_{\boldsymbol{y}}^{-1}\mathbf{\Sigma}_{w,\boldsymbol{y}}^{\top}\big) \tag{11}$$

## 4. Discussion

As this is an ongoing and multidisciplinary work, the current proposed model is a preliminary version of the complete model and as such, further progress to the final results is being made. Another preliminary and less refined version of this work has been communicated to the 36th International Workshop on Statistical Modelling in the form of a poster presentation and the submitted abstract was published in the proceedings that is available in (3).

The current preliminary model in this paper, although incomplete, starkly contrasts many contemporary methods in bio-acoustic analysis. The model accounts for the effects of the time-varying components in the observed vocalizations by using the entire bio-acoustic dataset as realizations of a spatio-temporal process, rather than reducing the high-dimensional time-frequency data into some independent features, such as absolute pitches, with arbitrarily assigned meanings. The main contribution is that the resulting mother call, i.e. the representative acoustic structure of a particular call-type from a species, can subsequently be used to measure the distances between the inherent acoustic structures of different species in the animal kingdom. Such quantitative measures can hopefully ease cross-species comparison in bio-acoustic analysis and facilitate biological studies on the evolutionary basis for the variations of the vocal repertoires of various species. The next steps being taken in this work are summarized as follow.

A closer inspection of the spectrograms of the observed signals in Figure 1 reveals that there are two major issues with the preliminary model in equation (2): (i.) the noticeable distortions of the observed time domains with respect to each other caused by the unique duration of each recordings. Though each recorded signal can be thought of a marginal realization of the same latent process over the observed regular grid, the unique duration of each recording entails distortions of the observed time-axis with respect to each other, and perhaps misalignments of the time domain for the observed process with respect to the time domain for the latent mother call process. The same earliest recorded time coordinate from the 1st recorded signal might not coincide with the very same earliest recorded time coordinate from another signal, for instance. In fact, each observed time domain may be treated as a somewhat stretched portion of the latent time domain for the mother call. (ii.) the oscillations along the time axis on the lower frequency spectrum that arises from time discretization and signal reconstruction during the initial stage of analogue signal pre-processing. When analogue signals are being discretized in time, if the time-step is less than the period of the true waveform of the low frequency, then the sampled frequency becomes an artifact because it does not capture the original periodicty of the true waveform of the sound. This leads to a periodic artefact that oscillates along the time domain of the reconstructured, discretized signal on the lower frequency spectrum.

Both (i.) and (ii.) render the model specification and inference for the mother call more difficult as the challenges posed by the misalignments of the time domains as well as the presence of the artefacts must be resolved. In light of (i.), the preliminary model and its separable covariance function must be

DRAFT

extended to incorporate a time-distortion function with parameters that can describe and quantify the distortions of the time domains of the observed processes with respect to the latent time domain of the mother call. In view of (ii.), an additional component that can explain the periodicity of the artefacts must be included into the covariance function of the observed processes. Considering that the sampling artefacts appear solely in the recorded discretized signals, care must be taken to ensure that only the covariance for the observed processes must possess this additional component, but not the covariance for the mother call. The parameters of the time-distortion function and this additional component must be part of the inference. Once the final model is formulated, it is tested on simulated data in order to demonstrate its efficacy and to recognize if the model suffers from the problem of non-identifiability or any other technical issue in the implementation stage. After all the implementation details are sorted, the final model is then ready to be implemented on the real dataset.

## References


[1] Datta A., Banerjee S., Finley A.O. and Gelfand A.E.: Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Am. Stat. Assoc. (2016) **111**(514), 800–812

[2] Valente D., De Gregorio C., Torti V., Miaretsoa L., Friard O., Randrianarison R.M., Giacoma C. and Gamba M.: Finding meanings in low dimensional structures: stochastic neighbor embedding applied to the analysis of Indri indri vocal repertoire. Animals(Basel). (2019) **9**(5):243. doi:10.3390/ani9050243. PMID: 31096675; PMCID: PMC6562776.

[3] Yip H.C., Mastrantonio G., Bibbona E., Gamba M., Valente D.: Nearest neighbours Gaussian process model for time-frequency data: An application in Bio-acoustic Analysis. Proceedings of the 36th International Workshop on Statistical Modelling. (July 18-22, 2022. Trieste, Italy.) Available via https://www.openstarts.units.it/handle/10077/33740


DRAFT

# An integrated space-time model to evaluate the innovation drivers in Italy

Emma Bruno[a], Rosalia Castellano[b], and Gennaro Punzo[a]

[a] University of Naples Parthenope, Department of Economic and Legal Studies, via Generale Parisi 13, 80132 Naples; emma.bruno@studenti.uniparthenope.it, gennaro.punzo@uniparthenope.it
[b] University of Naples Parthenope, Department of Management and Quantitative Studies, via Generale Parisi 13, 80132 Naples; lia.castellano@uniparthenope.it

**Abstract**

The focus on innovation has increased over the past few years due to its established role as an engine of economic growth, as well as its rising contribution in fostering sustainable development. By using the number of patent applications to measure innovation activity, the paper identifies a set of variables influencing innovation and quantifies their spatial spillover effects within Italian provinces in the period from 2010 to 2017. Panel data models are extended with spatial data estimation techniques to assess the innovation patterns over time under the assumption of global spillover effects. As a result of the study, a wide range of factors contributes to local innovation capacity, from education levels to institutional characteristics, from economic resources to productive structure, as well as the externalities associated with agglomeration economies.

**Keywords:** local innovation, spatial panel, spillover effects, sustainable development

DRAFT

## 1. Introduction

The topic of innovation is increasingly debated worldwide, as it is one of the key engines for economic growth and competitiveness (Huggins and Thompson, 2015; Capello and Nijkamp 2009). Scholars have discussed innovation from different perspectives, given the phenomenon's multifaceted nature and its effect on the many dimensions of a country's development. Supporting innovation is one of the goals of the 2030 Agenda for Sustainable Development Goals (SDGs), adopted by the United Nations in 2015. Goal 9 stresses the crucial role of resilient infrastructure, inclusive and sustainable industrialisation, and innovation in achieving economic development, creating job opportunities, and fostering sustainable growth. Therefore, innovation plays a crucial role not only in promoting growth but also in sustainable development since it is the factor through which organisations, industries, communities, regions, and countries can implement innovative multidisciplinary approaches to solve the current sustainability challenges (Silvestre, 2015a). To improve sustainability performance, processes, products, and management approaches need to be adapted and innovated. The literature recognises the prominent role of innovation-focused approaches in addressing the topic of sustainable development (Silvestre, 2015b). Traditionally, firms have considered sustainability as a potential obstacle to their growth and competitiveness (Andersen, 2004). Nowadays, consumers' environmental awareness as well as social and government pressure on companies to reduce their environmental impact are increasing (Bocken et al., 2011). In this regard, companies steadily direct their innovation process from a sustainable perspective. As a result, innovation and sustainability in the broadest sense have become increasingly intertwined, so much so that they feed off each other.

Based on the above, this paper aims to assess and quantify the spatial direct and spillover effects of several determinants of innovation processes in Italian provinces.

Our approach to identifying a suitable statistical model for analysing innovation patterns is based on the well-established assumption that innovation is a spatially embedded phenomenon characterised by global spillovers (Furková, 2019; Paci et al., 2014). Further, it results from a series of processes and investments that occur over time and which take time to manifest into innovations (Capello and Lenzi, 2018). This implies that the double dimension of the phenomenon, both spatial and temporal, cannot be neglected, along with the control of all time-invariant and time-specific effects. In this regard, panel

data models are extended to account for spatial interactions between units by estimating the panel spatial Durbin model extended with spatial and time-period specific effects.

It is necessary to delve into the drivers of innovation to provide private actors and governments with information on the key levers to be forced to foster the innovation process from a sustainable growth perspective. Several authors have attempted to explain innovation and its drivers in the existing literature, mainly focusing on the role of (i) agglomeration economies within Jacob and Marshall's theory of externalities (De Groot et al., 2016) (ii) investments in R&D and human capital endowments based on the regional knowledge production function (RKPF) theoretical approach (Aronica et al., 2022; Paci et al., 2014). The present study proposes a contribution to the literature by investigating whether Jacob's or Marshall's externalities prevail in stimulating innovation in the Italian provinces, jointly considering the spatial and temporal dimensions. Furthermore, the contribution of socio-economic variables as well as their spatial lag to local innovative activity are assessed within the framework of global spillover effects.

The remainder of the article is structured as follows: Section 2 provides an overview of the reference framework; Section 3 discusses the statistical methodology and presents the data used in the analysis; finally, Section 4 concludes by presenting and discussing the empirical findings of the research.

## 2. Framework

A wide variety of factors contribute to a firm's innovativeness, such as internal resources or firm-specific characteristics, as well as the interaction between the firm and its external operating environment. A comprehensive analysis of innovative patterns cannot fail to consider the spatial scope of the agglomeration externalities. Economic theory has long recognised that agglomeration economies help determine and improve a territory's productive structure by fostering processes of spatial concentration of productive activity.

The theoretical framework proposed by Marshall (1890), Arrow (1962) and Romer (1986) considers agglomeration economies that result from the geographic proximity of firms in the same industry. These externalities, also referred to as MAR externalities by the Marshall-Arrow-Romer model, result from increased industrial specialisation within a specific region that facilitates knowledge sharing among firms and promotes innovativeness in the region. Geographic proximity and labour mobility facilitate the transmission of knowledge between spatially concentrated enterprises and among workers, constituting an engine of innovation development and productivity growth. Moreover, creating a local market for skilled workers advantages both firms and workers.

In opposition to MAR, Jacobs' (1969) theory argues that the spatial concentration of firms from different industries within a region promotes innovation. These economies of diversification, called Jacobs' externalities, are based on the idea that the diversity and variety of neighbouring firms promote information transfers and productivity growth through the exchange of complementary knowledge among heterogeneous agents. The literature, however, continues to be inconclusive in determining which of these two concepts is likely to create a more a more conducive environment for innovation in the long run.

Finally, the contribution of spatial interactions in determining innovation patterns cannot be neglected. The results of innovativeness in one unit can spread to other units, influencing their innovation performance, and geographic proximity allows for faster knowledge diffusion. Thus, innovation processes in a given area generate global spillover effects that trigger chain reactions in other spatial units (potentially in all of them). Spatial interactions are nowadays an intrinsic feature of an innovation-oriented economy where globalising contexts prompt a wide array of spatial spillovers. Consequently, local innovation patterns require spatial econometric techniques to be assessed and to address the spatial dependence (Silvestre and Ţîrcă, 2019; Cabrer-Borras and Serrano-Domingo, 2007).

## 3. Methodology and Data

The spatial dependence between panel units cannot be ignored, making it necessary to use appropriate estimation techniques to consider the data structure. With this in mind, we extend panel data models to spatial specification to follow the spatio-temporal patterns of innovation over time. Spatial interactions are modelled by taking into account the assumption of global spillover effects of innovativeness by estimating the Spatial Durbin Model (SDM). It is a spatial model configuration that entails global and flexible spillover effects (Elhorst and Halleck Vega, 2017; Firmino Costa da Silva et

DRAFT

al., 2017) obtained through the introduction of spatially lagged dependent and independent variables, which account for the endogenous interaction effect of the dependent variable and the exogenous interaction effects of the independent variables, respectively.

The SDM model extended with spatial specific and time-period specific effects is written as (Elhorst 2014):

$$Y_t = \rho W Y_t + \alpha \iota_N + X_t \beta + W X_t \theta + \mu + \xi_t \iota_N + \varepsilon_t$$
$$\varepsilon_{it} \backsim N(0, \sigma^2) \quad i.i.d$$
$$\mu = (\mu_1, \dots, \mu_N)^T \tag{1}$$

Where $Y_t$ is an $N \times 1$ vector consisting of one observation on the patent intensity for each province during year $t$ ($i$ = 1, …, $N$; $t$ = 1, ..., $T$); $W$ is a non-negative and non-stochastic $N \times N$ spatial weights matrix whose generic element $w_{ij}$ is set to 1 if the generic observations i and j share an administrative boundary of non-zero length or have borders that touch the first-order neighbours and to 0 otherwise; $\iota_N$ is the $N \times 1$ vector of ones associated with the constant term α; $\beta$ is the $K \times 1$ vector of parameters for exogenous covariates $X_t$ $(N \times K)$; $\rho$ and $\theta$ are, respectively, the scalar for the endogenous interaction effect ($WY$) referred to as spatial autoregressive coefficient and the $K \times 1$ vector of parameters for the exogenous interaction effects in the regressors $WX$; $\mu$ is the vector of spatial specific fixed effects, while $\xi_t$ denotes the time-period specific fixed effects; $\varepsilon_t$ is the vector of independently and identically distributed error terms with zero mean and constant variance.

Spatial-specific effects control for all time-invariant variables whose omission could bias the estimates in a typical cross-sectional study, while time-period specific effects control for all time-specific effects whose omission could bias the estimates in a typical time–series study (Baltagi, 2008).

The presence of spatially lagged variables does not allow the parameters to be interpreted as in the usual framework of the linear model since the change in a covariate in a given spatial unit affects the dependent variable in that unit itself (direct effects) and the dependent variables in other units (spillover effects). Therefore, the direct and indirect effects of the SDM model associated with each independent variable will be presented and discussed, rather than the model parameter estimates.

The average direct effect associated with the generic $k_{th}$ explanatory variable is given by the sum of the diagonal elements of the partial derivatives' matrix (2) of the expected value of the dependent variable, $E(Y_t)$, for that variable. Conversely, the sum of the off-diagonal elements represents the average spillover effect (LeSage and Pace, 2009).

$$\left[\frac{\partial E(Y_t)}{\partial X_{1k_t}} \dots \frac{\partial E(Y_t)}{\partial X_{Nk_t}}\right] = (I_N - \rho W)^{-1}(I_N \beta_k + W \theta_k) \tag{2}$$

The space-time analysis of innovation activities in the 107 Italian provinces is performed on panel data obtained from official sources over the time span from 2010 to 2017. Innovation output is proxied by patent intensity, defined as the number of patent applications to the European Patent Office (EPO) per million population. Despite some limitations associated with patent intensity as an indicator of innovation production, it is considered to provide a reliable measure of innovative territorial activity (Acs et al., 2002) and a direct measure of innovation output (Ascani et al., 2020). The potential explanatory variables of innovation activities (Table 1) are taken from Istat (Italian Institute of Statistics) and SIEPI (Italian Society of Economics and Industrial Policy)

Table 1: Explanatory variables of innovation activities

| Label | Indicator | Description | Source |
|---|---|---|---|
| EDU | Higher education rate | Share of the population (aged 24-39) with tertiary-level education. | Istat |
| IQI | Institutional Quality Index | A measure of Italian institutional quality that ranges between 0 and 1. The closer the IQI is to 1, the higher the quality of the local institution. It is composed of five dimensions: Voice and accountability; Government effectiveness; Regulatory quality; Rule of law; Corruption. | SIEPI |
| R&D | Research and development. | Expenditure on R&D expressed as a percentage of gross domestic product (GDP). | Istat |
| GDP | Change in gross domestic product | Relative one-year-change in provincial GDP | Istat |
| NEET | Youth not in | Share of the young population (aged 15- 29) who are | Istat |

DRAFT

| | | | |
|---|---|---|---|
| | employment, education, or training | not in education, employment, or training to the population of the corresponding age group 15 to 29. | |
| REG | Registered companies | Share of registered enterprises minus terminated enterprises in the total number of enterprises in the commercial register. | Istat |
| SPE | Specialisation Index | An index of dissimilarity that takes the value 0 if the local unit has the same employment composition by sectors as its neighbouring units (no specialisation) and, on the contrary, the value 1 if all the employees of the local unit are concentrated in a single sector, unlike its neighbourhood. | Own elaboration of Istat data |

## 4. Results and Conclusions

The assumed global spillover effects of innovation are confirmed by the results of the locally robust panel Lagrange Multiplier tests (Elhorst, 2014; Anselin et al., 1996) for spatial lag ($LM_{lag}$=12.836, p-value < 0.001) or spatial error ($LM_{err}$=0.006, p-value=0.940) dependence. The advisability of using the SDM model with fixed effects rather than random effects is tested through Hausman's robust test (Hausman, 1978) of spatial autocorrelation (Mutl and Pfaffermayr, 2011). Given the statistic test of 108.42 (p-value=0.000), the random effects model is inconsistent; thus, a fixed effect model was chosen.

Table 2 shows the estimated direct and indirect effects of the SDM model extended with spatial specific and time-period specific effects (eq. 1). According to a stepwise procedure, the covariates presented in Table 1 were properly selected, also checking for multicollinearity. The significant spatial autoregressive coefficient (ρ) confirms that interaction between local units contributes to shaping their innovative profiles. Therefore, spatial analysis techniques best modelling the innovation patterns since they capture ties within provincial networks that promote the transfer of knowledge, expertise, and experience (Baycan et al., 2017). All explanatory variables exert significant direct effects on local innovation and, most of them, potentially influence the innovativeness of all other provinces due to significant global spillover effects.

Table 2: Direct and Spillover Effects for SDM

| Variable | Direct effects | Spillover Effects |
|---|---|---|
| EDU | 0.2597 *** | 0.4627*** |
| NEET | -0.8108*** | - 0.0771 |
| R&D | 0.1520*** | -0.0910 |
| IQI | 0.3809*** | 0.6831** |
| GDP | 1.9520*** | -1.7940*** |
| REG | 0.1542*** | 0.2942*** |
| SPE | -0.9548* | -1.6773* |
| $\rho$ | -0.2321*** | |
| N | 107 | |
| LogLik | -902.40 | |

The education rate (EDU), a proxy for the level of human capital, positively influences provincial innovativeness both directly and indirectly. Higher levels of education mean greater skills and knowledge that can be leveraged to generate innovation. Furthermore, education is considered to be both an "engine of innovation" and a "catalyst for sustainability" (Cai et al., 2020) due to the growing focus of universities and educational institutions on providing sustainability-oriented training. Supporting this result are those showing that better innovative performances are in the provinces characterised by a lower share of NEETs and a higher R&D expenditure. It is worth noting that better education systems can facilitate local development by widening access to education, human capital development and highly skilled labour (Raileanu Szeles and Simionescu, 2022). NEETs' lower access to the necessary knowledge and tools to foster innovation clearly inhibits innovative processes in contexts with a higher share of NEETs in the population. Additionally, researchers have shown that companies led by qualified

DRAFT

administrators holding a degree or post-graduate degree invest more in R&D (Wang et al., 2019), thereby increasing the percentage of patentable innovations (Gallié and Legros, 2012).

As regards the links between institutions and innovation, IQI positively affects provincial innovativeness, meaning that an increase in institutional quality leads, on average, to an increase in patent applications. Stable institutions and effective government regulations can profoundly influence and foster innovation activities. Conversely, political uncertainty, deficient rule of law, high corruption, and weak regulatory quality strongly undermine innovation probability and intensity (Bhattacharya et al., 2017; Rodríguez-Pose and Zhang, 2020). The spillover effects of IQI also significantly affect innovation patterns. That is, better local institutions in neighbouring areas foster innovation intensity in the area of interest. Therefore, stable local governments' cooperation and mutual influence can form a competitive environment capable of experimenting with sustainable and lasting innovation patterns (He and Tian, 2020).

The change in GDP expresses the economic growth of an area and as expected, positively affects the local propensity to innovate. It is assumed that wealthier provinces have more resources and economic incentives to invest in innovation to trigger innovation-oriented virtuous mechanisms (Gössling and Rutten, 2007). While the innovation propensity of a province is directly affected by its GDP, it is negatively impacted by the GDP of its neighbouring provinces. In light of this result, it appears that economic growth within one local unit may attract highly-skilled employees and draw resources, and potential investments away from neighbouring units, thereby hindering their ability to innovate. Entrepreneurs can recognise opportunities in more prosperous areas and move resources there that can support economic development and promote innovation (Feldman, 2014).

The net share of registered companies in each province (REG) is a driver of innovation both in the province of reference and the surrounding provinces. The proximity of a larger number of firms within the same province can produce positive externalities that facilitate the diffusion of knowledge and the development of local innovative activities (Glaeser et al., 1992). In particular, the specialisation index (SPE) has a negative impact on patenting propensity both directly and indirectly. Therefore, the more provinces have a different employment composition per sector than their neighbours (no specialisation - direct effect), and the more their neighbouring provinces have a different employment composition from each other (no specialisation - spillover effect), the more local innovativeness increases. Although the debate on whether a specialised or diversified productive structure fosters innovation remains open and requires further empirical evidence, the analysis shows that a diversified production structure within the same province can stimulate territorial innovation capacity. Therefore, a diversified environment promoting Jacobs' externalities can stimulate patenting activity as the output of innovation activities.

## References


[1] Acs, Z. J., Anselin, L., Varga, A.: Patents and innovation counts as measures of regional production of new knowledge. *Research policy*, *31*(7), 1069--1085 (2002).

[2] Andersen, M. M.: An innovation system approach to eco-innovation-Aligning policy rationales. In *The greening of policies-interlinkages and policy integration conference* (pp. 1-28) (2004).

[3] Anselin, L., Bera, A. K., Florax, R., Yoon, M. J.: Simple diagnostic tests for spatial dependence. *Regional science and urban economics*, *26*(1), 77-104 (1996).

[4] Aronica, M., Fazio, G., Piacentino, D.: A micro-founded approach to regional innovation in Italy. *Technological Forecasting and Social Change*, *176*, 121494 (2022).

[5] Arrow, K. J.: The economic implications of learning by doing. *The review of economic studies*, *29*(3), 155-173 (1962).

[6] Ascani, A., Balland, P. A., Morrison, A.: Heterogeneous foreign direct investment and local innovation in Italian Provinces. *Structural Change and Economic Dynamics*, *53*, 388-401 (2020).

[7] Baltagi, B. H.: *Econometric analysis of panel data* (Vol. 4). Chichester: Wiley (2008).

[8] Baycan, T., Nijkamp, P., Stough, R..: Spatial spillovers revisited: Innovation, human capital and local dynamics. *International Journal of Urban and Regional Research*, *41*(6), 962-975 (2017).

[9] Bocken, N. M. P., Allwood, J. M., Willey, A. R., King, J. M. H.: Development of an eco-ideation tool to identify stepwise greenhouse gas emissions reduction options for consumer goods. *Journal of Cleaner Production*, *19*(12), 1279-1287 (2011).

[10] Cai, Y., Ma, J., Chen, Q.: Higher education in innovation ecosystems. *Sustainability*, *12*(11), 4376 (2020).

[11] Cabrer-Borras, B., Serrano-Domingo, G.: Innovation and R&D spillover effects in Spanish regions: A spatial approach. *Research Policy*, *36*(9), 1357-1371 (2007).


DRAFT

[12] Capello, R., Lenzi, C.: Regional innovation patterns from an evolutionary perspective. *Regional Studies*, *52*(2), 159-171 (2018).
[13] Capello, R., Nijkamp, P.: Introduction: regional growth and development theories in the twenty-first century–recent theoretical advances and future challenges. In *Handbook of regional growth and development theories*. Edward Elgar Publishing (2009).
[14] De Groot, H. L., Poot, J., Smit, M. J.: Which agglomeration externalities matter most and why?. *Journal of Economic Surveys*, *30*(4), 756-782 (2016).
[15] Elhorst, J. P.: *Spatial econometrics: from cross-sectional data to spatial panels* (Vol. 479, p. 480). Heidelberg: Springer (2014).
[16] Elhorst, J. P., Halleck Vega, S.: The SLX model: extensions and the sensitivity of spatial spillovers to W. Papeles de Economía Española, 152, 34-50 (2017).
[17] Feldman, M. P.: The character of innovative places: entrepreneurial strategy, economic development, and prosperity. *Small Business Economics*, *43*, 9-20 (2014).
[18] Firmino Costa da Silva, D., Elhorst, J. P., Silveira Neto, R. D. M.: Urban and rural population growth in a spatial panel of municipalities. *Regional Studies*, *51*(6), 894-908 (2017).
[19] Furková, A.: Spatial spillovers and European Union regional innovation activities. *Central European Journal of Operations Research*, *27*, 815-834 (2019).
[20] Gallié, E. P., Legros, D.: Firms' human capital, R&D and innovation: a study on French firms. *Empirical Economics*, *43*, 581-596 (2012).
[21] Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., Shleifer, A.: Growth in cities. *Journal of political economy*, *100*(6), 1126-1152 (1992).
[22] Gössling, T., Rutten, R: Innovation in regions. *European planning studies*, *15*(2), 253-270 (2007).
[23] Hausman, J. A.: Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251-1271 (1978).
[24] He, J., Tian, X.: Institutions and innovation. *Annual Review of Financial Economics*, *12*, 377-398 (2020).
[25] Huggins, R., Thompson, P.: Entrepreneurship, innovation and regional growth: a network theory. *Small business economics*, *45*, 103-128 (2015).
[26] Jacobs J.: The economy of cities. Random House, New York (1969).
[27] LeSage, J., Pace, R. K.: *Introduction to spatial econometrics*. Chapman and Hall/CRC (2009).
[28] Marshall, A.: Principles of economics Macmillan. *London (8th ed. Published in 1920)* (1890).
[29] Mutl, J., Pfaffermayr, M.: The Hausman test in a Cliff and Ord panel model. *The Econometrics Journal*, *14*(1), 48-76 (2011).
[30] Paci, R., Marrocu, E., Usai, S.: The complementary effects of proximity dimensions on knowledge spillovers. *Spatial Economic Analysis*, *9*(1), 9-30 (2014).
[31] Raileanu Szeles, M., Simionescu, M.: Improving the school-to-work transition for young people by closing the digital divide: evidence from the EU regions. *International Journal of Manpower*, *43*(7), 1540-1555 (2022).
[32] Rodríguez-Pose, A., Zhang, M.: The cost of weak institutions for innovation in China. *Technological Forecasting and Social Change*, *153*, 119937 (2020).
[33] Romer, P. M.: Increasing returns and long-run growth. *Journal of political economy*, *94*(5), 1002-1037 (1986).
[34] Silvestre, B. S.: Sustainable supply chain management in emerging economies: Environmental turbulence, institutional voids and sustainability trajectories. *International Journal of Production Economics*, *167*, 156-169 (2015a).
[35] Silvestre, B. S.: A hard nut to crack! Implementing supply chain sustainability in an emerging economy. *Journal of cleaner production*, *96*, 171-181 (2015b).
[36] Silvestre, B. S., Ţîrcă, D. M.: Innovations for sustainable development: Moving toward a sustainable future. *Journal of cleaner production*, *208*, 325-332 (2019).
[37] Wang, C., Yang, J., Cheng, Z., Ni, C.: Postgraduate education of board members and R&D investment—Evidence from China. Sustainability, 11(22), 6524 (2019).

DRAFT

# Revealing the dynamic relations between traffic and crowding using big data from mobile phone network

Selene Perazzini[a], Rodolfo Metulini[b], and Maurizio Carpita[a]

[a]DMS Statlab, Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, Brescia; selene.perazzini@unibs.it, maurizio.carpita@unibs.it
[b]Department of Economics, University of Bergamo, Via Caniana, 2, Bergamo; rodolfo.metulini@unibg.it

DRAFT

**Abstract**

In this work, we use three sources of mobile phone data to monitor traffic between three neighboring small areas - three "Aree di Censimento" (ACE) - in the Province of Brescia. Two indicators aimed at capturing crowding and traffic intensity are defined and their relationship is analyzed. Then, their impact on traffic flows is investigated. To this scope, traffic flows between pairs of ACEs are estimated by means of a vector autoregressive model with crowding and traffic intensity indicators as regressors. Moreover, seasonal components are included in the model as exogenous variables modeled as dynamic harmonic regressions. We find that our model always satisfactorily captures traffic flows, although the effect of the seasonal components varies considerably among the pairs.

***Keywords:*** Vector AutoRegressive model, Dynamic harmonic regression, Traffic monitoring, Small area estimation, Miminization Drive Test.

## 1. Introduction

The Italian Government has recently set the goal of implementing an advanced digital system to monitor traffic in Italy as part of the National Recovery and Resilience Plan, which is part of the Next Generation EU Programme. Achieving this goal is of main relevance, as it would also constitute an important step toward the achievement of some of the United Nations Sustainable Development Goals (in particular goal 11, "Sustainable cities and communities"). To this aim, mobile phone data might play a key role. Indeed, they allow for capturing both the temporal and spatial dynamics that characterize social phenomena in urban areas. For this reason, they are increasingly adopted in the statistical literature for the analysis of people's presences and movements (e.g., 1; 2; 3; 4; 5; 6; 7; 11).

In this paper, we explore the potential of mobile phone data in traffic monitoring by analyzing their ability in forecasting flows between small areas. We consider three "Aree di CEnsimento" (ACE) (which correspond to small municipalities) in the Province of Brescia, namely Gussago, Castegnato, and Rodengo Saiano (see the left map of Figure 1), and estimate the traffic flows between pairs of them. To this aim, we refer to the Vector AutoRegressive model with eXogenous variables (VARX) with complex seasonality by (8), which we extend by introducing regressors defined over different types of mobile phone data. More in detail, we use "Origin-Destination" (OD) data to represent the number of people moving between the ACEs, the "Mobile Phone Density" (MPD) data to capture crowdings in the ACEs, and the "Minimization of Drive Test" (MDT) data to capture the level of traffic intensity. As we will

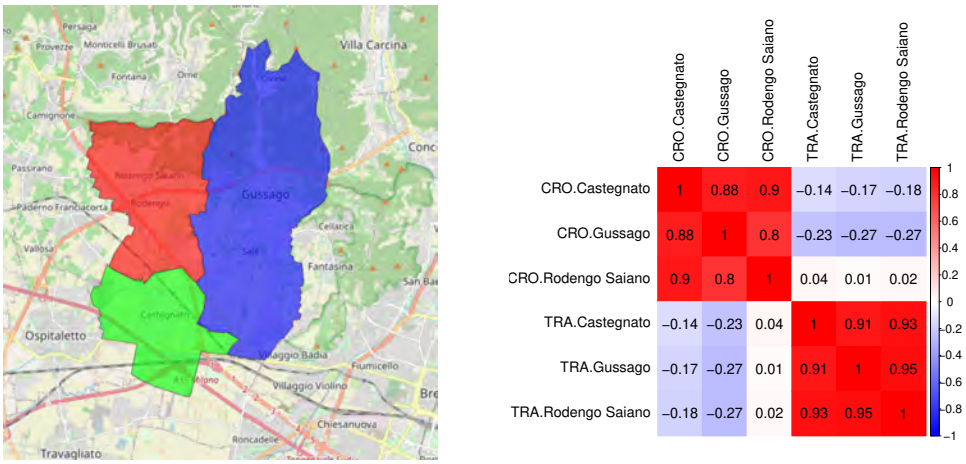



Figure 1: Left: Map of the three ACEs: Rodengo Saiano (red), Castegnato (green), and Gussago (blue). Right: Correlation plot of the indicators CRO and TRA.

discuss in detail in the next Section, the three sources of data present different temporal and spatial characteristics, which allow us to capture different important features of the phenomena, but also require proper preprocessing to be combined. Indeed, this constitutes a further contribution of this work to the literature.

The paper organizes as follows: Section 2 presents the data; Section 3 defines the model; Section 4 shows the results; Section 5 concludes.

DRAFT

## 2. Data

Three sources of data have been used for the analysis: the OD, MPD, and MDT data. All the data refer to users subscribed to TIM but have been provided to the DMS Statlab in the context of different projects. Specifically, the MPD data was provided by the Municipality of Brescia in the context of a territorial monitoring project between 2014 and 2016, while the OD and MDT data have been provided by Olivetti S.p.A. (www.olivetti.com) with the support of FasterNet S.r.l. (www.fasternet.it) for the MoSoRe Project 2020-2022. For this reason, the three datasets cover different periods of time: the MPD data refer to April 2014 - August 2016, the OD to September 2020 - August 2021, and the MDT to 5 days of November 2021 (namely Wednesday $10^{th}$, Friday $19^{th}$, Saturday $20^{th}$, Sunday $21^{st}$, and Monday $22^{nd}$). To guarantee the comparability of the data, we restrict our attention to traffic flows between October $1^{st}$, 2020 and March $31^{st}$, 2021. Three variables based on mobile phone data have been included in the model: the traffic flows, the crowding indicator, and the traffic intensity indicator. The three variables are described in detail in the next paragraphs.

**Traffic flows ($y_{ijt}$)** The flows of people $y_{ijt}$ are captured by the OD data, which report the number of SIM cards moving from ACE $i$ to ACE $j$ during the hour $t$. Unusual traffic flows characterize holidays. Therefore, according to (8), flows corresponding to holidays have been replaced with the corresponding values observed seven days before (the previous same day of the week). Substituted days are November $1^{st}$, December $8^{th}$, $25^{th}$, and $26^{th}$, and January $1^{st}$ and $6^{th}$.

**Crowding indicator ($CRO_{id}$)** The crowding indicator $CRO_{id}$ represents the average number of individuals in ACE $i$ during the $d$-th hour of the day of the week (i. e., Monday at hour 00-01 AM, Monday at 01-02 AM, ...). For the construction of this indicator, the MPD data have been used. Specifically, the MPD database reports the average number of mobile phone SIM cards in a $150 \times 150$ squared meters cell of a pixel grid during a 15-minute interval. Following the approach in (9), we restricted our attention to human SIM cards (i.e., machine SIM cards have been excluded) and to observations in November to

guarantee comparability with the MDT data. Then, for each 15-minute interval $t'$ the number of SIMs ($MPD$) in the $i$-th ACE has been computed by overlaying the pixel grid to a map of the administrative boundaries of the Province of Brescia and aggregating the portions of cells of the grid overlapping $i$. At last, observations have been averaged among hours and days of the week:

$$CRO_{id} = \sum_{t' \in d} \frac{1}{N_{t'}} \sum_{k} MPD_{kt'} \cdot \frac{Area(ACE_i \cap Cell_k)}{Area(Cell_k)} \tag{1}$$

where $N_{t'}$ is the number of 15-minute intervals of November 2020 corresponding to the $d$-th hour of the day of the week.

**Traffic intensity indicator ($TRA_{id}$)** The traffic intensity indicator $TRA_{id}$ has been defined using the MDT data, which collects the number of signals (i.e., phone calls, text messages, internet browsing, or technical operations on the network) transmitted over the 3G/4G mobile network from/to terminal devices with GPS enabled. The signals are collected in 15-minute intervals and geo-referenced on a grid of pixels measuring 10 meters per side. MDT data only represents a sample of users and a mobile phone can produce multiple signals in 15 minutes. However, the high accuracy in georeferencing allows for estimation at the small area level. Following (9), we overlay the grid of pixels with a polygon-based street map and identify the cells of the grid that correspond to streets. For each ACE $i$ and time interval $t'$, we count the number of street cells from which at least one signal originated during a 15-minute interval. Then, values have been averaged in intervals of length 1 hour. Therefore, we obtained 24 observations per each of the 5 observed days of the week:

$$TRA_{id} = \frac{1}{4} \sum_{t' \in d} (\text{Number of Street Cells with MDT signals}_{it'} \in ACE_i) \tag{2}$$

Since the data at our disposal do not include Tuesday and Thursday, values for each hour of the two days are set equal to the corresponding average of Monday, Wednesday, and Friday.

The relationship between the indicators $CRO_{id}$ and $TRA_{id}$ has been investigated. As shown in the right plot of Figure 1, the two indicators do not appear correlated. However, high positive values of the Pearson correlation emerge between the values of each indicator observed in the three ACEs. This evidence is not particularly surprising, as traffic conditions in neighboring ACEs are likely to be related and so do crowdings.

DRAFT

## 3. The model

To model traffic flows between pairs of ACEs, we refer to the vector autoregressive model with dynamic harmonic components defined by (8), to which we apply two major modifications. First, (8) focuses on the analysis of one ACE and models its inflows from and outflows to other 38 ACEs as well as its internal flows. Since we are interested in flows between ACEs, we do not model the internal flows. Second, the original work only includes lags of traffic flows and some dummies as regressors; here, we also introduce the indicators $CRO$ and $TRA$.

The estimated VARX model is:

$$\boldsymbol{y}_t = \boldsymbol{\nu} + \sum_{h} \boldsymbol{A}_h \boldsymbol{y}_{t-h} + \boldsymbol{B}\boldsymbol{x}_t + \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{\epsilon}_t \tag{3}$$

where $\boldsymbol{\nu}$ is a $2 \times 1$ vector of constants, $\boldsymbol{y}_t$ is a $2 \times 1$ vector containing the flows between and within ACEs $i$ and $j$, namely $y_{ijt}$, $y_{jit}$, and $\boldsymbol{\epsilon}_t$ is a $2 \times 1$ vector of the error terms at time $t$. $h$ indicates a set of lags of $y_{ijt}$ and $y_{jit}$. For this parameter, we refer to the analyzes in (8) and include 3 daily and 4 weekly lags, therefore we have $h = (24, 48, 72, 168, 336, 504, 672)$. For each value of $h$ there is a $2 \times 2$ matrix of coefficients $\boldsymbol{A}_h$ to be estimated. $\boldsymbol{x}_t$ is a vector of exogenous variables containing the crowding and

| | $R^2$ | |
|---|---|---|
| **Model** | with CRO, TRA | no CRO, TRA |
| Castegnato - Gussago | [0.939, 0.936] | [0.936, 0.933] |
| Rodengo - Gussago | [0.949, 0.948] | [0.942, 0.940] |
| Rodengo - Castegnato | [0.942, 0.941] | [0.939, 0.939] |

Table 1: $R^2$ of the estimated models. Each row represents a pair of ACEs, column "with CRO, TRA" indicates the model in Eq.s (3)-(4), and column "no CRO, TRA" the model in (8). For each model and each pair of ACEs, two values of the $R^2$ are reported (one per each flow in Eq. 3, i.e., $y_{ij}$ and $y_{ji}$).

traffic indicators, 6 daily dummies (from Tuesday to Sunday), and 5 monthly dummies (from November 2020 to March 2021). In order to avoid multicollinearity issues, the average values of the two indicators in the ACEs $i$ and $j$ at time $d$ are considered, namely $\overline{CRO}_{ijd}$ and $\overline{TRA}_{ijd}$. $\boldsymbol{B}$ is the $2 \times 13$ matrix of coefficients associated to $\boldsymbol{x}_t$. At last, $\boldsymbol{z}_t$ is a $l \times 1$ vector of exogenous variables capturing seasonality, and $\boldsymbol{C}$ is the corresponding $2 \times l$ matrix of coefficients. The elements of the vector $\boldsymbol{C}\boldsymbol{z}_t$ are modeled using the dynamic harmonic regression components:

$$\beta_0^{(r)} + \sum_{k=1}^{K} \left[ \alpha_k^{(r)} sin\left(\frac{2\pi kt}{m}\right) + \gamma_k^{(r)} cos\left(\frac{2\pi kt}{m}\right) \right], \qquad r = 1, 2 \tag{4}$$

where $\beta_0$ is a constant term, $K$ is the optimal numbers of Fourier bases, $\alpha_k$ and $\gamma_k$ are regression coefficients to be estimated, and $m$ is the seasonal period. Note that $\beta_0$, $\alpha_k$, and $\gamma_k$ are allowed to assume different values in the two flows $y_{ijt}$ and $y_{jit}$. Since we account for both daily and weekly seasonal components, we include two sets of Fourier basis, one considering $m = 24$ and one with $m = 168$.

At last, all the dependent variables in the model excluding dummies and Fourier bases have been transformed into z-scores such that we can compare the effect of the variables on different pairs of ACEs.

DRAFT

## 4. Results

The model defined in Eq.s (3)-(4) has been estimated on pairs of ACEs, namely: Gussago-Rondengo Saiano, Castegnato-Gussago, and Castegnato-Rodengo Saiano. The performance of the here proposed model has been evaluated with respect to the original model in (8) where regressors $\overline{CRO}$ and $\overline{TRA}$ were not included. In this respect, Table 1 shows the $R^2$ for the two equations of the VARX model (Eq. 3) (i.e., flows from the $i$-th to the $j$-th ACE and flows from $j$ to $i$). It could be noticed that, for each combination of ACEs, the values of the $R^2$ just slightly increase when $\overline{CRO}$ and $\overline{TRA}$ are included. Although the improvement in the $R^2$ is not so high, the Akaike and Bayesian Information Criteria decrease when $\overline{CRO}$ and $\overline{TRA}$ are introduced, as shown in Table 2. Therefore, we find evidence that our model should be preferred, and that accounting for crowding and traffic intensity improves the estimates.

The estimated parameters are shown in Figure 2. In general, we find that the monthly and daily dummies are among the main determinants of traffic flows $y_{ijt}$. Along with the dummies, highly significative coefficients are also found for the indicator $\overline{TRA}$. The crowding indicator $\overline{CRO}$ also generally appears significant, although it plays a minor negative effect. As far as these variables are concerned, the estimated coefficients appear stable among the pairs of ACEs. By contrast, major differences among the three cases emerge in the lagged variables $\boldsymbol{y}_{t-h}$. Indeed, each pair appears affected by different lags $h$. However, many lagged variables are associated with coefficients close to 0 or large confidence intervals and are therefore not significant. This is particularly the case of the flows between Gussago and Rodengo Saiano.

| | AIC | | BIC | |
|---|---|---|---|---|
| **Model** | with CRO, TRA | no CRO, TRA | with CRO, TRA | no CRO, TRA |
| Castegnato - Gussago | -8.464 | - 8.421 | -8.306 | - 8.269 |
| Rodengo - Gussago | -8.443 | -8.301 | -8.285 | -8.150 |
| Rodengo - Castegnato | -7.797 | -7.736 | -7.640 | -7.584 |

Table 2: Comparison of the performance of the model in Eq.s (3)-(4) (i.e., with regressors $\overline{CRO}$ and $\overline{TRA}$) and the model in (8) (i.e., without $\overline{CRO}$ and $\overline{TRA}$) estimated on the three pairs of ACEs. The second and third columns report the Akaike Information Criterion (AIC), and the last two columns the Bayesian Information Criterion (BIC).

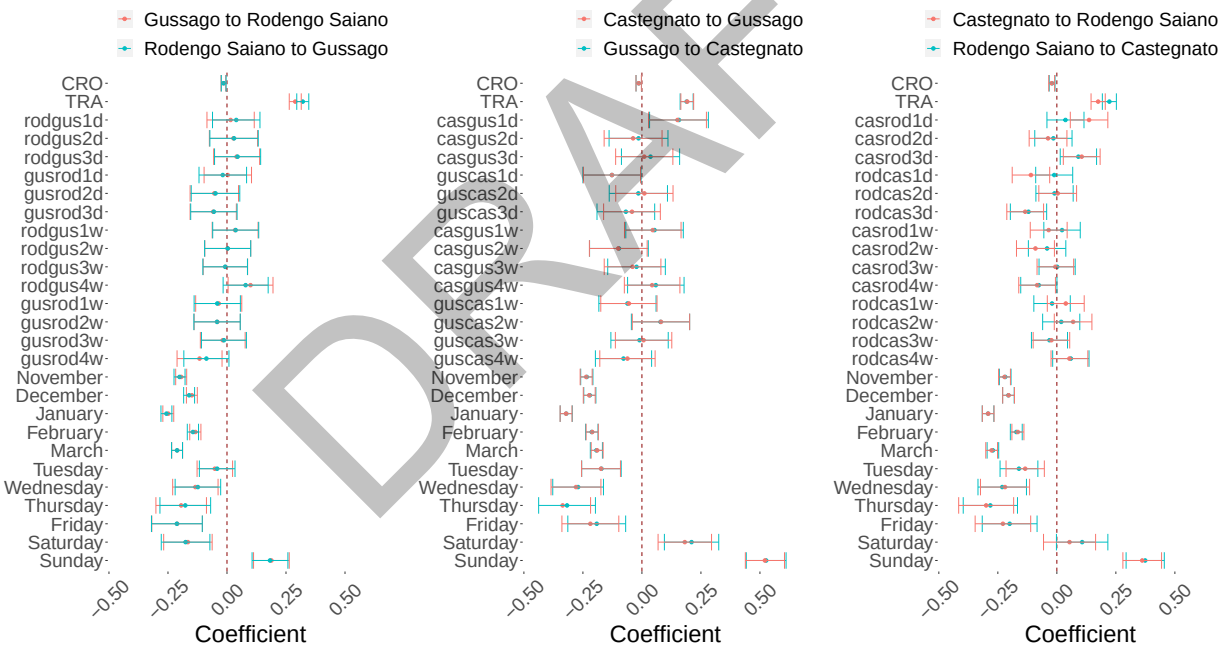



Figure 2: Estimated coefficients of the model in Eq.s (3)-(4). Each column represents one of the analyzed pairs of ACEs. For each pair, the variables included in the model are reported in this order: a $\overline{CRO}$ indicator, a $\overline{TRA}$ indicator, 14 lagged flows, 5 monthly dummies, and 6 daily dummies. Lagged variables are indicated by the three initial letters of the ACEs of origin and destination (e.g., rodgus = flows from Rodengo Saiano to Gussago) and two digits representing the lag (1d corresponds to $h = 24$, 2d to $h = 48$, 3d to $h = 72$, 1w to $h = 168$, 2w to $h = 336$, 3w to $h = 504$, 4w to $h = 672$). For each variable, the point represents the estimated coefficient, and the line reports the associated confidence interval.

DRAFT

## 5. Conclusion

A VARX model with complex seasonality and regressors representing crowding and traffic has been proposed for traffic flow estimation based on mobile phone data. The model has been applied to three ACEs in the Province of Brescia, with satisfactory results. In particular, we found that the two mobile phone data indicators of crowding and traffic intensity are among the main determinants of traffic flows. The seasonal component also appears to affect the flows, and its effect is captured by both monthly and daily dummies and the lagged flows. We find similar results for the three pairs of ACEs, although each pair appears affected by different lags of the flows. As a future development, the model forecasting ability will be analyzed using a blocked k-folds cross-validation along with the mean absolute percentage error and the hit rate.

**Acknowledgments** This contribution has been developed for the European Union (EU) and Italian Ministry for Universities and Research (MUR), National Recovery and Resilience Plan (NRRP), within the project "Sustainable Mobility Center (MOST)" 2022-2026, CUP D83C22000690001, Spoke N 7 "CCAM, Connected networks and Smart Infrastructures".

DRAFT

## References

[1] Bibri, S.E., and Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable cities and society. **31**, 183–212 (2017)

[2] Carpita, M., Manisera, M., and Zuccolotto, P.: Mobile Phone Data to Monitor the Impact of Social and Cultural Events of Brescia. In: Lombardo R., Camminatiello I., Simonacci V. eds. IES 2022: Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment, Book of Short Papers of the 10th Scientific Conference of the SVQS, 575-581. PKE Press, Milano (2022)

[3] Carpita, M., and Simonetto, A.: Big data to monitor big social events: Analysing the mobile phone signals in the Brescia smart city. Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation. **5 (1)**, 31–41 (2014)

[4] Curci, F., Kërçuku, A., Zanfi, F., Novak, C. et al.: Permanent and seasonal human presence in the coastal settlements of Lecce. An analysis using mobile phone tracking data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 57–71 (2022)

[5] Manfredini, F., Lanza, G., Curci, F., et al.: Mobile phone traffic data for territorial research. Opportunities and challenges for urban sensing and territorial fragilities analysis. TeMA-Journal of Land Use, Mobility and Environment. **2**, 9–23 (2022)

[6] Mariotti, I., Giavarini, V., Rossi, F., and Akhavan, M.: Exploring the "15-Minute City" and near working in Milan using mobile phone data. TeMA-Journal of Land Use, Mobility and Environment. **2**, 39–56 (2022)

[7] Metulini, R., and Carpita, M.: A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Social Indicators Research. **156 (2)**, 761–781 (2021)

[8] Metulini, R., and Carpita, M.: Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making. Annals of Operations Research. 1–26, online first (2023)

[9] Perazzini, S., Metulini, R., Carpita, M. Statistical indicators based on mobile phone and street maps data for risk management in small urban areas. Submitted to journal.

[10] Pucci, P., Gargiulo, C., Manfredini, F., Carpentieri, G., et al.: Mobile phone data for exploring spatio-temporal transformations in contemporary territories. TeMA-Journal of Land Use, Mobility and Environment. **2**, 6–12 (2022)

[11] Tettamanti, T., and Varga, I.: Mobile phone location area based traffic flow estimation in urban road traffic. Advances in Civil and Environmental Engineering. **1 (1)**, 1–15 (2014)

# SMaC: Spatial Matrix Completion method

Giulio Grossi[a], Alessandra Mattei[a], and Georgia Papadogeorgou[b]

[a]Viale Morgagni 59, 50134, Florence, Italy; giulio.grossi@unifi.it,
alessandra.mattei@unifi.it,
[b]102 Griffin-Floyd Hall, 32611, Gainesville, Florida; gpapadogeorgou@ufl.edu,

**Abstract**

Synthetic control methods are commonly used in panel data settings to evaluate the effect of an intervention. In many of these cases, the treated and control time series correspond to spatial areas such as regions or neighborhoods. Synthetic control methods can be used to evaluate the effect that the treatment had in the treated area, but it is often unclear how far the treatment's effect propagates, as this approach ignores the spatial structure of the data, and can lead to efficiency loss in spatial settings. We propose to deal with these issues by developing a Bayesian spatial matrix completion framework that allows us to predict the missing potential outcomes in the different areas around the intervention point while accounting for the spatial structure of the data. Specifically, the missing time series in the absence of treatment for the treated areas of all sizes are imputed using a weighted average of control time series, where the weights are assumed to vary smoothly over space according to a Gaussian process.

***Keywords:*** Causal inference, Spatial Econometrics, Synthetic Control Method, Gaussian Process, Potential Outcomes

DRAFT

## 1. Introduction

The Synthetic Control method (SCM hereinafter) is a widespread methodology to estimate causal effects in presence of a single treated unit and many control units, observed over time (2). With this method, the impact of an intervention is evaluated as the difference between the observed value of some primary outcome and its counterfactual value, imputed by using a weighted average of control units.

Evidence of interest in SCM is the flurry of methodological developments. Recently, the exploration of SCM alternatives heads toward Bayesian regression models. (6), (5), (7) and (8) use Bayesian methods for causal effects estimation, illustrating a simple and effective proposal for inference in SCM-like settings. Finally, recent work from (3) investigates the use of multitask Gaussian Processes for weights estimations. In Many fields where SCM is commonly used study outcomes which are measured in spatial areas such as municipalities, states or regions. (1) suggests these as the specific framework of application for SCM-like methods. In such contexts, it is common to see treatment assigned to a single area, and the focus being to estimate the treatment effect on this treated unit. Usually, scholars consider no second-round effects from the treatment, neither in terms of spillovers nor in terms of effect propagation. However, no previous work has addressed spatial treatment effect propagation explicitly within the scope of SCM. In practice, researchers often evaluate the extent to which treatment effects propagate through space by applying SCM to areas of different sizes around the treated location. In this work, we propose a Bayesian estimator for missing potential outcomes in presence of spatial correlation among treated units. We exploit a Gaussian process prior for the vertical regression coefficients that take into

account spatial correlation, encouraging regression coefficients across similar areas to be similar. We aim to exploit this spatial information to estimate counterfactual quantities that are still unbiased, but have improved properties in terms of mean bias and mean square error of the point estimate with respect to the separated SCM or vertical regression methods. We refer to this method as *Spatial Matrix Completion* or SMaC. Our motivating application is the impact evaluation arising from the construction of the first line of the Florentine tramway network. In particular, we wish to assess the infrastructural impact on the commercial vitality of the treated neighbourhood, measured as the number of stores located within some distance $d$ from a tramway stop.

## 2. Causal Framework

Consider a space $\Omega$ that can be partitioned into N areas, indexed in $i \in \mathbf{N} = \{1, \dots, i, \dots, N\}$, such that $\bigcup_{i=1}^{N} \Omega^i = \Omega$ and $\Omega^i \cap \Omega^j = \emptyset$ for each couple $i, j \in \{1, \dots, N\}$. In our study, we consider the natural partition of our sample space into the clusters representing the Florentine neighbourhoods. We observe treatment arising from some specific locations $\boldsymbol{\omega}_1 \in \Omega_1$. We can consider treatment locations as a point treatment (e.g.: pollution created by a power plant), a linear treatment or even a polygonal treatment. Let be $\boldsymbol{\omega}^1$ the set of treatment locations, in our application we consider $\boldsymbol{\omega}^1$ as the tramway stops located in the treated area. We also consider sets of locations $\boldsymbol{\omega}^i, i \in \{2, \dots, N\}$ as sets of locations located in neighbourhoods located far away from the tramway line, in streets similar to the one that receives the treatment. We define our observation units as the areas around the treatment sites $\boldsymbol{\omega}$. Therefore, for each neighbourhood $i$ we construct a set of buffers areas $\boldsymbol{A}_i = \{A_i^1, \dots, A_i^d, \dots A_i^H\}$ around the treatment locations $\boldsymbol{\omega}_i$, using the vector of distances $\mathbf{D} = (d_1, \dots, d_h, \dots, d_H)$ representing the distance of the $h-$th area from the treatment site. We sort units and distances such that $d_{h+1} \geq d_h \quad \forall h \in (1, 2, \dots, H-1)$. Let also $d$ denote a generic distance between a treated area and the treatment site. We repeatedly observe units over time, so we consider a panel data setting, with $H \times N$ areas observed for $T^0 = (1, \dots, t_0 - 1)$ pre-treatment periods, and $T^1 = (t_0, \dots, T)$ post-treatment periods. Let $Y_{i,t}^d$ be our primary outcome, the number of stores in neighbourhood $i$ within distance $d$ from the tramway stops in each time period $t \in T$. Let be $\boldsymbol{z} = \{z_i^d\}_{i \in N}^{d \in D} \quad z_i^d \in \{0, 1\}$ be a neighbourhood-level treatment for each area $A_i^d$ considered. Thus following, units belonging to the same cluster $i$ can be only treated or not-treated together. We consider two alternative situations for $\boldsymbol{z}$: $\boldsymbol{z}^1$ is the scenario in which each area $A_1^d \in \Omega_1$ receives the treatment, and no one outside. Instead, $\boldsymbol{z}^0$ represents the scenario in which no area results treated, in our scenario the situation in which the tramway was never built in Florence. We consider that areas $\boldsymbol{A}_1 = \{A_1^1, \dots, A_1^d, \dots, A_1^H\} \in \Omega_1$ will receive the treatment starting from the period $t_0$, and remain treated afterwards. In our application, we consider the treated space as the Legnaia neighbourhood in which the tramway stops are located, and $t_0$ =2006. Units located in other part of Florence will be considered non-treated units with $\boldsymbol{A}^0 = \{A_2^1, \dots, A_i^d, \dots, A_N^H\} \notin \Omega_1$. We adopt the potential outcome approach to causal inference (9). Under consistency assumption, for each unit $A_i^d$ in each period $t$ we define the following couple of potential outcomes: $Y_{i,t}^d(1) \equiv Y_{i,t}(\boldsymbol{z}^1) \quad$ as the potential outcome under $\boldsymbol{z}^1$ assignment and $Y_{i,t}^d(0) \equiv Y_{i,t}(\boldsymbol{z}^0) \quad$ as the potential outcome under $\boldsymbol{z}^0$ assignment. In contexts with cluster-level treatment allocation, scholars often invoke a partial interference assumption (10), which rules that interference may occur, but not within groups. Moreover, we exploit the *non-anticipating treatment* assumption to rule out anticipatory effects. We define the causal effect for the treated units as

$$\Delta_{1,t}^d = Y_{1,t}^d(\boldsymbol{z}^1) - Y_{1,t}^d(\boldsymbol{z}^0) \quad \forall t \in T^1, d \in \mathbf{D} \tag{1}$$

For the treated units we observe $Y_{1,t}^d = Y_{1,t}^d(\boldsymbol{z}^1)$ when $t \geq t_0$, so we need to impute the missing quantity $Y_{1,t}^d(\boldsymbol{z}^0)$. From the comparison of effects at different distances from the treatment site, we can get precious insights into the transmission of treatment effects through space. In general, we could expect decaying treatment effects up to some boundary of spatial treatment.

DRAFT

## 3. Estimation of causal effects

One might be interested in understanding the effect that treating the specific location $\omega_i$ had on the area comprised within a specific distance $d \in \mathbf{D}$ versus not treating it. To do this, they can use synthetic control methodology. Specifically, one can find $\beta_{0d} \in \mathbb{R}$ and $\boldsymbol{\beta}_d = (\beta_{2d}, \ldots, \beta_{Nd})^T \in \mathbb{R}^{N-1}$ such that:

$$\begin{pmatrix} \beta_{0d} \\ \boldsymbol{\beta}_d^T \end{pmatrix} = \underset{\boldsymbol{\beta}_d \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ \sum_{t=1}^{t_0-1} \left( Y_{1,t}^d - (1 \; \mathbf{Y}_{i,t}^T)^T \boldsymbol{\beta}_d \right)^2 \right\}. \tag{2}$$

The synthetic control weights and vertical regression coefficients can be calculated separately for different choices of $d \in (d_1, d_H)$. For example, to find the synthetic control weights at distances $d_1 < d_2 < \cdots < d_H$, one could solve the minimization problem in 2 using a constrained optimization procedure, separately for each of these distances. Alternatively, the $H$ different minimization problems could be stacked, and one could solve the combined minimization problem

$$\begin{pmatrix} \beta_{0d_1} \\ \boldsymbol{\beta}_{d_H}^T \\ \beta_{0d_2} \\ \boldsymbol{\beta}_{d_2}^T \\ \vdots \\ \beta_{0d_H} \\ \boldsymbol{\beta}_{d_H}^T \end{pmatrix} = \underset{\boldsymbol{\beta}_0, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_N \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ \sum_{d=1}^{H} \sum_{t=1}^{T_0-1} \left( Y_{1,t}^d - (1 \; \mathbf{Y}_{i.t}^T)^T \boldsymbol{\beta}_i \right)^2 \right\} \tag{3}$$

which will return the exact same solutions as solving 2 separately for each distance. Thus we can obtain weights that minimise the pre-treatment distance between treated unit and the synthetic control, but ignore the spatial structure of data.

Exploting the spatial structure of data, we introduce the Bayesian framework we will use to impute the missing outcome $Y_{1,t}^d(z^0)$. Building on the vertical regression idea (2, (4)) we will propose a matrix completion algorithm that smooths regression coefficient values through contiguous treated units.

In order to consider the spatial structures of the observed treated units, yet being flexible in the parameter estimation, we follow a Bayesian regression approach to solve the optimization problem in 3, using Gaussian processes as priors for control unit coefficients. In our setting, Gaussian processes can be particularly useful, as we could exploit the spatial information in our data for the specification of regression coefficients. We consider $\beta$ varying smoothly through space, in particular, the vector of coefficients $\beta^d$ will be more similar for physically close units. We specify such structure by using a Gaussian process prior for $\beta$ such that

$$\boldsymbol{\beta}_i(\mathbf{D}) \sim \mathcal{GP}\left(\mathbf{0}, \mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})\right)$$

with $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ as a quadratic exponential smoothing kernel with parameter $\rho_i$. Thus, the $(p, q)$ entry of $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ is

$$[\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})]_{pq} = \alpha_i \exp\left\{ -\frac{(d_p - d_q)^2}{2\rho_i^2} \right\}$$

As stated above, other kernel specifications are possible, in order to consider different correlation structures between the treated units. To solve the pooled regression problem in 3, we specify the total vector of coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_N)$ with $\boldsymbol{\beta} \sim \mathcal{MVN}(0, \Sigma)$, where $\Sigma$ is an appropriate block covariance matrix for the pooled coefficient estimation for multiple treated units. With $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ on the

DRAFT

block diagonal. We define the Bayesian regression model as

$$\begin{aligned}
\mathbf{Y} &\sim \mathcal{N}(\beta^T \mathbf{X}, \sigma_y \mathbf{I}) \\
\beta_i &\sim \mathcal{GP}(\mathbf{0}, \mathcal{K}_{\alpha_i, \rho_i}(\mathbf{D})) \\
\alpha_i &\sim \Gamma^{-1}(50, 5) \\
\rho_i &\sim \Gamma^{-1}(5, 5) \\
\sigma_y &\sim \Gamma^{-1}(5, 5)
\end{aligned}$$

This framework has simple yet powerful relapses. In context with spatially correlated units, Gaussian process priors can improve the point estimate quality both in terms of bias and in terms of efficiency. Moreover, from the posterior distribution of $\beta_{\mathbf{i}}$ we can derive the smoothed path of the coefficient for some control unit $i$ across the treated units $d \in \mathbf{D}$. Lastly, we can easily derive credibility intervals for the posterior distribution of the causal effect, retrieving it from the posterior distribution of $\beta_i$.

## 4. Estimating the effect of the Florentine tramway construction

Figure 1 show the results of our computation. Our results show that the tramway has provoked generally an increase in the commercial vitality of the area considered. These results are particularly significant for the areas closer to the tramway stops, as we find significant average treatment effects for the areas within 50 and 100 meters of the treatment sites. The positive, yet non-statistically significant effects are present for the outer areas, from 150 to 400 meters away from the tramway stops. Worksites have not extensively damaged the commercial environment of the treated area. We can note a significant and negative effect for the area within 100 meters during the period 2006-2010. That time span was the construction period of the tramway, and thus we could expect worse outcomes for areas close to the construction site. However, the number of stores steadily recovered in 2010, the inauguration year, and the overall effect, even for this particularly affected area, is still positive. For this purpose, it is worth noting that in the closer area to the treatment site, the positive effect is present since the start of the construction period, some retailers anticipate their competitors by locating the shops in the most served areas even before the start of tramway operations. The effect on the outer bands is similar to the ones found for inner areas. In particular, we notice that worksites has not affected the commercial environment of the outer areas, while the tramway has improved the accessibility of the area, leading to an increase in the number of shops present. The estimated causal effect has a growing tendency, especially for the outer areas, that exhibits statistically significant effects in the last observational periods. Concluding, in this work we propose a framework for matrix completion with spatial data, an open challenge in policy evaluation literature. We provide convincing results for our motivating application, showing the spatial diffusion of the causal effect. Simulation results, not provided here, confirms the good properties of the proposed estimation framework.

DRAFT

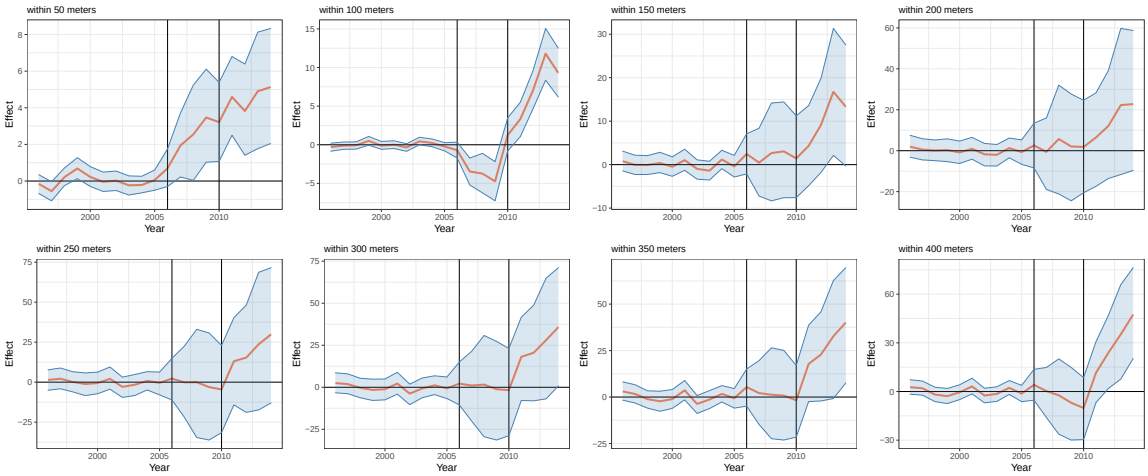



Figure 1: Treatment effect for areas within $d$ meters from a tramway stop, Red line: Treatment effect, Blue area: 90% Credibility interval - First vertical line: tramway worksite starts (2006) - Second vertical line: tramway operational (2010)

DRAFT

## References

[1] Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

[2] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

[3] Arbour, D., Ben-Michael, E., Feller, A., Franks, A., and Raphael, S. (2021). Using multitask gaussian processes to estimate the effect of a targeted effort to remove firearms. *arXiv preprint arXiv:2110.07006*.

[4] Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

[5] Kim, S., Lee, C., and Gupta, S. (2020). Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852.

[6] Menchetti, F. and Bojinov, I. (2020). Estimating causal effects in the presence of partial interference using multivariate bayesian structural time series models. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (21-048).

[7] Pang, X., Liu, L., and Xu, Y. (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288.

[8] Pinkney, S. (2021). An improved and extended bayesian synthetic control. *arXiv preprint arXiv:2103.16244*.

[9] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[10] Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.

# The impact of traffic flow and road signs on road accidents: an approach based on spatiotemporal point pattern analysis on linear networks

Andrea Gilardi[a] and Riccardo Borgoni[a]

[a]University of Milano - Bicocca; Department of Economics, Management and Statistics;
andrea.gilardi@unimib.it, riccardo.borgoni@unimib.it

**Abstract**

Road accidents represent a concern for modern societies, especially in poor and developing countries. In this paper, we develop a road safety model assuming that the car crashes recorded in Milan (Italy) during 2019 can be appropriately modelled as a realisation of a spatio-temporal point process on a linear network. We adopt a separable first-order intensity function with spatial and temporal components. The temporal dimension is estimated semi-parametrically using an additive Poisson regression model. The spatial dimension is estimated semi-parametrically considering a b-spline transformation of two potentially relevant space-varying covariates, namely the traffic flows and the distance to the closest road sign. This approach permits us to analyse traffic accidents at a very granular spatial scale, hence avoiding potential biases due to data aggregation.

***Keywords:*** Car crashes, Linear network, Poisson process

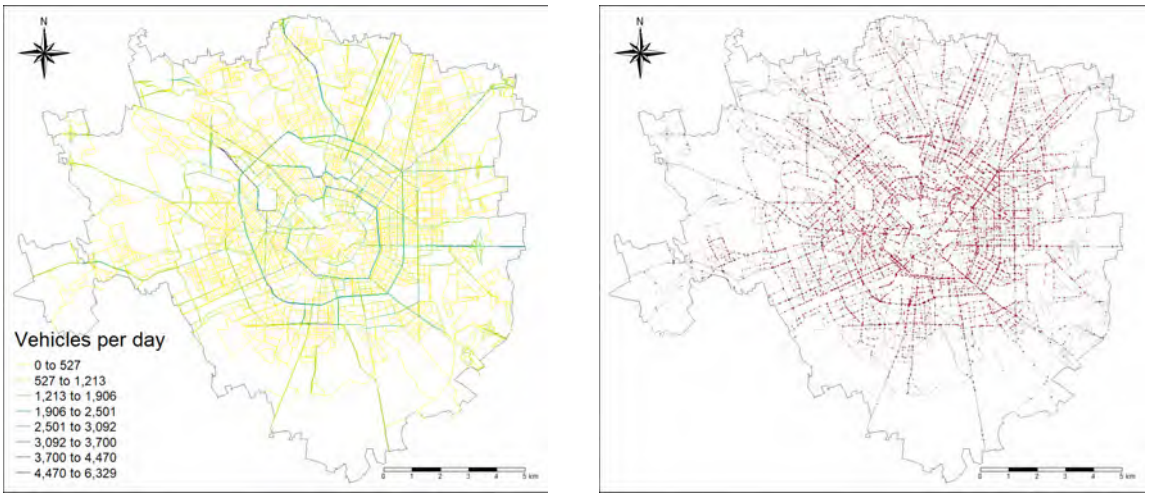
DRAFT

## 1. Introduction

According to the World Health Organisation, car crashes are responsible for more than 1 million casualties each year, representing "the leading cause of death for children and young adults aged 5-29 years" (7). The statistics highlight that the burden is disproportionally borne by vulnerable road users (such as cyclists or pedestrians) and that these problems are particularly relevant in low- and middle-income countries. Therefore, road injuries were named "neglected and silent epidemic", demanding evidence-based interventions and innovative approaches to understand how to reduce the huge annual death toll and establish its determinants.

Car accidents represent a typical example of a point pattern occurring on a linear network, i.e. a set of linear features corresponding to the street segments of a road network. In the last years, we observed a surge of interest in the statistical analysis of such processes, which typically exhibit several complexities due to the restricted spatial domain and the geometrical complexity of the network. The first approaches were introduced in a purely spatial context, mainly focusing on non-parametric techniques for estimating the intensity function (8). More recently, a few authors focused on defining a parametric model for the first-order intensity function that takes into account a set of spatially-varying covariates (3). Following their work, this paper proposes a spatio-temporal semi-parametric Poisson model to analyse the car crashes that occurred in the road network of Milan (Italy) during 2019, relating the intensity function with two variables representing traffic flows and (shortest-path) distances to road signs.

The rest of the article is structured as follows. The car crashes database and the relevant covariates are presented in Section 2. The statistical model is introduced in Section 3, whereas the results are summarised in Section 4, which ends the paper.




(a) (b)

Figure 1: (a) Choropleth map displaying the estimates of daily traffic volumes at the street segment level in the city of Milan. (b) Location of all car crashes that occurred in 2019 and required an ambulance intervention.

## 2. Data and covariate construction

The datasets analysed in this paper come from three different geo-referenced sources that were combined together into a unique object suitable for estimating the statistical model detailed below. In the remaining part of this section, we briefly introduce each provider and describe the procedures adopted to pre-process the data.

**Traffic flows:** The street network and the GPS counts, representing respectively the spatial domain and one of the covariates included in our model, were obtained from TomTom Move service (https://move.tomtom.com/). The network is composed of 33244 geo-referenced segments that are associated with traffic volume estimates obtained from mobile devices connected to cars and anonymous GPS-equipped smartphones. Traffic volumes on Milan's road network are depicted in Figure 1a. As we can see, the traffic estimates are provided with a very granular spatial coverage.

**Car crashes data:** We considered all car crashes that occurred in the city of Milan (IT) from 2019-01-01 to 2019-12-13 and required an ambulance intervention. The raw dataset was provided by the regional Emergency Medical System agency and it was processed by applying the following operations. First, we removed all records with missing spatial or temporal coordinates. Then, we excluded the events farther than 30 metres from the closest segment in the linear network since they are assumed to occur in streets not included in our database. The final sample included 8586 road accidents that were projected into the network. Accident locations are depicted in Figure 1b.

**Road signs:** The information regarding the road signs in the city of Milan was taken from Open Street Map (9). Geographical coordinates of traffic lights, pedestrian crossings and speed bumps were retrieved from the OSM servers and projected onto the spatial network. Similarly to the previous case, we adopted a threshold of 30m to decide which points must be excluded from our dataset. Eventually, we ended up with 21820 observations. The shortest path distance from each point in the network to the closest of these points has been calculated and included as a covariate in the statistical model.

DRAFT

## 3. Spatio-temporal modelling of the road accidents point process

As already mentioned, we modelled the car crashes occurrences as a spatio-temporal point pattern on a linear network. Formally, a linear network $L$ is defined as the union of a finite number of segments, say $l_i$, lying in a planar region $S$:

$$l_i = [\mathbf{u}_i, \mathbf{v}_i] = \{\mathbf{s} : \mathbf{s} = c\mathbf{u}_i + (1-c)\mathbf{v}_i;\ 0 \leq c \leq 1\},$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the endpoints of $l_i$ stored using an appropriate coordinate reference system. Let $\mathcal{T} = \{1, 2, \ldots, T\}$ be a discrete temporal dimension divided into intervals of, say, two hours. A spatio-temporal point pattern on $L$ is a finite realisation of a stochastic process on $L \times \mathcal{T}$. In this paper, we assume that for each $t \in \mathcal{T}$, the observed events represent a realisation of a *non-homogeneous Poisson Process* (NHPP) with *intensity function* $\lambda_t(\boldsymbol{s})$ (1; 4).

Following a classical hypothesis in the literature of spatio-temporal point patterns (4), we further assume the separability of the intensity function

$$\lambda_t(\boldsymbol{s}) = \mu_t g(\boldsymbol{s}) \quad \text{for } t \in \mathcal{T} \text{ and } \boldsymbol{s} \in L, \tag{1}$$

where $\mu_t$ and $g(\boldsymbol{s})$ represent the temporal and spatial dimension of the process at time $t$, respectively. More precisely, if we denote by $y_t$ the number of events at time $t$, under the NHPP assumption we have $y_t|\lambda_t \sim \text{Poisson}(\mu_t)$, implying that $\mu_t$ represents the *expected volume* of road crashes occurring all over the network at time $t$. Similarly, denoting by $\boldsymbol{s}_{t,i}$ the location of the $i$th, $\{\boldsymbol{s}_{t,i}\}|\lambda_t, y_t \overset{\text{iid}}{\sim} g(\boldsymbol{s})$, highlighting that $g(\boldsymbol{s})$ represents the (common) *spatial density function* of the road accidents. Hereinafter, we introduce two statistical models for $\mu_t$ and $g(\boldsymbol{s})$, respectively.

DRAFT

### 3.1 The temporal model

As mentioned above, $\mu_t$ represents the expected number of car crashes observed over the network during a two-hours interval $t$. Following the suggestions in (5), we modelled the bi-hourly counts using Poisson regression considering the hour of the day, the day of the week, and the week of the year as predictors. To incorporate smoothness into the model, Generalized Additive Models (GAMs) are used in the estimation process (10). GAMs extend Generalized Linear Models allowing for non-linear relationships between the response variable and the covariates.

Considering the previous assumptions, the (log-linear) Poisson additive regression model is given by

$$\log \mu_t = \beta_0 + \text{dow}_t + \text{dow}_t \times \beta_1(\text{hour}_t) + \beta_2(\text{week}_t), \tag{2}$$

where $\beta_0$ is the intercept, $\text{dow}_t$ is a factor variable denoting the day of the week, $\text{hour}_t$ represents the hour of the day (taking values from 0 to 23) while $\text{week}_t$ represents the week of the year (taking values from 1 to 53). The terms $\beta_j(x),\ j = 1, 2$, represent spline transformations, i.e. $\beta_j(x) = \sum_{r=1}^{k_j} b_{jr}\gamma_{jr}(x)$, where $\gamma_{jr}(x), r = 1, \ldots, k_j$ are the basis functions and $b_{jr}$ the unknown coefficients. In particular, a cyclic cubic regression spline is adopted since in our context it is appropriate to assume a smooth transition between the last hour of one day and the first hour of the next day as well as between the last week of one year and the first week of the next year.

### 3.2 The spatial model

The spatial component of the intensity function at location $\boldsymbol{s} \in L$, previously denoted by $g(\boldsymbol{s})$, is modelled semi-parametrically as a function of the two covariates described in Section 2: the traffic counts in the road segments ($Z_1$) and the shortest-path distance to a traffic sign ($Z_2$). The model of the spatial component reads

$$\log g(\boldsymbol{s}) = \alpha_0 + \alpha_1(\log Z_1(\boldsymbol{s})) + \alpha_2(Z_2(\boldsymbol{s})) \tag{3}$$

where $\alpha_0$ is an intercept and $\alpha_j(x),\ j = 1, 2$ are spline transformations defined as before.

The parameters in Equation (3) were estimated using a likelihood-based approach. In general, the log-likelihood function for an NHPP on a linear network L writes as

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log g(\boldsymbol{s}_i; \boldsymbol{\theta}) - \int_L g(\boldsymbol{s}; \boldsymbol{\theta}) \, \mathrm{d}_1 \boldsymbol{s} \tag{4}$$

where $n = \sum_{t \in \mathcal{T}} y_t$ represents the total number of points, $\boldsymbol{\theta}$ is the set of model parameters, and $\mathrm{d}_1 \boldsymbol{s}$ denotes integration with respect to arc-length measure (1). The MLE $\hat{\boldsymbol{\theta}}$ is typically not available in closed form since it depends on the form of $g(\boldsymbol{s}; \boldsymbol{\theta})$ and must be derived using numerical approximation techniques. In particular, in this paper, we employed a strategy named *Berman - Turner device*, originally developed for planar point patterns in (2) and recently extended to linear network data (1).

The first step of the aforementioned procedure involves approximating the integral in Equation (4) by a weighted sum over *quadrature points* $\{\tilde{\boldsymbol{s}}_j\}_{j=1}^m$ with weights $\{w_j\}_{j=1}^m$

$$\int_L g(\boldsymbol{s}; \boldsymbol{\theta}) \, \mathrm{d}_1(\boldsymbol{s}) \approx \sum_{j=1}^{m} w_j g(\tilde{\boldsymbol{s}}_j; \boldsymbol{\theta})$$

yielding to the following approximation for the log-likelihood function

$$\log L(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} \log g(\boldsymbol{s}_i; \boldsymbol{\theta}) - \sum_{j=1}^{m} w_j g(\tilde{\boldsymbol{s}}_j; \boldsymbol{\theta}). \tag{5}$$

Then, the key part of the Berman-Turner device is observing that if the quadrature points $\{\tilde{\boldsymbol{s}}_j\}_{j=1}^m$ contain the $n$ observed events, Equation (5) can be rewritten as

$$\log L(\boldsymbol{\theta}) \approx \sum_{j=1}^{m} \Big( Z_j \log g(\tilde{\boldsymbol{s}}_j; \boldsymbol{\theta}) - g(\tilde{\boldsymbol{s}}_j; \boldsymbol{\theta}) \Big) w_j \tag{6}$$

where $Z_j = I_j / w_j$ and $I_j$ is an indicator function which is equal to 1 if $\tilde{\boldsymbol{s}}_j$ denotes an observed event (instead of a quadrature point) and 0 otherwise. Equation (6) represents the likelihood function of $m$ independent Poisson random variables with mean $g(\tilde{\boldsymbol{s}}; \boldsymbol{\theta})$ and weight $w_j$, implying that $\hat{\boldsymbol{\theta}}$ can be derived using standard GLM tools.

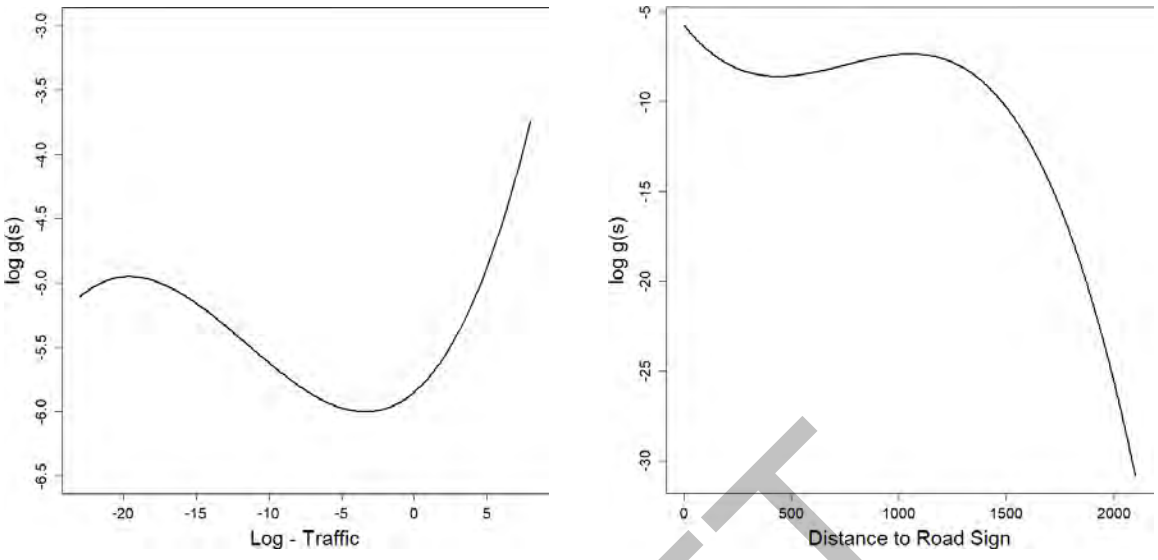
DRAFT

## 4. Results and Conclusions

We can now present the results obtained after applying the methodology described in Section 3. to the accident data presented in Section 2.

Figure 2a and 2b display the effects on the spatial density $g(\boldsymbol{s})$ due to variations in traffic flows and distance to the closest road sign. For low traffic flow, the accident density tends to increase slightly. It can be argued that low traffic flows are typically found in residential streets where, despite mild vehicular traffic, accidents can be present due to intensive pedestrian and cyclist activities. When the traffic gets moderate, which is typical in medium-sized streets, this effect declines. Finally, the car crashes density increases substantially on traffic-intensive roads. Figure 2b shows the impact of the distance to the (closest) road sign. It clearly appears that being away from a pedestrian crossing, a traffic light or a speed bump has a protective effect against accidents with a striking decline when the distance gets larger than, say, 1-1.5km, which is typical in fast roads and highways.

The methodology proposed in this paper allows us to estimate the proneness of accidents in the road network of the city of Milan. More specifically, by combining the estimated version of model (2) and (3), we are able to calculate the estimate of the intensity function of the spatio-temporal point process at any desired location of the network and time of the day. Hence, after discretising the linear network by a fine pixel grid, intensity estimates can be computed efficiently at any grid location of the raster to produce a map that graphically displays the spatio-temporal dynamic.




(a) (b)

Figure 2: Covariates' effects obtained after applying a spline transformation. a) road traffic (log scale); b) distance to the closest road sign (metre).

DRAFT

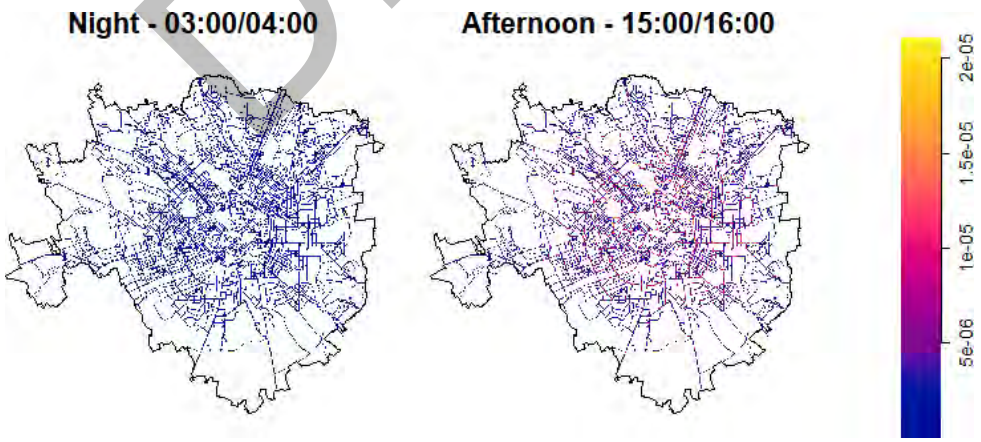



Figure 3: Choropleth maps showing the estimated intensity of events occurring in the street network of Milan for two particular time periods. The colour scale represents the expected number of car crashes occurring in a small linear neighbourhood around a point of the network.

Figure 3 represents the map obtained using this approach for two different temporal occasions. The values reported in the maps represent the expected number of car crashes occurring in a small linear neighbourhood around a point taken in the network.

We finally observed that our dataset does not include all those accidents that did not require medical intervention, hence the intensity function can be somewhat underestimated. However, these events are expected to be of lower severity, hence less crucial for a road safety program. In addition, we acknowledge that the separability assumption of our model is a bit simplistic since the spatial configuration of car crashes can, to some extent, change over time. This aspect has been considered by (6) in a similar context using a kernel-based non-parametric approach. Tackling this issue in a parametric or semi-parametric setting, however, is not an easy task and is material for future research.

## Acknowledgements

This study was carried out within the MOST - Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 - D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

## References

[1] Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G. and Davies, T.M., 2021. Analysing point patterns on networks - A review. Spatial Statistics, 42, p.100435.

[2] Berman, M. and Turner, T.R., 1992. Approximating point process likelihoods with GLIM. Journal of the Royal Statistical Society: Series C (Applied Statistics), 41(1), pp.31-38.

[3] D'Angelo, N., Adelfio, G., Abbruzzo, A. and Mateu, J., 2022. Inhomogeneous spatio-temporal point processes on linear networks for visitors' stops data. The Annals of Applied Statistics, 16(2), pp.791-815.

[4] Diggle, P.J., 2013. Statistical analysis of spatial and spatio-temporal point patterns. CRC press.

[5] Diggle, P., Rowlingson, B. and Su, T.L., 2005. Point process methodology for on-line spatio-temporal disease surveillance. Environmetrics: The official journal of the International Environmetrics Society, 16(5), pp.423-434.

[6] Gilardi, A., Borgoni, R., Mateu, J., 2021 A non-separable first-order spatio-temporal intensity for events on linear networks: an application to ambulance interventions: arXiv. https://arxiv.org/abs/2106.00457

[7] Global status report on road safety 2018. Geneva: World Health Organization; 2018.

[8] McSwiggan, G., Baddeley, A. and Nair, G., 2017. Kernel density estimation on a linear network. Scandinavian Journal of Statistics, 44(2), pp.324-345.

[9] OpenStreetMap contributors. (2015) Planet dump [Data file from 2022]. Retrieved from https://planet.openstreetmap.org.

[10] Wood, S.N. (2017). Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

DRAFT

# A clustering model for flow data: an application to international student mobility

Cinzia Di Nuzzo[a] and Donatella Vicari[a]

[a]Department of Statistical Sciences, Sapienza University of Rome ,
cinzia.dinuzzo@uniroma1.it, donatella.vicari@uniroma1.it

**Abstract**

A new clustering model for skew-symmetric matrices is introduced to analyse flow data. This model aims to find clusters of objects that have a significant flow, interpreted as exchange intensity. The model analyses the within-clusters effects between objects and provides the directions of the flows within clusters. Formally, it is based on the decomposition of the data skew-symmetric matrix into within-cluster components, i.e. the skew-symmetric matrix is decomposed into a sum of diagonal block skew-symmetric matrices. The model is estimated in a least-squares sense through the SVD of the skew-symmetric matrices. An application to the international student mobility is discussed.

***Keywords:*** Flow data, Skew-symmetry, Within-cluster effects

DRAFT

## 1. Motivating example and Introduction

In order to analyse the international mobility of students in the OECD countries, flow data models can be useful tools for identifying clusters of countries where international students come from and move to in order to enrol in different education programs abroad.

Studying abroad has become a key differentiating experience for university students, and international student mobility has received increasing policy attention in recent years. Studying abroad is an opportunity to access a high-quality education, acquiring skills that may not be taught in the country of origin and it is also seen as a way to broaden one's knowledge and improve language skills, especially English. For their countries of origin, mobile students could be seen as lost talent (the classic "brain drain"). However, they can contribute to knowledge improvement and technological upgrading in their origin country. Mobile students acquire knowledge that is often shared through direct personal interactions and can enable their origin country to integrate into global knowledge networks. Some research suggests that students abroad are a good predictor of future flows of scientists, providing evidence for the "brain circulation" effect. Student mobility appears to shape international scientific cooperation networks more profoundly than a common language or geographical or scientific proximity (3).

In this context, in order to analyse whether a country has more international students in incoming or outgoing mobility, it becomes useful to understand if there exist clusters of countries that exchange more mobile students with each other and above all it is important to provide the directions of such flows. To this end, in this work, we introduce a new model to identify clusters of countries that are similar in terms of flows within clusters and able to give information also on what are the directions within each cluster. Specifically, the model proposed here analyses the within-cluster effects of a skew-symmetric matrix describing the exchanges between objects within clusters. In order to estimate the model, singular value decompositions (SVD) are considered which are implemented in an Alternating

Least Squares algorithm. Furthermore, this model allows for a graphical interpretation of the results in terms of amounts and directions of the imbalances within clusters.

The rest of the work is organized as follows: in Section 2. the model is formalized, and in Section 3. an application to the flows of the international student mobility in tertiary education in the OECD founding countries is analysed by studying the number of students enrolled by country of origin and destination.

## 2. The Model

Let $\mathbf{K} = (k_{ij})$ be a $(N \times N)$ skew-symmetric matrix, i.e. $k_{ij} = -k_{ji}$, for all $i, j = 1, \ldots, N$ (for a review on methodologies for studying asymmetric and skew-symmetric data see (1)). Clustering models for fitting skew-symmetries have been proposed in (4) and (5) following different approaches to account for both within and between cluster effects.

Here, we introduce a clustering model that takes into account the within cluster effects and aims to cluster the skew-symmetric matrix $\mathbf{K}$ by considering a partition of the $N$ objects into $C$ disjoint clusters which can be identified by an $(N \times C)$ binary membership matrix $\mathbf{U} = (u_{ic})$ for $i = 1, \ldots N$ and $c = 1, \ldots, C$.

Given a partition $\mathbf{U}$, we are interested in modelling the imbalances within clusters as follows

$$\mathbf{K} = \mathbf{W} + \mathbf{E}, \tag{1}$$

where $\mathbf{W}$ is the $(N \times N)$ skew-symmetric diagonal block matrix of the imbalances *within* clusters.

The within matrix $\mathbf{W}$ can be decomposed as

$$\mathbf{W} = \sum_{c=1}^{C} \mathbf{W}^{(c)}, \tag{2}$$

where $\mathbf{W}^{(c)}$ is the $(N \times N)$ skew-symmetric matrix of the exchanges within cluster $c$. Any skew-symmetric matrix $\mathbf{W}^{(c)}$ can be approximated as

$$\mathbf{W}^{(c)} = \mathbf{v}_1^{(c)} \mathbf{v}_2^{(c)\prime} - \mathbf{v}_2^{(c)} \mathbf{v}_1^{(c)\prime} + \mathbf{E}_W^{(c)}, \qquad \text{for } c = 1, \ldots, C, \tag{3}$$

where, $\mathbf{v}_1^{(c)}$ and $\mathbf{v}_2^{(c)}$ are orthogonal vectors of size $N$, and $\mathbf{E}_W^{(c)}$ is the $(N \times N)$ residual matrix.

Therefore, model (1) can be written as

$$\mathbf{K} = \sum_{c=1}^{C} \left[ \mathbf{v}_1^{(c)} \mathbf{v}_2^{(c)\prime} - \mathbf{v}_2^{(c)} \mathbf{v}_1^{(c)\prime} \right] + \mathbf{\Xi}, \tag{4}$$

subject to

$$u_{ic} \in \{0, 1\}, \quad \sum_{c=1}^{C} u_{ic} = 1, \text{ for } c = 1, \ldots, C, \quad i = 1, \ldots, N,$$

$$\mathbf{v}_1^{(c)\prime} \mathbf{v}_2^{(c)} = 0, \text{ for } c = 1, \ldots, C.$$

Model (4) is fitted in the least squares-sense thanks to the SVD of any within skew-symmetric matrix.

## 3. A case of study: the international student mobility flow in OECD countries

In this section, model (4) is fitted to study the flow of the enrolment of international students between countries. Specifically, the data can be downloaded from the website http://www.oecd.

DRAFT

org/education/education-at-a-glance/ and refer to the number of students enrolled in different education programs abroad by country of origin and destination in 2017 for the 20 founding countries of the OECD, namely, Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, UK, USA, which represent around 84% of the international university student mobility flows within all OECD countries in 2017.

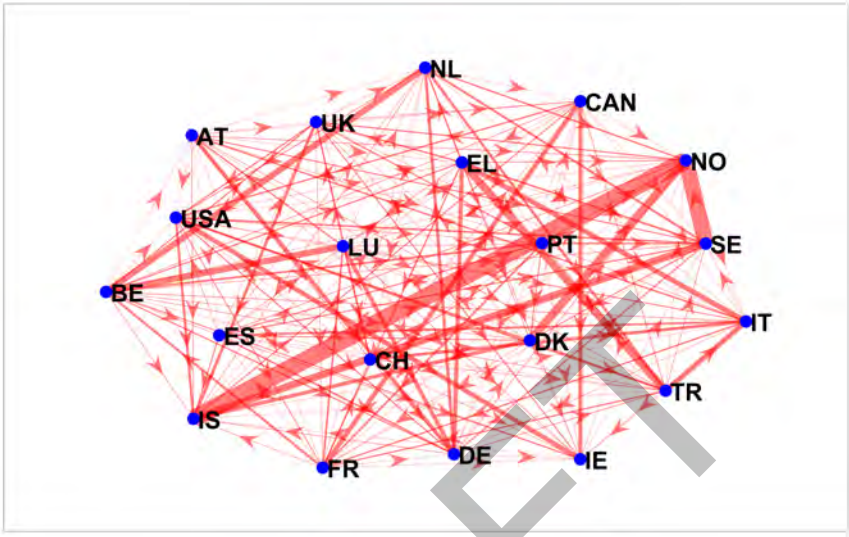



Figure 1: Directed graph of the skew-symmetric matrix **K** of the international student flows for the 20 founding countries of the OECD in 2017.

DRAFT

The original mobility data of the asymmetric origin/destination matrix **A** have been firstly pre-processed to incorporate the information on the amounts of the flows, and then transformed into a skew-symmetric matrix **K** as follows

$$A^*_{ij} = \frac{A_{ij} \sum_{i=1}^{20} \sum_{j=1}^{20} A_{ij}}{\sum_{i=1}^{20} A_{ij} \sum_{j=1}^{20} A_{ij}}, \qquad \text{for } i, j = 1, \ldots, 20,$$

$$K_{ij} = \frac{A^*_{ij} - A^*_{ji}}{2}, \qquad \text{for } i, j = 1, \ldots, 20.$$

Once computed matrix **K**, the clustering model (4) has been fitted. Matrix **K** is displayed in the graph in Fig.1 where some relations between countries are quite evident. In fact, it is possible to distinguish high mobility flows between some pairs of countries, such as Iceland-Norway or Turkey-Greece.

The proposed model (4) has been fitted by varying $C = 1, \ldots, 8$ and, from the loss function values, the model suggests to select 4 clusters of countries with a goodness-of-fit equal to 85%. The clustering results are described in Table 1 and the directions of the international student mobility within clusters of countries are represented in Fig.3.

Table 1: Clustering of the founding OECD countries.

| **CLUSTERS** | **COUNTRIES** |
|---|---|
| $C_1$ | Spain, Sweden, Island, Norway, UK, Denmark |
| $C_2$ | Greece, Turkey, Switzerland, Italy |
| $C_3$ | USA, Canada, Ireland, France |
| $C_4$ | Netherlands, Germany, Belgium, Luxembourg, Austria, Portugal |

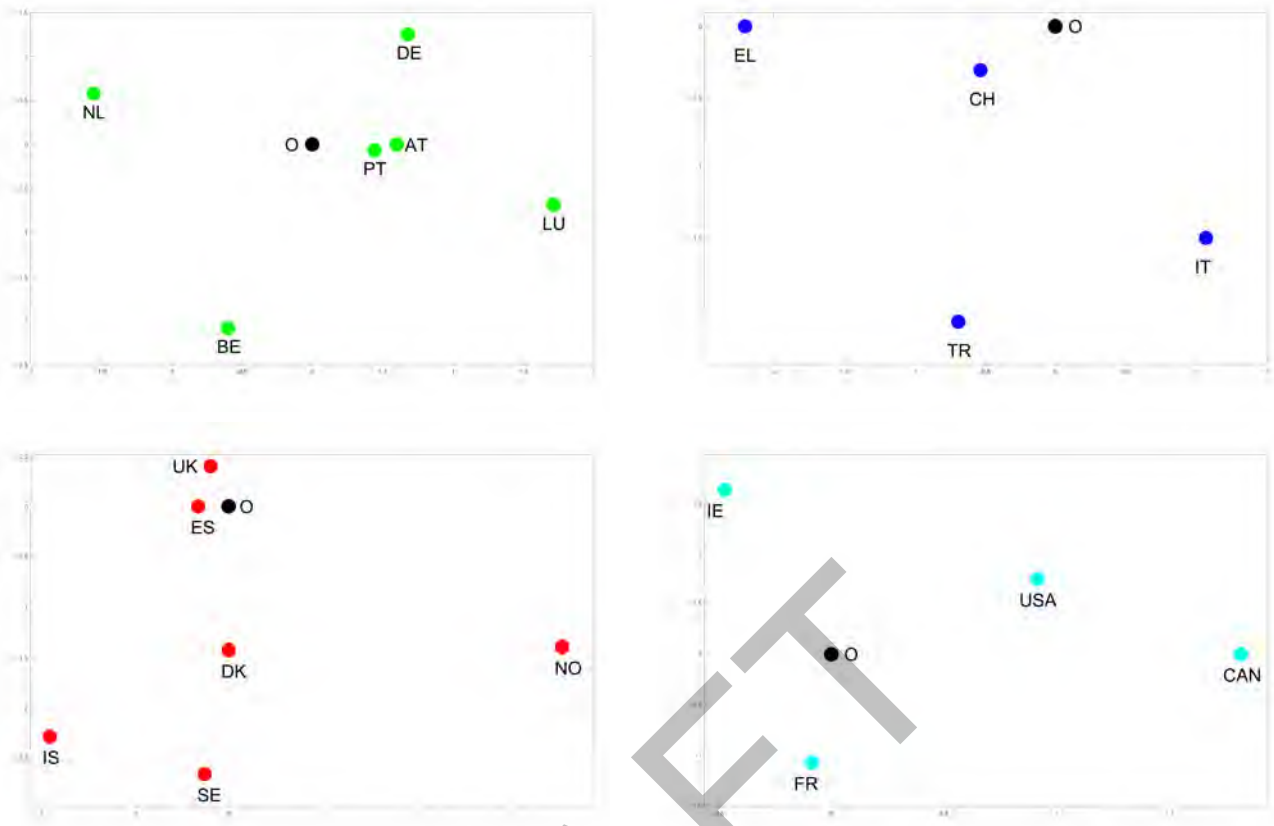



Figure 2: Scatter plot of the OECD countries. Colours identify countries in same clusters

DRAFT

One of the advantages of this model is the possibility to make a graphical representation and provide an interpretation in terms of directions and amounts of the skew-symmetries within clusters (2), (1). Therefore, in Fig.2 the positions of the countries can be interpreted in terms of areas: the areas of the triangles that all pairs of countries form with the origin O are approximatively proportional to the amounts of the skew-symmetries, i.e. to the imbalances between flows of international students between countries. Moreover, it is possible to give an interpretation in terms of directions: by convention, 1) a clockwise direction denotes a negative skew-symmetry, i.e. the destination country of the student flow; 2) an anticlockwise direction indicates a positive skew-symmetry, i.e. the country from which the student flow originates. Thus, according to the interpretation of the Gower diagram (2), from Fig.2 we can deduce the directions of the student mobility within each cluster of countries summarized in Fig.3. In particular, in Fig.2 we point out that Spain-UK, Portugal-Austria have small skew-symmetries because are positioned close to the origin: there is a low flow of students between these two pairs of countries, i.e., for example, few Portuguese students go to study in Austria. Similarly, it can be noted that Netherlands, Portugal, and Luxembourg are nearly collinear with the origin, so they represent small skew-symmetries, i.e. there is a low flow of students between these countries. Moreover, it can be seen that the area of the triangle between Norway, Sweden, and the origin is very large and Norway has a clockwise direction towards Sweden, which means that there are many student exchanges from Sweden to Norway. In general, Fig.2 reveals a large flow of students between the Scandinavian countries. Similar considerations can be made for Turkey and Greece, as the student flow from Greece to Turkey is high. These findings are easily interpretable due to both the geographical proximity of the countries and their culture. Finally, it is also important to point out that USA, Canada and Ireland have a rather high flow of students exchanged between them, both for geographical proximity and for the common language: the clockwise direction from Ireland to USA connotes Ireland as a country of destination from USA, while the anticlockwise direction from Canada to USA qualifies Canada as the origin country for students going to the United States. Another group of countries clearly closely linked to the geographical aspect

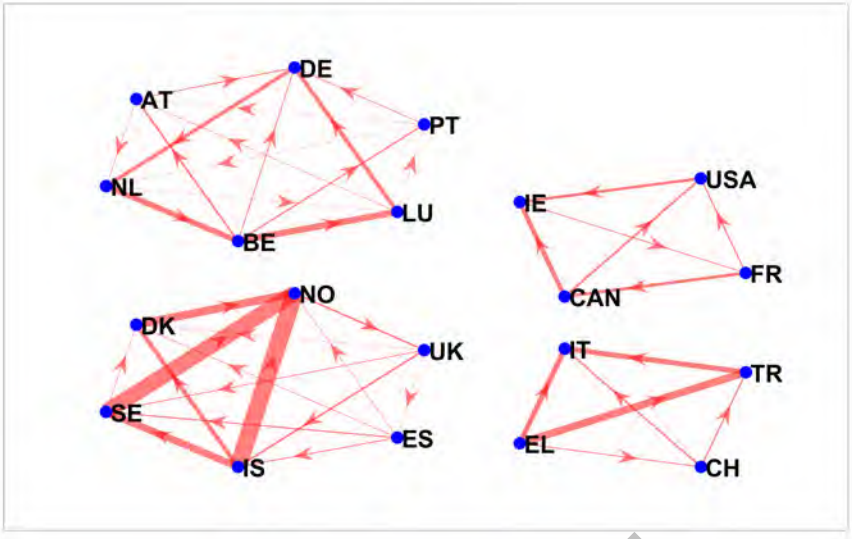



Figure 3: Directions of the flows within each cluster.

and the language spoken is group $C_4$ which mainly contains countries where French and German are spoken and which are bordering each other. All in all, the clustering results (Fig.3) summarize the main links of Fig.1 which agree with the OECD report (3).

DRAFT

## References

[1] Bove G., Okada A., Vicari D., Methods for the Analysis of Asymmetric Proximity Data, Springer Nature Singapore Pte Ltd (2021)

[2] Gower J.C., Skew symmetry in retrospect. Advances in Data Analysis and Classification 12: 33–41 (2018)

[3] OECD, Education at a Glance 2019: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/f8d7880d-en (2019)

[4] Vicari D., Classification of asymmetric proximity data. J Classif 31(3): 386–420 (2014)

[5] Vicari D., Modeling Asymmetric Exchanges Between Clusters. In: Imaizumi, T., Nakayama, A., Yokoyama, S. (eds) Advanced Studies in Behaviormetrics and Data Science. Behaviormetrics: Quantitative Approaches to Human Behavior, vol 5. 297–313. Springer, Singapore (2020)

# Contingency tables with structural zeros and discrete copulas

Roberto Fontana[a], Elisa Perrone[b], and Fabio Rapallo[c]

[a]Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy, roberto.fontana@polito.it

[b]Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands, e.perrone@tue.nl

[c]Dipartimento DIEC, Università di Genova, via Vivaldi 5, 16126 Genova, Italy, fabio.rapallo@unige.it

DRAFT

**Abstract**

In this work, we analyze the connection between contingency table analysis and copulas in a discrete framework. We focus on the impact of structural zeros on the general theory presented by Geenens (2020) based on a new idea of copula models for discrete variables. Through examples, we investigate the pros and cons of applying the theory developed by Geenens (2020) and discuss some open questions for future research.

***Keywords:*** Bubble plot, Categorical data analysis, Discrete copulas, Iterated Proportional Fitting (IPF), Structural zeros

## 1. Introduction

Contingency tables appear in many applied fields, such as biology, health care, and social science. Due to their importance in application, the analysis of such tables was studied in statistics, where researchers developed methodological tools to extract information about the relation between the variables involved (9). Recent work by Geenens (4) highlights interesting connections between standard methods in contingency table analysis and *copulas* in a discrete setting. Copulas are popular tools to model dependencies between random variables. Their popularity is due to Sklar's theorem (11), which states that, for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$, the joint distribution function $F_{\mathbf{X}}$ of any $d$-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can be written as $F_{\mathbf{X}}(x_1, \ldots, x_d) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$, where the function $C : [0,1]^d \to [0,1]$ is a *d-dimensional copula* and $F_{X_1}, \ldots, F_{X_d}$ are univariate marginal distributions. In a nutshell, copulas can be seen as joint probability distributions with uniform margins in $[0,1]$. When the random vector $\mathbf{X}$ is discrete, the copula associated with $F_{\mathbf{X}}$ is uniquely defined on the $\text{Range}(F_{X_1}) \times \cdots \times \text{Range}(F_{X_d})$. Thus, it is possible to associate any contingency table with the restriction of a full-domain copula on a grid domain, i.e., a so-called *discrete copula*; see, e.g., (6; 7) and references therein. Any discrete copula can be extended into a full-domain copula by simply spreading the probability mass on each hyper-rectangle of their grid domain. However, the extension is not unique (1), which leads to problems while applying copula theory in discrete settings. In this work, we elaborate on a novel approach to using copulas in contingency table analysis presented in (4). In Sect. 2, we recall basic definitions (in the bivariate case) and results on the topic with special attention to contingency tables with structural zeros, and we discuss the connections with some classical log-linear models for incomplete tables. Some interesting examples are illustrated in Sect. 3, where we also discuss some open questions for future research on the topic.

## 2. Background

In this section, we provide the mathematical framework and notation to work with discrete copulas. We consider $R \in \mathbb{Z}_{>0}$ and denote $I_R = \{0, 1/R, \dots, (R-1)/R, 1\}$, $[R] = \{1, \dots, R\}$, and $\langle R \rangle = \{0, \dots, R\}$. For $R$ and $S$ in $\mathbb{Z}_{>0}$, we define $U_R = \{u_0 = 0, u_1, \dots, u_{R-1}, u_R = 1\}$, $u_0 < \dots < u_R$ and $V_S = \{v_0 = 0, v_1, \dots, v_{S-1}, v_S = 1\}$, $v_0 < \dots < v_S$ as two finite grid partitions of the unit interval. A discrete copula $C_{U_R,V_S}$ is a function defined on $U_R \times V_S$ that satisfies the properties of a copula function on the grid domain $U_R \times V_S$. As highlighted in (6), there are interesting connections between the space of discrete copulas and convex polytopes called *transportation polytopes*, which are also linked with contingency tables analysis (2) . Namely, considering two vectors $\tilde{u} = (\tilde{u}_1, \dots, \tilde{u}_R) \in \mathbb{R}^R_{>0}$ and $\tilde{v} = (\tilde{v}_1, \dots, \tilde{v}_S) \in \mathbb{R}^S_{>0}$, we can define the transportation polytope $\mathcal{T}(\tilde{u}, \tilde{v})$ as the convex polytope in the $RS$ variables $x_{i,j}$ satisfying, for all $i \in [R]$ and $j \in [S]$, the following conditions: $x_{i,j} \geq 0$, $\sum_{h=1}^{S} x_{i,h} = \tilde{u}_i$, $\sum_{\ell=1}^{R} x_{\ell,j} = \tilde{v}_j$. The two vectors $\tilde{u}$ and $\tilde{v}$ are called the margins of $\mathcal{T}(\tilde{u}, \tilde{v})$. In (8), the authors show that any discrete copula $C_{U_R,V_S}$ corresponds to a matrix within a transportation polytope $\mathcal{T}(\tilde{u}, \tilde{v})$, and viceversa. Intuitively, the transportation matrix directly relates to the probability mass function of the discrete random vector, while the corresponding discrete copula relates to the cumulative distribution function.

We now show how to derive the discrete copula associated with a given contingency table. We consider an example taken from (10) whose data is reported in Table 1. The table shows the cross-classification of a father's and his son's occupational status categories in Japan in 1955. There are four categories (1: Professional and Managers; 2: Clerical and Sales; 3: Skilled manual and Semiskilled manual; 4: Unskilled manual and Farmers). Since this table is analyzed in (10) under quasi-symmetry models, we have removed the diagonal, because the diagonal cells are fitted exactly, and thus there is no variability. We get $N = 799$ values. In this example, $R = S = 4$, the vectors $\tilde{u}$ and $\tilde{v}$ are the margins of the contingency table, i.e., $\tilde{u} = (128, 136, 144, 391)$ and $\tilde{v} = (139, 301, 264, 95)$, while the defining grids of the corresponding discrete copula are $U_R = \frac{1}{N}\{0, \tilde{u}_1, \tilde{u}_1 + \tilde{u}_2, \dots\} = \{0, 0.16, 0.33, 0.51, 1\}$ and $V_S = \{0, 0.17, 0.55, 0.88, 1\}$. The entries of the discrete copula $C_1 = C_{U_R,V_S} = (c_{i,j})$, $i \in [R]$ and $j \in [S]$ are computed from the entries of the contingency table $(x_{i,j})$ by summing up and normalizing, i.e., $c_{i,j} = \frac{1}{N} \sum_{\ell=1}^{i} \sum_{h=1}^{j} x_{\ell,h}$, while $c_{0,0} = c_{i,0} = c_{0,j} = 0$, for $i \in [R]$ and $j \in [S]$.

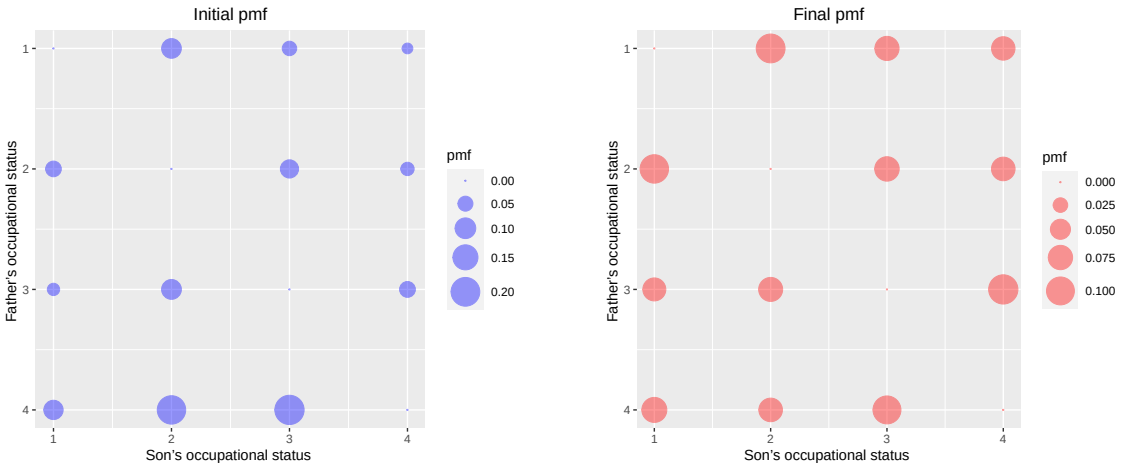
In (4), the author highlights the difficulty of drawing conclusions on the dependence from tables that have non-uniform margins as the one reported in Table 1. Therefore, in the spirit of copula theory for continuous random variables, the author suggests searching for a representative $\bar{\mathbf{p}}$ of all $(R \times S)$ probability distributions that (1) preserves the inter-dependencies of a contingency table in terms of odds-ratios, and (2) has uniform margins equal to $1/R$ and $1/S$. In discrete copula terms, this is equivalent to searching for a discrete copula defined on the rectangular grid $I_R \times I_S$ which preserves the dependence structure of the original discrete copula computed from a given contingency table. Looking at the example above, we would search for a discrete copula $\tilde{C}_1$ with support $I_4 \times I_4$ and margins $I_4$ associated with $C_1$ in a meaningful way. A natural question that arises is whether or not such an element exists and is unique. The answer to this question is given in the theorem below presented in (4). We here report the cases that are relevant to our examples while the original formulation of the theorem is more general and presents more scenarios. The cardinality of a set $A$ is denoted by $|A|$.

**Theorem 1.** *Let* $\mathbf{p}$ *be in the set* $\mathcal{P}_{R\times S}$ *of all* $(R \times S)$ *probability distributions. We define Supp*$(\mathbf{p}) = \{(i,j) \in [R]\times[S]$ *s.t.* $p_{i,j} > 0\}$ *and* $N(\mathbf{p}) = \{(v_X \times v_Y) : v_X \subset [R], v_Y \subset [S] s.t. \sum_{(i,j)\in v_X \times v_Y} p_{i,j} = 0\}$, *the set of rectangular subset of* $[R] \times [S]$ *where* $\mathbf{p}$ *is null.*

1. *Suppose that for all* $(v_X \times v_Y) \in N(\mathbf{p})$, $\frac{|v_X|}{R} + \frac{|v_Y|}{S} < 1$, *then there exists a unique* $\bar{\mathbf{p}}$ *with uniform margins, same odds-ratio structure of* $\mathbf{p}$, *and associated with a discrete copula* $C_{I_R\times I_S}$.
2. *Suppose that there exists* $\tilde{\tilde{v}}_X \times \tilde{\tilde{v}}_Y \in N(\mathbf{p})$ *such that* $\frac{|\tilde{\tilde{v}}_X|}{R} + \frac{|\tilde{\tilde{v}}_Y|}{S} > 1$. *Then there is no element* $\bar{\mathbf{p}}$ *with uniform margins such that it has same odds-ratio structure of* $\mathbf{p}$ *and is associated with a discrete copula* $C_{I_R\times I_S}$.

The element $\bar{\mathbf{p}}$ can be obtained by using the iterated proportional fitting (IPF) procedure, which is a standard method in contingency table analysis for a meaningful comparison of tables with different

DRAFT




(a) observed pmf

(b) final pmf as output of IPF

Figure 1: Bubble plots of the probability mass functions

DRAFT

margins and same dependence structure in terms of corresponding odds-ratios (4; 9). Theorem 1 states that zero-count cells in a contingency table impact the existence of the element $\bar{\mathbf{p}}$ of interest.

When a contingency table contains structural zeros, classical statistical methods become difficult to apply. Some of the odds-ratios can not be defined, and the standard independence model can not be used, because it implies a sample space in the form of a Cartesian product $[R] \times [S]$. For the analysis of incomplete square tables with structural zeros on the main diagonal, as in the example in Table 2, one can use quasi-independence or quasi-symmetry log-linear models. The latter is defined by

$$\log(p_{i,j}) = \lambda + \lambda_i^{(X)} + \lambda_j^{(Y)} + \lambda_{i,j}^{(XY)} \tag{1}$$

for $(i, j) \in [R] \times [S]$, where $\lambda$ is a mean parameter, $\lambda_i^{(X)}$ are the row-parameters, $\lambda_j^{(Y)}$ are the column-parameters, and $\lambda_{i,j}^{(XY)}$, with the constraints $\lambda_{i,j}^{(XY)} = \lambda_{j,i}^{(XY)}$, measure the symmetry beyond the marginal contributions. Although the expression of the model is simple, the practical interpretation of the values of the $\lambda$ parameters is not easy, and usually, log-linear models of this kind are used only for a global goodness-of-fit test. Using the discrete copulas and their graphical visualization through the bubble plots as the ones reported in Fig. 1, we will be able to focus on the dependence described by the $\lambda_{i,j}^{(XY)}$ parameters. In the next section, we further explore this aspect by analyzing two examples of contingency tables with special zero-count cell structures.

Table 1: Father's and son's occupational status categories in Japan in 1955, adapted from (10). Observed data in the left panel (dashes denote structural zeros), and associated discrete copula in the right panel.

| | Son | | | | |
|---|---|---|---|---|---|
| Father | 1 | 2 | 3 | 4 | Total |
| 1 | – | 72 | 37 | 19 | 128 |
| 2 | 44 | – | 61 | 31 | 136 |
| 3 | 26 | 73 | – | 45 | 144 |
| 4 | 69 | 156 | 166 | – | 391 |
| Total | 139 | 301 | 264 | 95 | 799 |

$$C_1 = \begin{pmatrix} 0 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.09 & 0.14 & 0.16 \\ 0 & 0.06 & 0.15 & 0.27 & 0.33 \\ 0 & 0.09 & 0.27 & 0.39 & 0.51 \\ 0 & 0.17 & 0.55 & 0.88 & 1.00 \end{pmatrix}$$

## 3. Examples and discussion

The first example we consider here is taken from (10) and we have already described it in Section 2. The corresponding data are reported in Table 1. We have $R = S = 4$, $\mathrm{N}(\mathbf{p}) = \{(v_X \times v_Y) : v_X = v_Y = \{i\}, i = 1, \ldots, 4\}$ and for all $(v_X \times v_Y) \in N(\mathbf{p})$, $\frac{|v_X|}{R} + \frac{|v_Y|}{S} = 1/4 + 1/4 < 1$. This example falls under case 1 of Theorem 1. As expected, we are able to find a representative element $\bar{\mathbf{p}}$ with uniform margins $1/4$ and a corresponding discrete copula $\tilde{C}_1$ defined on the uniform grid $I_4^2$. The results are reported in Table 2. Fig. 1 displays the bubble plots for this example, a graphical representation of the probability mass function (pmf) for the observed data and the transformed pmf obtained through IPF. In the right-hand panel, exploiting the IPF algorithm, we have removed the effects of marginal non-homogeneity and the red plot unveils the symmetry structure, which is masked in the blue panel by the different values of the marginal frequencies. The role of the frequencies in the right-hand panel, i.e., to unveil symmetry beyond marginal non-homogeneity, is the same role of the residuals in the log-linear quasi-symmetry model in Eq. (1). Nevertheless, the graphical representation of the discrete copula is immediate to read and provides an easy way to detect symmetries in the data table. In this example, we can observe that the bubbles in symmetric cells have nearly the same radius, and thus the data table has a strong symmetry.

The second example is taken from (3), and the data are displayed in Table 3. The table here differ from previous case, as here we do not have a square table. The data concern the relationship between the locular composition (the number of locules of the ovary with odd or even numbers of ovules) and radial symmetry (root mean square deviation of the number of ovules from the mean number in the individual ovary) for the fruit of the American Bladder Nut, *Staphylea trifolia*, which has three locules per ovary. Some of the combinations are biologically impossible, which implies that the table has structural zeros. Note that after a reordering of the rows and columns, the observed table in Table 3 can be split into two separate complete sub-tables without structural zeros. This approach has been developed in, e.g., (5). We have retained the original structure since ordinal random variables are relevant in the copula framework.

This example falls under the assumption of case 2 in Theorem 1. Indeed, we have $R = 4$, $S = 9$ and with $v(X) = \{1, 4\}$, $v(Y) = \{2, 3, 5, 6, 8\}$ we get a sub-table with structural zeros with $|v(X)|/R + |v(Y)|/S = 2/4 + 5/9 > 1$. Therefore, the existence of a unique discrete copula is not guaranteed. Running the IPF algorithm on the original table, we obtain the probabilities in Table 4, top panel, where only the row variable is uniform. Running the IPF algorithm on the transposed of the observed table, we obtain the probabilities in Table 4, bottom panel, where, again, only the row variable is uniform. Thus, we have two solutions. Both of them share the same odd-ratio structure of the observed table due to the properties of the IPF algorithm. However, it is not possible to construct a unique discrete copula on $I_4 \times I_9$ (or on $I_9 \times I_4$) which is associated to the table. This implies that the identifiability issue of a unique copula model in a discrete setting is not completely solved by the approach presented in (4). More research is needed to shed light on such cases and further inquire into other invariant properties of the discrete copulas that can be relevant for more scenarios.

In this work we have only discussed two examples. Though, the theory is far more rich and several interesting examples do not fall within the cases of Theorem 1. For instance, we have not explored here the situation where $\frac{|v_X|}{R} + \frac{|v_Y|}{S} \leq 1$ for all $(v_X \times v_Y) \in N(\mathbf{p})$, but there are $\tilde{v}_X, \tilde{v}_Y$ such that

DRAFT

Table 2: IPF output of the data reported in Table 1 in the left panel, and corresponding discrete copula in the right panel

| | Son | | | | |
|---|---|---|---|---|---|
| Father | 1 | 2 | 3 | 4 | Total |
| 1 | 0.0000 | 0.1069 | 0.0739 | 0.0693 | 0.25 |
| 2 | 0.1040 | 0.0000 | 0.0757 | 0.0703 | 0.25 |
| 3 | 0.0666 | 0.0730 | 0.0000 | 0.1105 | 0.25 |
| 4 | 0.0794 | 0.0702 | 0.1004 | 0.0000 | 0.25 |
| Total | 0.25 | 0.25 | 0.25 | 0.25 | 1 |

$$\tilde{C}_1 = \begin{pmatrix} 0 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0 & 0.00 & 0.11 & 0.18 & 0.25 \\ 0 & 0.10 & 0.21 & 0.36 & 0.50 \\ 0 & 0.17 & 0.35 & 0.50 & 0.75 \\ 0 & 0.25 & 0.50 & 0.75 & 1.00 \end{pmatrix}$$

Table 3: Locular composition versus radial symmetry, from (3). Observed data in the top panel dashes denote structural zeros), IPF output in the central panel, and IPF output of the transposed table in the bottom panel.

| | Radial symmetry | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Locular c. | .00 | .47 | .82 | .94 | 1.25 | 1.41 | 1.63 | 1.70 | 1.89 | Total |
| 3 even 0 odd | 462 | – | – | 130 | – | – | 2 | – | 1 | 595 |
| 2 even 1 odd | – | 614 | 138 | – | 21 | 14 | – | 1 | – | 788 |
| 1 even 2 odd | – | 4413 | 95 | – | 22 | 8 | – | 5 | – | 4543 |
| 0 even 3 odd | 103 | – | – | 35 | – | – | 1 | – | 0 | 139 |
| Total | 565 | 5027 | 233 | 165 | 43 | 22 | 3 | 6 | 1 | 6065 |

DRAFT

Table 4: Locular composition versus radial symmetry, from (3). IPF output in the top panel, and IPF output of the transposed table in the bottom panel.

| | Radial symmetry | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Locular c. | .00 | .47 | .82 | .94 | 1.25 | 1.41 | 1.63 | 1.70 | 1.89 | Total |
| 3 even 0 odd | 0.0451 | 0.0000 | 0.0000 | 0.0401 | 0.0000 | 0.0000 | 0.0259 | 0.0000 | 0.1111 | 0.2222 |
| 2 even 1 odd | 0.0000 | 0.0208 | 0.0785 | 0.0000 | 0.0681 | 0.0826 | 0.0000 | 0.0277 | 0.0000 | 0.2778 |
| 1 even 2 odd | 0.0000 | 0.0903 | 0.0326 | 0.0000 | 0.0430 | 0.0285 | 0.0000 | 0.0834 | 0.0000 | 0.2778 |
| 0 even 3 odd | 0.0660 | 0.0000 | 0.0000 | 0.0710 | 0.0000 | 0.0000 | 0.0852 | 0.0000 | 0.0000 | 0.2222 |
| Total | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 0.1111 | 1 |

| | Locular composition | | | | |
|---|---|---|---|---|---|
| Radial s. | 3 even 0 odd | 2 even 1 odd | 1 even 2 odd | 0 even 3 odd | Total |
| .00 | 0.0507 | 0.0000 | 0.0000 | 0.0743 | 0.125 |
| .47 | 0.0000 | 0.0188 | 0.0812 | 0.0000 | 0.100 |
| .82 | 0.0000 | 0.0707 | 0.0293 | 0.0000 | 0.100 |
| .94 | 0.0451 | 0.0000 | 0.0000 | 0.0799 | 0.125 |
| 1.25 | 0.0000 | 0.0613 | 0.0387 | 0.0000 | 0.100 |
| 1.41 | 0.0000 | 0.0744 | 0.0256 | 0.0000 | 0.100 |
| 1.63 | 0.0292 | 0.0000 | 0.0000 | 0.0958 | 0.125 |
| 1.70 | 0.0000 | 0.0249 | 0.0751 | 0.0000 | 0.100 |
| 1.89 | 0.1250 | 0.0000 | 0.0000 | 0.0000 | 0.125 |
| Total | 0.25 | 0.25 | 0.25 | 0.25 | 1 |

$\frac{|\tilde{v}_X|}{R} + \frac{|\tilde{v}_Y|}{S} = 1$. This case happens for instance in the framework of triangular tables, which are often observed in applications. A complete description of the copulas with different patterns of structural zeros will be the subject of further research.

## References


[1] de Amo, E., Díaz Carrillo, M., Durante, F., Fernández Sánchez, J.: Extensions of subcopulas. J. Math. Anal. **452**(1), 1–15 (2017)

[2] De Loera, J.A., Kim, E.D.: Combinatorics and geometry of transportation polytopes: An update. In: Discrete Geometry and Algebraic Combinatorics, pp. 37–76. American Mathematical Society, Providence, RI (2014)

[3] Fienberg, S.E.: The analysis of incomplete multi-way contingency tables. Biometrics **28**(1), 177–202 (1972)

[4] Geenens, G.: Copula modeling for discrete random vectors. Depend. Model. **8**(1), 417–440 (2020)

[5] Goodman, L.A.: The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing cells. J. Amer. Statist. Ass. **63**, 1091–1131 (1968)

[6] Perrone, E.: Polytopes of discrete copulas and applications. In: L.A. García-Escudero, A. Gordaliza, A. Mayo, M.A. Lubiano Gomez, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (eds.) Building Bridges between Soft and Statistical Methodologies for Data Science, pp. 319–325. Springer International Publishing, Cham (2023)

[7] Perrone, E., Durante, F.: Extreme points of polytopes of discrete copulas. In: Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP), pp. 596–601. Atlantis Press (2021)

[8] Perrone, E., Solus, L., Uhler, C.: Geometry of discrete copulas. J. Multivar. Anal. **172**, 162–179 (2019)

[9] Rudas, T.: Lectures on categorical data analysis. Springer (2018)

[10] Saigusa, Y., Fukumoto, N., Nakagawa, T., Tomizawa, S.: Measure of departure from conditional partial symmetry for square contingency tables. J. Math. Stat. **18**, 138–142 (2022). DOI 10.3844/jmssp.2022.138.142

[11] Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de Paris **8**, 229–231 (1959)


DRAFT

# Levels Merging in the Latent Class Model

Christophe Biernacki[a]

[a]Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France
christophe.biernacki@inria.fr,

**Abstract**

The latent class model (LCM), dedicated to cluster categorical variables, suffers for the curse of dimension when the number of levels is large, situation frequently encountered in practice. We propose to extent LCM to a natural modeling which limits the number of levels by merging them, process which is also equivalent to a specific levels clustering. Related estimation and model selection processes are also presented and discussed.

***Keywords:*** Mixture models, categorical data, numerous levels, model selection

DRAFT

## 1. Introduction

Cluster analysis is one of the main data analysis method. It aims at partitioning into $K$ groups a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ composed by $n$ individuals and lying in a space $\mathcal{X}$ of dimension $d$. This partition is denoted by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, lying in a space $\mathcal{Z}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ is a vector of $\{0,1\}^K$ such that $z_{ik} = 1$ if individual $\mathbf{x}_i$ belongs to the $k$th group, and $z_{ik} = 0$ otherwise ($i = 1, \ldots, n$, $k = 1, \ldots, K$). Model-based clustering allows to reformulate cluster analysis as a well-posed estimation problem both for the partition $\mathbf{z}$ and for the number $K$ of groups. See for instance (13) among many other references on this topic.

Thanks to the nice mathematical background it provides, model-based clustering has led also to numerous and significant practical successes, even in some challenging situations. It is typically the case for the high dimensional context ("large" $d$ value), which is very pregnant in applications such as transcriptomics, image analysis, text analysis, and so on. We can for instance refer to (5) and references therein. However, in the categorical data case, meaning that $\mathcal{X}$ corresponds to $d$ categorical variables, the $j$th one having $m_j$ response levels, another kind of dimensionality curse can appear as soon as some $m_j$s are large. This situation appears in many practical situations also, for instance for features related to the states in the US, to the French departments, to the professional social categories, to medical nomenclature, and so on. But, existing modelings as the standard latent class model (LCM, (9)), are not really adapted to this problematic situation. As far as we know, this question is seldom addressed until now, except some early work in (11) (see a description later in the paper) and also a current practice which usually simplifies data set at hand by merging manually some levels before the clustering step itself. In this work, we thus propose to formalize this empirical procedure by a specific variant of LCM which includes natively a levels merging process, which can also be equivalently re-interpret like a levels partioning process, as we will explain later in the paper. Our proposal can also be seen as an extension of (11).

We will present this LCM extension in Section 2, before to estimate parameters (Section 3) and to select the levels partitioning re-interpreted as a model selection process (Section 4). Section 5. presents preliminary numerical experiments and draws possible extensions of this preliminary work.

## 2. A Latent Class Model (LCM) in case of numerous levels

### 2.1 The classical LCM and variants

We consider data sets $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, each $\mathbf{x}_i$ containing $d$ categorical variables, the $j$th having $m_j$ response levels. The coding $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ indicates that $x_i^{jh} = 1$ if $i$ has response level $h$ for variable $j$ and $x_i^{jh} = 0$ otherwise. The standard model for clustering observations described through categorical variables is the so-called latent class model (see for instance (9)). Data are assumed to arise independently from a mixture of $K$ multivariate multinomial distributions with pdf

$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}, \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ denotes the vector parameter of the latent class model to be estimated, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$, $\alpha_k^{jh}$ denoting the probability that variable $j$ has level $h$ if object $i$ is in cluster $k$. Thus, the latent class model assumes that the variables are *conditionally independent* knowing the latent groups.

The number of parameters for LCM is equal to $K - 1 + K \sum_{j=1}^{d} (m_j - 1)$. For limiting the effect of dimension $d$, some approaches proposed parsimonious variants including for instance variable selection or variable clustering (see (5) for an overview). However, the effect of $d$ and the $m_j$s are similar on the number of parameters count, thus it makes sense also to limit the number of levels for each variable. Less works are focused in this direction, except an early one of (11) which proposes four parsimonious versions of LCM. They correspond to an extension of the parameterization of Bernoulli distributions used by (7) for clustering and also by (1) for kernel discriminant analysis. The basic idea is to impose the vector $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ to take the form $(\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j)$ with $\gamma_k^j > \beta_k^j$. Since $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$, we have $(m_j - 1)\beta_k^j + \gamma_k^j = 1$ and, consequently, $\beta_k^j = (1 - \gamma_k^j)/(m_j - 1)$. The constraint $\gamma_k^j > \beta_k^j$ becomes finally $\gamma_k^j > 1/m_j$. It leads to a very parsimonious (only $K - 1 + Kd$ parameters) but poor flexible model, and three other still more parsimonious and less flexible variants are also proposed. Thus, there is a need for a new modeling allowing to tune more finely the trade-off between parsimony and flexibility concerning the number of free levels by variable.

### 2.2 A parsimonious LCM by levels merging (LCM-LM)

We propose to follow the previous idea of constraining some levels to share the same parameter value, but we drastically extend it. More precisely, for each variable $j = 1, \ldots, d$, we define $L_j$ clusters of levels through a vector $\mathbf{w}_j = (\mathbf{w}_j^1, \ldots, \mathbf{w}_j^{m_j})$. Here $\mathbf{w}_j^h = (w_j^{h1}, \ldots, w_j^{hL_j})$ is a binary vector such $w_j^{h\ell} = 1$ if the level $h$ of the variable $j$ belongs to the level cluster $\ell$ ($\ell = 1, \ldots, L_j$), and $w_j^{h\ell} = 0$ otherwise. The LCM parameters are linked to this levels clustering $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$ as follows: $\alpha_k^{jh} = \beta_k^{j\ell} / (\sum_{h'=1}^{m_j} w_j^{h'\ell}) \propto \beta_k^{j\ell}$ if $w_j^{h\ell} = 1$, where $\sum_{\ell=1}^{L_j} \beta_k^{j\ell} = 1$. The new parameters of this constraint multinomial model are then gathered in $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, with $\boldsymbol{\beta}_k = (\beta_k^{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots, L_j)$. Denoting by $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\beta})$ the new mixture parameter, the pdf of this novel LCM is expressed as

$$\mathrm{p}(\mathbf{x}_i | \mathbf{w}; \boldsymbol{\vartheta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{\ell=1}^{L_j} \left( \frac{\beta_k^{j\ell}}{\sum_{h=1}^{m_j} w_j^{h\ell}} \right)^{\sum_{h=1}^{m_j} w_j^{h\ell} x_i^{jh}}. \tag{2}$$

The number of parameters associated to this constraint model is equal to $K - 1 + K \sum_{j=1}^{d} (L_j - 1)$, thus showing a varying parsimony degree depending on the numbers of clusters $L_j$s. We call hereafter this constraint LCM as LCM-LM for *Latent Class Model by Levels Merging*.

**Remark** Notice that the extreme parsimonious version of LCM-LM where $L_j = 1$ could be interpreted as having no effect of variable $j$ on the clustering. Consequently, it is also a model where variable $j$ has no effect on the clustering, playing the role of a specific variable selection process.

DRAFT

# 3. Parameter estimation of LCM-LM

## 3.1 Reformulating the log-likelihood of LCM-LM as a classical LCM

We propose to estimate $\boldsymbol{\vartheta}$ through $\hat{\boldsymbol{\vartheta}}_{\mathbf{w}}$ which maximizes the observed log-likelihood

$$L(\boldsymbol{\vartheta}|\mathbf{w};\mathbf{x}) = \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{\ell=1}^{L_j} \left( \frac{\beta_k^{j\ell}}{\sum_{h=1}^{m_j} w_j^{h\ell}} \right)^{\sum_{h=1}^{m_j} w_j^{h\ell} x_i^{jh}} \right). \tag{3}$$

But it is of practical interest to notice at this step that the log-likelihood $L(\boldsymbol{\vartheta}|\mathbf{w};\mathbf{x})$ of LCM-LM can be reformulated as the log-likelihood $L(\boldsymbol{\vartheta}|\mathbf{w};\mathbf{y})$[1] of a classical LCM where the levels of the same cluster $\ell$ are merged into a unique new level with an additional term $c_{\mathbf{w}}$ independent of the parameter $\boldsymbol{\vartheta}$. Indeed, we can write

$$\begin{aligned}
L(\boldsymbol{\vartheta}|\mathbf{w};\mathbf{x}) &= \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{\ell=1}^{L_j} \left( \beta_k^{j\ell} \right)^{y_{\mathbf{w},i}^{jl}} \right) + c_{\mathbf{w}} && (4)\\
&= \sum_{i=1}^{n} \ln \left( \sum_{k=1}^{K} \pi_k \mathrm{p}(\mathbf{y}_{\mathbf{w},i};\boldsymbol{\beta}_k) \right) + c_{\mathbf{w}} && (5)\\
&= \sum_{i=1}^{n} \ln \mathrm{p}(\mathbf{y}_{\mathbf{w},i};\boldsymbol{\vartheta}) + c_{\mathbf{w}} && (6)\\
&= L(\boldsymbol{\vartheta};\mathbf{y}_{\mathbf{w}}) + c_{\mathbf{w}} && (7)
\end{aligned}$$

where $c_{\mathbf{w}} = -\sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{\ell=1}^{L_j} y_{\mathbf{w},i}^{j\ell} \ln \left( \sum_{h=1}^{m_j} w_j^{h\ell} \right)$ is a constant with regards to the parameter $\boldsymbol{\vartheta}$ and where $y_{\mathbf{w},i}^{j\ell} = \sum_{h=1}^{m_j} w_j^{h\ell} x_i^{jh}$, $\mathbf{y}_{\mathbf{w},i} = (y_{\mathbf{w},i}^{j\ell}; j = 1,\ldots,d; \ell = 1,\ldots,L_j)$ and $\mathbf{y}_{\mathbf{w}} = (\mathbf{y}_{\mathbf{w},1},\ldots,\mathbf{y}_{\mathbf{w},n})$.

## 3.2 EM algorithm description

As a consequence of the preeceeding remark, estimating $\boldsymbol{\vartheta}$ can be performed by any classical EM algorithm (8) for LCM implemented in any classical existing package (for instance Rmixmod (11) or RMixtComp[2]). Concretely, starting from $\boldsymbol{\vartheta}^0 = (\boldsymbol{\pi}^0, \boldsymbol{\beta}^0)$, iteration $r > 0$ of this EM is expressed as:

**E step:** Conditional probability that individual $i$ arose from cluster $k$

$$t_{ik}(\boldsymbol{\vartheta}^r) = \frac{\pi_k^r \mathrm{p}(\mathbf{y}_{\mathbf{w},i};\boldsymbol{\beta}_k^r)}{\mathrm{p}(\mathbf{y}_{\mathbf{w},i};\boldsymbol{\vartheta}^r)}. \tag{8}$$

**M step:** Updating the mixture parameter estimates by maximizing the expected complete log-likelihood

$$\pi_k^{r+1} = \frac{\sum_i t_{ik}(\boldsymbol{\vartheta}^r)}{n} \quad \text{and} \quad (\beta_k^{j\ell})^{r+1} = \frac{\sum_{i=1}^{n} t_{ik}(\boldsymbol{\vartheta}^r) y_{\mathbf{w},i}^{j\ell}}{\sum_{i=1}^{n} t_{ik}(\boldsymbol{\vartheta}^r)}. \tag{9}$$

The stopping rule can rely on a plateau of the log-likelihood (equally $L(\boldsymbol{\vartheta};\mathbf{y}_{\mathbf{w}})$ or $L(\boldsymbol{\vartheta}|\mathbf{w};\mathbf{x})$) or a given iteration number. Several runs should be performed also from different starting parameters for avoiding local maxima traps.

[1] For simplifying notations, both p and $L$ stand for *generic* pdf and log-likelihood functions, respectively.
[2] https://cran.r-project.org/web/packages/RMixtComp/index.html

DRAFT

# 4. Model selection for LCM-LM

## 4.1 Exact ICL criterion for LCM-LM

Beyond parameter estimation, there is a need also to estimate the number of individual clusters $K$ and the levels partioning $\mathbf{w}$ (which includes itself the levels clusters $L_j$s). Hereafter we omit the (implicit) index $K$ in notations for focalizing attention on the index $\mathbf{w}$ which is the novel part of this work.

Usually, the BIC criterion (16) or the ICLbic criterion[3] (3) appear to give a reasonable answer to the model selection problem in mixtures, especially the ICLbic criterion when the clustering point of view is of first interest. However, some previous works dealing with LCM (see for instance (15) or (4)) suggest that both BIC and ICLbic need particular large sample sizes to reach its expected asymptotic behaviour in practical situations. In (4), it is taken profit from the possibility to avoid the asymptotic approximation ICLbic of the observed integrated likelihood to propose an alternative non-asymptotic ICL criterion, leading to a subsequent improvement in model selection for moderate sample sizes, while having a particular easy to calculate closed-form expression of the corresponding ICL. We propose thus to adapt now this non-asymptotic ICL criterion calculation from LCM to the new modeling LCM-LM.

Denoting by $\hat{\mathbf{z}}_{\mathbf{w}}$ the maximum *a posteriori* (MAP) estimate of $\mathbf{z}$[4] and by $\Theta_{\mathbf{w}}$ the space of $\boldsymbol{\vartheta}$ when considering the model $\mathbf{w}$, the ICL criterion for LCM-LM is defined as

$$\begin{aligned}
\text{ICL}_{\mathbf{w}}^{\text{LCM-LM}} &= \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{w}}|\mathbf{w}) && (10)\\
&= \ln \int_{\Theta_{\mathbf{w}}} \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{w}}|\mathbf{w};\boldsymbol{\vartheta})\mathrm{p}(\boldsymbol{\vartheta}|\mathbf{w})d\boldsymbol{\vartheta} && (11)\\
&= \ln \int_{\Theta_{\mathbf{w}}} \mathrm{p}(\mathbf{y}_{\mathbf{w}}, \hat{\mathbf{z}}_{\mathbf{w}};\boldsymbol{\vartheta})\mathrm{p}(\boldsymbol{\vartheta}|\mathbf{w})d\boldsymbol{\vartheta} + c_{\mathbf{w}} && (12)\\
&= \text{ICL}_{\mathbf{w}}^{\text{LCM}} + c_{\mathbf{w}} && (13)
\end{aligned}$$

where we have used a similar development as in 7 for obtaining this last expression, and where $\text{ICL}_{\mathbf{w}}^{\text{LCM}}$ is the ICL criterion for LCM. This latter is classically defined as follows (see for instance (4))

$$\begin{aligned}
\text{ICL}_{\mathbf{w}}^{\text{LCM}} = \sum_{k=1}^{K}\sum_{j=1}^{d}\left\{\sum_{\ell=1}^{L_j}\ln\Gamma\left(\hat{n}_{\mathbf{w},k}^{j\ell}+\tfrac{1}{2}\right) - \ln\Gamma(\hat{n}_{\mathbf{w},k}+\tfrac{L_j}{2})\right\} - \ln\Gamma(n+\tfrac{K}{2}) + \ln\Gamma(\tfrac{K}{2})\\
+K\sum_{j=1}^{d}\left\{\ln\Gamma(\tfrac{L_j}{2}) - L_j\ln\Gamma(\tfrac{1}{2})\right\} + \sum_{k=1}^{K}\ln\Gamma(\hat{n}_{\mathbf{w},k}+\tfrac{1}{2}) - K\ln\Gamma(\tfrac{1}{2}), \qquad (14)
\end{aligned}$$

where $\hat{n}_{\mathbf{w},k} = \#\{i : \hat{z}_{\mathbf{w},ik} = 1\}$ and $\hat{n}_{\mathbf{w},k}^{j\ell} = \#\{i : \hat{z}_{\mathbf{w},ik} = 1, y_{\mathbf{w},i}^{j\ell} = 1\}$. As a consequence from 13, the $\text{ICL}_{\mathbf{w}}^{\text{LCM-LM}}$ is itself easily calculated also.

## 4.2 Labels partitioning estimation without multiple parameter estimation

Since the space $\mathcal{W}$ where lie models $\mathbf{w}$ is extremely large (including or not the number labels clusters $L_j$), it is obviously impossible to calculate the value of $\text{ICL}_{\mathbf{w}}^{\text{LCM-LM}}$ for each of them. In addition, a single evaluation of $\text{ICL}_{\mathbf{w}}^{\text{LCM-LM}}$ requires the availability of $\hat{\mathbf{z}}_{\mathbf{w}}$, thus to perform potentially a full run of the EM algorithm for evaluating $\hat{\boldsymbol{\vartheta}}_{\mathbf{w}}$.

An original strategy is proposed by (12) for avoiding parameter estimation for all models which compete, thus limiting the computing time. Then a parameter estimation is just performed for the retained model at the end of their process. Their strategy is applied to mixture models where the within component variable independence is assumed, as it the case in LCM.

[3] ICLbic corresponds to the classical asymptotic approximation of the ICL (*Integrated Complete-data Likelihood*) criterion.

[4] It means that $\hat{z}_{\mathbf{w},ik} = 1$ iif $k \in \arg\max_{k'} t_{ik'}(\hat{\boldsymbol{\vartheta}}_{\mathbf{w}})$.

DRAFT

Their strategy relies on a variant of the ICL criterion, the so-called MICL criterion (*Maximum Integrated Complete-data Likelihood*), and defined in our case by

$$\text{MICL}_{\mathbf{w}}^{\text{LCM-LM}} = \ln \text{p}(\mathbf{x}, \mathbf{z}_{\mathbf{w}}^*|\mathbf{w}) \quad \text{with} \quad \mathbf{z}_{\mathbf{w}}^* = \arg\max_{\mathbf{z}\in\mathcal{Z}} \ln \text{p}(\mathbf{x}, \mathbf{z}|\mathbf{w}).$$

Then, the model $\mathbf{w}^*$ maximizing $\text{MICL}_{\mathbf{w}}^{\text{LCM-LM}}$ is retained:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\in\mathcal{W}} \text{MICL}_{\mathbf{w}}.$$

We expect that MICL is consistent for choosing $\mathbf{w}$, at least when the cluster numbers ($K$ and $L_j$s) are known, as being an extension of the proof of (12) available in a more restricted by quite similar situation.

The question of maximizing $\text{MICL}_{\mathbf{w}}^{\text{LCM-LM}}$ on $\mathbf{w}$ is obviously the crucial difficulty. Thus, mimiking the idea of (12), we implement the following simple alternate procedure, for a fixed $K$ value (thus this algorithm has to be run for different candidate values of $K$). Starting from a value $\mathbf{w}^0$ uniformly sampled in the corresponding space and then a value $\mathbf{z}^0$ being deduced from the MAP rule of the associated maximum likelihood estimate, iteration $s > 0$ of the algorithm is composed by the following two steps:

**Partition step** Fix $\mathbf{z}^{s+1}$ such that $\ln \text{p}(\mathbf{x}, \mathbf{z}^{s+1}|\mathbf{w}^s) \geq \ln \text{p}(\mathbf{x}, \mathbf{z}^s|\mathbf{w}^s)$.
**Model step** Fix $\mathbf{w}^{s+1}$ such that $\ln \text{p}(\mathbf{x}, \mathbf{z}^{s+1}|\mathbf{w}^{s+1}) \geq \ln \text{p}(\mathbf{x}, \mathbf{z}^{s+1}|\mathbf{w}^s)$.

In practice, the partition step can be performed by sampling uniformly an individual which is affiliated to the component maximizing the integrated complete-data likelihood, while the other component memberships are unchanged (details are given in (12)). The model step can be performed in two successive phases: First, estimate all $\boldsymbol{\beta}_k$ in the case where all $L_j = m_j$ (no labels partitioning case), which is obtained in closed-form since the individual partition $\mathbf{z}^{s+1}$ is fixed; Second, different labels partition candidates $\mathbf{w}$ can be proposed thanks to a simple kmeans of $\boldsymbol{\beta}_k$, variable by variable, for different values of $L_j$s. Note that this procedure can be trapped within local maxima and thus several run are required.

DRAFT

## 5. Preliminary numerical experiments and conclusion

Here are some early numerical experiments on a real data set for evaluating the practical interest of our LCM-LM proposal. Authors in (10) (see also (14) p. 139–142) considered the clustering of patients on the basis of petrial variates alone for the prostate cancer clinical trial data of (6) which is reproduced in (2) p. 261–274. This data set was obtained from a randomized clinical trial comparing four treatments for $n = 506$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease (there are 62 missing values, so about 1% of the whole sample). As reported by (6), Stage 3 represents local extension of the disease without evidence of distance metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, X-ray evidence, or both. Twelve pre-trial variates were measured on each patient, composed by eight continuous variables and four categorical variables with various numbers of levels. We limit here our study to these latter, namely performance rating (so-called PF, with 4 levels), cardiovascular disease history (HX, 2 levels), electrocardiogram code (EKG, 7 levels) and bone metastases (BM, 2 levels). Invoking the RMixtcomp package[5], which is a clustering package able implementing LCM categorical data (and also dealing with missing values), we obtain a classification error equal to 49.2% and an ICL value equal to -1728.4 (we have fixed $K = 2$, because there are two stages of the desease).

We see below the frequency of each level for all these variables (missing values are denoted by a "?" and appear below as a level, but mathematical speaking it is not a level obviously).

```
    PF       HX           EKG       BM
?:  4   ?:  4   1      :168   ?:  4
1:450   1:289   5      :150   1:420
2: 37   2:213   6      : 75   2: 82
3: 13          3      : 51
4:  2          4      : 26
               2      : 23
               (Other): 13
```

[5] https://cran.r-project.org/web/packages/RMixtComp/index.html

As observed through this level distribution, level 4 of variable PF is under-represented, thus we propose to merge it with level 3 of the same variable. Running the corresponding LCM-LM (still through RMixtcomp) on this new data set, we obtain now a classification error equal to 28.8% and an ICL value equal to -1618.4 (we used the ICL formulation given in Section 4. in such a way that this LCM-LM ICL value can be compared to the ICL value of LCM). We thus observe that merging these two levels of PF leads to a very significant improvement of both the error rate and the ICL value, illustrating in that way all the potential interest of the proposed LCM-LM model for the clustering of categorical variables.

This paper should be considered as a preliminary work for levels clustering in the LCM context. We need now to numerically experiment the LCM-LM proposal on simulated and real data sets. We expect however that the alternating algorithm proposed for implementing the MICL criterion could be quite low. An extension could thus be to reinterpret LCM-LM within a new co-clustering paradigm where the levels partitioning should replace the classical variable partitioning involved in co-clustering. In that way, we could consider levels partitioning as a latent variable, and use in this context an efficient Variational EM algorithm for instance.

## References


[1] Aitchinson, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420.

[2] Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many. Fields for the Student and Research Worker*. Springer-Verlag.

[3] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

[4] Biernacki, C., Celeux, G., and Govaert, G. (2011). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

[5] Biernacki, C. and Maugis, C. (2017). High-dimensional clustering. In Droesbeke, J.-J., Saporta, G., and Thomas-Agnan, C., editors, *Choix de modèles et agrégation*. Technip.

[6] Byar, D. and Green, S. (1980). The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. *Bulletin du Cancer*, 67:477–490.

[7] Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176.

[8] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

[9] Goodman, L. A. (1974). Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

[10] Hunt, L. and Jorgensen, M. (1999). Mixture model clustering: a brief introduction to the multimix program. *Australian and New Zealand Journal of Statistics*, 41(2):153–171.

[11] Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and Govaert, G. (2015). Rmixmod: The r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. *Journal of Statistical Software*, in press.

[12] Marbac, M. and Sedki, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statitics and Computing*, 27:1049â1063.

[13] McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

[14] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New-York.

[15] Nadif, M. and Govaert, G. (1998). Clustering for binary data and mixture models: Choice of the model. *Applied Stochastic Models and Data Analysis*, 13:269–278.

[16] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.


DRAFT

# Model-based clustering of count processes with multiple change points

Shuchismita Sarkar[a] and Xuwen Zhu[b]

[a]Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, OH, USA; ssarkar@bgsu.edu

[b]Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL, USA; xzhu20@cba.ua.edu

**Abstract**

A model-based clustering algorithm based on finite mixture of negative binomial Lévy processes is proposed. The algorithm models heterogeneous long-time count process data that can have multiple change points. The design of the algorithm involves several stages to avoid the infeasible exhaustive search. The proposed model is applied to the COVID ICU cases in the state of California USA with promising results.

*Abstract in italiano.*

***Keywords:*** Categorical data analysis, Model-based clustering...

DRAFT

## 1. Introduction

The field of change point has a long history and most of the papers study the existence of single change point. Multiple change point estimation was first proposed by [1] and can be found in more recent literature. Model-based clustering [4] is an unsupervised learning tool for partitioning a heterogeneous population. Model-based clustering for stochastic process data in its functional form has recently received a lot of attention. Specifically, [3] proposed a novel extension of the Poisson mixture model for clustering count process data using intensity functions. To the authors' knowledge, change point estimation and model-based clustering methodologies have only been studied in the common framework by [5]. The method relies on matrix-variate mixture and exhaustively searches for a shift in the mean or variance. In this paper, our focus is on long-time count process data that can have multiple change points. An exhaustive search would be computationally infeasible. The authors take an alternative approach and search for an optimal interval at each step of the algorithm which contains a change point most likely. Then the change point is exhaustively tested in the interval. We assume a mixed type of change point where there is a gap between two logit-transformed segments.

## 2. Methodology

### 2.1 Lévy processes

A stochastic process $\boldsymbol{X} = \{X_t, t \geq 0\}$ is called a Lévy process, if it has right continuous left limited path with independent, stationary increments. The Poisson process is the most popular example of a

Lévy process used for count data. Since the Poisson distribution has a strong restriction that its mean and variance are equal, a negative binomial (NB) process with parameters $r = \alpha$ and $p = 1/(\beta + 1)$ is usually adopted with more flexibility. $p$ can be considered as the success probability that some event happened and $r$ is viewed as the number of failures until the experiment is stopped. In this paper, we consider a negative binomial time-inhomogeneous process whose increments are independent, but not stationary.

## 2.2 Finite mixture model of negative binomial process

Assume $\underline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ is an observed random sample from a mixture model

$$g(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{y}; \boldsymbol{\Psi}_k) \tag{2.1}$$

with $K$ components. Time interval $(t_0, t_T]$ is split into small intervals of equal length $t$ given by $\{(t_0, t_1]$, $\ldots$, $(t_{T-1}, t_T]\}$. Let $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{T}$ denote a count process associated to subject $i$ where $y_{ij}$ is the number of events happening in interval $(t_{j-1}, t_j]$. We assume that $f_k(\boldsymbol{y}; \boldsymbol{\Psi}_k)$ from Equation 2.1 is a negative binomial (NB) process with the following form

$$f_k(\boldsymbol{y}; \boldsymbol{\Psi}_k) = \prod_{j=1}^{T} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_k)^{r_k} p_k^{y_{ij}}, \tag{2.2}$$

with parameters $r_k (\geq 0)$ and $p_k$ $(0 \leq p_k \leq 1)$. The complete-data likelihood function is given by:

$$\mathcal{L}_c(\boldsymbol{\Psi}; \underline{\boldsymbol{y}}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left( \tau_k \prod_{j=1}^{T} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_k)^{r_k} p_k^{y_{ij}} \right)^{I(Z_i = k)}, \tag{2.3}$$

where $Z_i$ denotes the unknown membership of the $i^{th}$ count process. We model the changes in process through the success probability $(p_k)$. During the observed time interval $(t_0, t_T]$, suppose there are $m_k$ change points $(0 \leq m_k < T)$ denoted as $\zeta_{k1}, \ldots, \zeta_{km_k}$ and the shift occurs at the beginning of the $m_k$ sub intervals of the form $(t_{j-1}, t_j]; j = 1, 2, \ldots, T-1$. The change in the process is introduced through the change in the logit of $(1 - p_k)$ as follows

$$\log\left(\frac{1 - p_k}{p_k}\right) = \begin{cases} \eta_{k0} + \nu_{k0} j, & \text{if } \zeta_{k0} < j \leq \zeta_{k1} \\ \eta_{k1} + \nu_{k1} j, & \text{if } \zeta_{k1} < j \leq \zeta_{k2} \\ \vdots & \\ \eta_{km} + \nu_{km} j, & \text{if } \zeta_{km} < j \leq \zeta_{k(m_k+1)}. \end{cases} \tag{2.4}$$

The logit of $(1 - p_k)$ is assumed to be a linear function of time. The parameters associated to this function indicate the change in the process.

Denote $p_{kl}$ as the success probability in the interval $(\zeta_{kl}, \zeta_{k(l+1)}]$. The complete data likelihood becomes

$$\mathcal{L}_c(\boldsymbol{\Psi}; \underline{\boldsymbol{y}}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left( \tau_k \prod_{l=0}^{m_k} \prod_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}} \binom{r_k + y_{ij} - 1}{y_{ij}} (1 - p_{kl})^{r_k} p_{kl}^{y_{ij}} \right)^{I(Z_i = k)}. \tag{2.5}$$

The posterior probabilities are given by

$$\ddot{\pi}_{ik} = \frac{\dot{\tau}_k f_k(\boldsymbol{y}_i; \dot{\boldsymbol{\Psi}}_k)}{\sum_{k'=1}^{K} \dot{\tau}_{k'} f_{k'}(\boldsymbol{y}_i; \dot{\boldsymbol{\Psi}}_{k'})} \tag{2.6}$$

DRAFT

in the E-step. In the M-step, the mixing proportion $\ddot{\tau}_k$ is obtained as $\frac{1}{n}\sum_{i=1}^{n}\ddot{\pi}_{ik}$. The maximum likelihood estimates of $\eta_{kl}$ and $\nu_{kl}$ can be found by numerically maximizing

$$Q_{kl} = \sum_{i=1}^{n} \ddot{\pi}_{ik} \sum_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}} (\dot{r}_k \log(1-\dot{p}_{kl}) + y_{ij}\log(\dot{p}_{kl})). \tag{2.7}$$

For estimating $r_k$

$$\begin{aligned} Q_k = &\sum_{i=1}^{n} \ddot{\pi}_{ik} \Bigg( \sum_{j=1}^{T} \log \binom{\dot{r}_k + y_{ij} - 1}{y_{ij}} \\ &+ \sum_{l=0}^{m_k} \sum_{j=\zeta_{kl}+1}^{\zeta_{k(l+1)}} \dot{r}_k \log(1-\ddot{p}_{kl}) \Bigg) \end{aligned} \tag{2.8}$$

needs to be numerically maximized. If there is no change point associated to the $k^{th}$ component *i.e.* if $m_k = 0$, Equation 2.5 becomes equivalent to Equation 2.3.

## 2.3 Algorithm for finding multiple change points

For change point estimation in a time series with moderate length, an exhaustive search at each time point is extremely time consuming and maybe computationally infeasible. Employing a sequential approach for multiple change point estimation has been suggested by the studies of [2]. Application of segmentation algorithm that involves partitioning the entire time series into smaller pockets [3] is particularly useful in this context.

In the beginning, the list of change points $\mathcal{CP}list$ is $NULL$ and a null model $\mathcal{M}_{null}$ with no change point is fitted according to Equation 2.3. The observed interval is partitioned into smaller intervals of length $h$ with break points $\{b_0, \ldots, b_d\}$ where $b_0 = t_0$, $b_d = t_T$ and $d = \lfloor T/h \rfloor + 1$. Here, the $\lfloor . \rfloor$ operator returns the integer part of a real number. For each mixture component, $d-2$ models according to Equation 2.5 are fitted with the assumption that there exist change points at$\{b_i, b_{i+1}\} \cup \mathcal{CP}list$ where $1 \leq i \leq d-2$. The best $c$ models ($c$ is user specified) among these $d-2$ candidates are chosen based on BIC value. The intervals associated to these $c$ models are combined to get a favorable interval ($\mathcal{I}_{fav}$). An exhaustive search is performed for each time point within the interval which involves fitting $c(h+1)$ models with the assumption that change points exist at $\{cp\} \cup \mathcal{CP}list$ where $cp \in \mathcal{I}_{fav}$. If the BIC value for a fitted model is less than the BIC value of $\mathcal{M}_{null}$, the associated time point $\mathcal{CP}_{est}$ is added to $\mathcal{CP}list$ for the specific mixture component and the corresponding model is denoted as $\mathcal{M}_{current}$. This process is repeated for each component. Then a search for the next set of change points is conducted until all searches yield higher BIC.

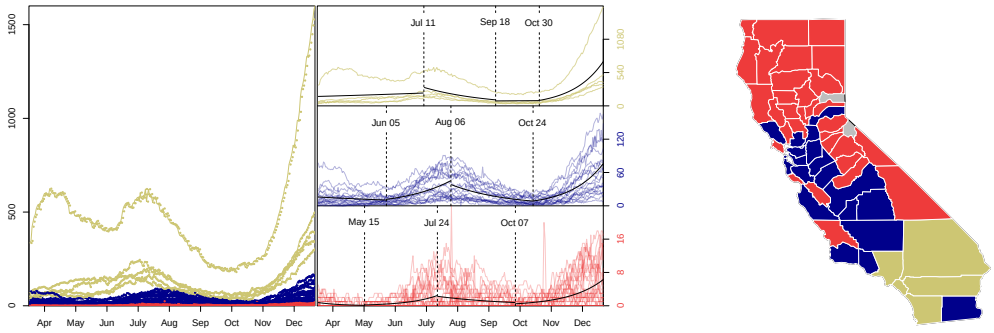
## 2.4 Computational issues

In this paper, we use k-means with one iteration to initialize the algorithm. Several random starts are run till convergence and the solution with the minimum BIC is adopted as the null model. During the search for change points, each alternative model is initialized by the model adopted in the previous step so that the iteration required for convergence is much fewer than if initialized randomly. The computational time for analyzing one dataset (in the following application section) with $n = 56$ and $T = 275$ is 50327.706 seconds (14 hours).

# 3. Application

We apply the proposed methodology to the California county COVID-19 daily ICU data which includes a total of 275 days in 2022. Models are fitted for $K = 1, 2, 3, \ldots$ components and $\hat{K} = 3$ is the

DRAFT




(a) (b)

Figure 1: (a) Grouped daily ICU positive patients curves for 56 counties in California, (b) 3-cluster solution presented on California map.

chosen mixture order based on its lowest entropy value [1]. Figure 1 (a) demonstrates the obtained partition on the left. The khaki group has the highest daily ICU positive patients, followed by the blue one. The red group has the lowest ICU counts. The corresponding geographical location of these counties are shown in California map in Figure 1 (b). South California counties are mostly khaki with highest positive patients. The blue group along the west coast including the Bay area has moderate control of cases. The North California area, which consists of red counties, has the lowest transmission rate.

The change points estimated by the algorithm are reflected by dashed lines in Figure 1 (a). The solid black lines represent mean profiles fitted by the model. There are three change points detected in each group and they happen at different times across different groups. Generally speaking, the numbers are stable (slightly dipped for the red and blue groups) due to regulations and stay-at-home orders from the state governor. After those orders get lifted, the numbers increased for all three groups during May – July. The state slowly reopened public places in an attempt to recover from economy depression. During the mid of summer, several close-down orders were re-imposed.

In khaki counties, this results in a decline in numbers starting from July 11. On August 28, the state governor released the "Blueprint for a Safer Economy" guideline which classified all counties into 4 tiers based on the number of daily cases and current test positivity rate. The assignment can move to a lower or higher tier if recent numbers change. Re-openings will be permitted according to the assigned category. This smart movement effectively controlled the numbers during September – October. We can clearly see an increasing trend for all three groups towards the end of the year starting from October, which is probably driven by the holiday season and the presidential election.

DRAFT

## References

[1] J. Ng, T. B. Murphy, Model-based clustering of count processes, Journal of Classification (2020). doi:10.1007/s00357-020-09363-4.

[2] L. C. Zhao, P. R. Krishnaiah, Z. D. Bai, On detection of the number of signals in presence of white noise, Journal of Multivariate Analysis 20 (1986a) 1-25.

[3] P. Fryzlewicz, Wild binary segmentation for multiple change-point detection, Annals of Statistics 42 (6) (2014) 155 2243-2281.

[4] V. Melnykov, Challenges in model-based clustering, WIREs: Computational Statistics 5 (2013) 135-148.

[5] X. Zhu, Y. Melnykov, On finite mixture modeling of change-point processes, Journal of Classification (2021) 39 3-22.

# Similarity Measures and Internal Evaluation Criteria in Hierarchical Clustering of Categorical Data

Jana Cibulková[a], Zdeněk Šulc[a], Hana Řezanková[a], and Jaroslav Horníček[a]

[a] Prague University of Economics and Business, Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic;
jana.cibulkova@vse.cz, zdenek.sulc@vse.cz, hana.rezankova@vse.cz, jaroslav.hornicek@vse.cz

**Abstract**

The contribution focuses on the hierarchical clustering of objects characterized by categorical variables and on the evaluation of the obtained results by selected internal criteria. The aim is to assess the dependence of the evaluation criteria on the dataset properties and the used similarity measure. In the experiment, 270 generated datasets (with the controlled number of variables, between-cluster distances, and the number of clusters) are analysed using the average linkage method and twelve similarity measures. Moreover, the correlation of obtained values of the five examined internal criteria is studied as well. We apply the ANOVA method to assess the dependence of evaluation criteria on the dataset parameters and the similarity measure. We found that the number of clusters influences the dependence intensity of internal criteria on the type of similarity measure. The dependence intensity of internal criteria on the dataset properties varies greatly according to the particular dataset parameter and the type of evaluation criteria.

**Keywords:** categorical data, cluster analysis, similarity measures, evaluation criteria

DRAFT

## 1. Introduction

For the investigation of the data structure, cluster analysis is often used. This type of data analysis includes many different techniques. The most frequent task of clustering is the identification of groups (clusters) of similar objects, whereas each object is represented by a vector of variables. The usual input for analyses is a data matrix in which each row characterizes one object, and each column represents one variable. Methods of cluster analysis can be classified according to what types of variables they can handle.

This contribution focuses on the analysis of datasets with categorical variables whose values are on a nominal scale, i.e., unordered. For this type of data, several kinds of methods were proposed. In our research, we investigate methods of hierarchical cluster analysis, which are based on a proximity matrix with the distances for each pair of objects. In the case of categorical data, dissimilarities are used instead of distances. In the further text, we will use the term "distance" in a broader sense, including also dissimilarity.

The relationships of two objects characterized by values of categorical variables are usually expressed by a similarity, which is transformed into the distance for the purpose of cluster analysis. Recently, different similarity measures were proposed for categorical data. Šulc and Řezanková (2019) proposed two new similarity measures and compared them with similarity measures proposed by some other authors in terms of the quality of the obtained clusters. All these measures are implemented in the *nomclust* R package (Šulc et al., 2022) in the R environment (R Core Team, 2020).

The main aim of this contribution is to assess the dependence of the selected internal evaluation criteria on the properties of the datasets and the used similarity measure. The internal evaluation criteria

are usually constructed to satisfy the principles of compactness and separation of the obtained clusters (e.g., Liu et al., 2010). We also investigate dependencies and differences between the examined internal criteria.

## 2. Theoretical Background of Experiment

In this section, we describe the way of dataset generation, cluster analysis using different similarity measure, and used internal evaluation criteria. The generated datasets for the experiment were obtained using the *genRandomClust* function from the *clusterGeneration* R package (Qiu and Joe, 2020) and the discretize function from the *arules* R package (Hahsler et al., 2021). The generation process is realized in two steps. In the first step, a dataset with quantitative variables and a multidimensional correlation structure reflecting the studied dataset properties is created. In the second step, the variables are categorized using the equal-width intervals approach.

### 2.1 Hierarchical Cluster Analysis and Similarity Measures for Categorical Data

For cluster analysis, we applied the average linkage method and twelve similarity measures for categorical data. The average linkage method was chosen, since it avoids the extremes of either large or tight compact clusters and of unwanted chain effect. There are the Goodall 1 (G1), Goodall 2 (G2), Goodall 3 (G3), and Goodall 4 (G4) based on the original Goodall measure (Goodall, 1966), and Lin 1 (LIN1), which were introduced by Boriah et al. (2008). The names of the other measures are Eskin (ES) proposed by Eskin et al. (2002), Lin (LIN) introduced by Lin (1998), the simple matching coefficient (SM) proposed by Sokal and Michener (1958), the measures occurrence frequency (OF) and inverse occurrence frequency (IOF) introduced by Spärck Jones (1972), and the measures variable entropy (VE) and variable mutability (VM) proposed by Šulc and Řezanková (2019).

When two objects are compared, for each variable, two values for the examined two objects are compared. In the case of the categorical data, we can only distinguish whether the two values are the same or not. Thus, for the expression of object similarity, two certain values (e.g. 1 or 0) or ways of their calculations are given: one for the equality, and the second for the inequality of the variable values.

For a certain similarity measure, the relationship of the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be computed either as

$$S(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}\,, \tag{1}$$

or as

$$S(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m} \left(\ln p(x_{ic}) + \ln p(x_{jc})\right)}\,, \tag{2}$$

where $c$ is the index of a certain variable, $S_c$ is the comparative value for the $c$th variable, $x_{ic}$ is the value for the $i$th object in the $c$th variable, $m$ is the number of variables, and $p(x_{ic})$ is the relative frequency of the value $x_{ic}$ (similarly for $x_{jc}$).

For the similarity measures that take on values from zero to one, the distance between the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is computed as

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - S(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{3}$$

For the similarity measures whose maximum is lower than one, the distance between the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is calculated as

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{S(\boldsymbol{x}_i, \boldsymbol{x}_j)} - 1. \tag{4}$$

DRAFT

## 2.2 Internal Evaluation Criteria for Clustering Results with Categorical Data

For the evaluation of clustering results, we used five internal criteria, two of them are based on the comparison of within-cluster and between-cluster variability, two criteria are based on the likelihood, and the last criterion is based on the distances. Distance based criteria are usually applied when quantitative data are analysed. Their obtained values depend on the used type of the distance measure.

### 2.2.1 Variability-based Criteria

When data with categorical variables are analysed, special variability measure should be applied. There are usually the Gini measure (mutability) and entropy. Mutability of the whole dataset (the total mutability) is expressed as

$$TM = \sum_{c=1}^{m}\left(1 - \sum_{u=1}^{K_c}\left(\frac{n_{cu}}{n}\right)^2\right), \tag{5}$$

where $c$ is the index of a certain variable, $m$ is the number of variables, $u$ is the index of a certain category (a certain value from the set of unique values of the analysed variable), $K_c$ is the number of categories of the $c$th variable, $n_{cu}$ is the number of values equal to the $u$th category of the $c$th variable, and $n$ is the number of objects in the dataset. Entropy of the whole dataset (the total entropy) is then expressed as

$$TE = \sum_{c=1}^{m}\left(-\sum_{u=1}^{K_c}\left(\frac{n_{cu}}{n}\ln\frac{n_{cu}}{n}\right)\right). \tag{6}$$

The within-cluster mutability for $k$-cluster solution of clustering is calculated as

$$WCM(k) = \sum_{g=1}^{k}\frac{n_g}{n}\sum_{c=1}^{m}\left(1 - \sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\right)^2\right), \tag{7}$$

where $g$ is the index of a certain cluster, $n_g$ is the number of objects in the $g$th cluster, and $n_{gcu}$ is the number of values equal to the $u$th category of the $c$th variable in the $g$th cluster. Similarly, the within-cluster entropy for $k$-cluster solution is calculated as

$$WCE(k) = \sum_{g=1}^{k}\frac{n_g}{n}\sum_{c=1}^{m}\left(-\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right)\right). \tag{8}$$

On the basis of the formulas mentioned above, the pseudo F coefficients (for $k$ clusters) inspired by Milligan and Cooper (1987) can be expressed either as the pseudo F coefficient based on mutability

$$PSFM(k) = \frac{(n-k)(TM - WCM(k))}{(k-1)WCM(k)} \tag{9}$$

or as the pseudo F coefficient based on mutability

$$PSFE(k) = \frac{(n-k)(TE - WCE(k))}{(k-1)WCE(k)}\ . \tag{10}$$

These modifications of the pseudo F coefficient were proposed by Řezanková et al. (2011). The maximal value across all the examined cluster solutions suggests the optimal number of clusters for both criteria.

### 2.2.2 Likelihood-based Criteria

We applied two criteria based on the likelihood in our experiment: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Clustering evaluation by these criteria is explained by Bacher (2004). Both the criteria indicate the optimal number of clusters by their minimal value.

With the notation mentioned above, we can write the AIC (Akaike, 1973) as

DRAFT

$$AIC(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m}\left(-\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right)\right) + 2k\ \sum_{c=1}^{m}(K_c - 1) \quad (11)$$

and the BIC (Schwarz, 1978) as

$$BIC(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m}\left(-\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right)\right) + k\sum_{c=1}^{m}(K_c - 1)\ln n\ . \quad (12)$$

#### 2.2.3 A Distance-based Criterion

As a representative of distance-based criteria, we used the silhouette index (SI), proposed by Rousseeuw (1987) as the average silhouette width. It takes on values from –1 to 1. Values close to one indicate well-separated clusters with low within-cluster distances and high between-cluster distances. The maximal value of the criterion across all the examined cluster solutions indicates the optimal number of clusters. It is calculated as

$$SI(k) = \frac{1}{n}\sum_{i=1}^{n}\frac{b(i)-a(i)}{\max(a(i),b(i))}, \quad (13)$$

where $a(i)$ is the average distance of the $i$th object from the other objects in the same cluster, and $b(i)$ is the minimal average distance of the $i$th object from other objects in any cluster not containing the $i$th object. For the analysis and for result evaluation, the same type of the distance measure should be used.

## 3. Results of Experiment

The datasets with two, four and six clusters were generated. The other parameters of the datasets were a number of variables (4, 7, and 10) a between-cluster distance (0.21, 0.34, and 0.5). The number of categories was five for all variables. For the experiment, 27 different dataset settings were used. Each dataset setting combination was replicated ten times, it means that 270 datasets were generated. Each dataset was analysed using twelve similarity measures.

Correlation analysis was performed to determine whether the applied evaluation criteria assess the quality of clusters in a similar way. We found that the values of the correlation coefficient express medium to high linear dependences between the examined criteria. There was almost a direct linear dependence ($r$ = 0.998) between the AIC and BIC and between PSFE and PSFM ($r$ = 0.99). Therefore, only one representative for each group of evaluation criteria, namely PSFE, BIC, and SI, are presented for the following analyses.

Table 1 shows the values of the eta coefficient expressing the dependence intensity of a particular evaluation criterion on the similarity measure. The values are obtained based on the ratio of the between-group and total variability in the ANOVA method. The method compares the between-group sum of squares ($SS_B$) with the within-group sum of squares ($SS_W$). The higher the $SS_B$ variability, the more strongly the quantitative variable depends on the categorical one. The *eta coefficient* can be expressed by the formula

$$\eta(v,a) = \sqrt{\frac{SS_B}{SS_B+SS_W}}, \quad (14)$$

where $v$ is a partition with values of an evaluation criterion, $a$ is the categorical variable expressing a given property of the clustered dataset, e.g., the number of clusters or variables.

The outputs are broken down by the number of generated clusters. Whereas the BIC values barely depend on the used similarity measure, the SI values are strongly determined by it. The eta coefficient values decrease with the increasing number of clusters.

DRAFT

Table 1: Eta coefficient values evaluating the dependence intensity of internal criteria on the type of the similarity measure (broken down by the number of clusters)

| Criterion | 2 clusters | 4 clusters | 6 clusters | Total |
|---|---|---|---|---|
| PSFE | 0.529 | 0.427 | 0.320 | 0.345 |
| BIC | 0.214 | 0.193 | 0.187 | 0.190 |
| SI | 0.887 | 0.653 | 0.489 | 0.632 |

Table 2 shows the values of the eta coefficient expressing the dependence intensity of a particular evaluation criterion on the parameters of the dataset, namely the between-cluster distance (clu_dist), the number of variables (n_var), and the number of clusters (n_clu). The outputs are broken down by the similarity measure. We can see that the dependence intensity of internal criteria on the dataset properties varies greatly according to the particular dataset parameter and the type of evaluation criteria.

Table 2: Eta coefficient values evaluating the dependence intensity of internal criteria on the dataset parameters (broken down by the similarity measure)

| | PSFE | | | BIC | | | SI | | |
|---|---|---|---|---|---|---|---|---|---|
| Measure | clu_dist | n_var | n_clu | clu_dist | n_var | n_clu | clu_dist | n_var | n_clu |
| ES | 0.430 | 0.445 | 0.669 | 0.199 | 0.966 | 0.031 | 0.504 | 0.603 | 0.420 |
| G1 | 0.451 | 0.507 | 0.631 | 0.190 | 0.969 | 0.070 | 0.535 | 0.652 | 0.320 |
| G2 | 0.419 | 0.440 | 0.691 | 0.199 | 0.965 | 0.025 | 0.465 | 0.595 | 0.502 |
| G3 | 0.433 | 0.507 | 0.646 | 0.184 | 0.971 | 0.046 | 0.507 | 0.648 | 0.369 |
| G4 | 0.346 | 0.200 | 0.266 | 0.114 | 0.967 | 0.031 | 0.123 | 0.294 | 0.105 |
| IOF | 0.438 | 0.419 | 0.688 | 0.196 | 0.969 | 0.024 | 0.408 | 0.369 | 0.724 |
| LIN | 0.441 | 0.420 | 0.636 | 0.191 | 0.968 | 0.083 | 0.577 | 0.555 | 0.389 |
| LIN1 | 0.363 | 0.702 | 0.176 | 0.068 | 0.976 | 0.163 | 0.243 | 0.809 | 0.090 |
| OF | 0.432 | 0.430 | 0.464 | 0.209 | 0.949 | 0.048 | 0.460 | 0.517 | 0.468 |
| SM | 0.427 | 0.449 | 0.669 | 0.198 | 0.966 | 0.031 | 0.511 | 0.616 | 0.415 |
| VE | 0.418 | 0.467 | 0.683 | 0.185 | 0.970 | 0.018 | 0.470 | 0.639 | 0.414 |
| VM | 0.424 | 0.469 | 0.675 | 0.192 | 0.969 | 0.022 | 0.485 | 0.636 | 0.417 |
| Total | 0.388 | 0.407 | 0.554 | 0.172 | 0.948 | 0.038 | 0.333 | 0.405 | 0.316 |

DRAFT

The total dependence of the BIC criterion on the number of variables (eta = 0.948) is more than twice as strong as by PSFE (eta = 0.407) and SI (eta = 0.405) criteria. On the other hand, the total dependence of this criterion on the between-cluster distance is weak, and the dependency on the number of clusters is minimal. The eta value for this criterion is highest with the number of variables for all similarity measures.

For most similarity measures, the values of PSFE depend mostly on the number of clusters. However, moderate dependences are also with the number of variables and between-cluster distance. For almost all similarity measures, the values of PSFE depend mostly on the number of variables. However, moderate dependences are also with the between-cluster distance and number of variables.

## 4. Conclusion

The results show that the dependence intensity of internal criteria on the type of similarity measure is influenced by the number of clusters. The dependence intensity of internal criteria on the dataset properties varies greatly according to the particular dataset parameter and the type of evaluation criteria. The highest values of the eta coefficient evaluating these dependences were found in the case of the BIC criterion and the number of variables – for all twelve examined similarity measures.

In the case of SI, the values of the eta coefficient are mostly the highest in the case of the number of variables, too. However, the dependence intensity is twice as small in comparison with the BIS criterion. In the case of PSFE, the values of the eta coefficient are mostly the highest in the case of the number of clusters. For this evaluation criterion, comparing the values of the eta coefficient for the particular similarity measures, the values for the LIN1 measure are very different from the others. Due to space limitations, this contribution considers average linkage only. However, it could be extended to study and compare dependence intensity of internal criteria on the type of similarity measure across multiple linkage methods. It is expected by authors, based on the preliminary studies, that the results presented in this contribution would be stable across various linkage methods.

## References


[1] Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike, pp. 199–213. Springer, New York (1973)

[2] Bacher, J., Wenzig, K., Vogler, M.: SPSS TwoStep Cluster – a First Evaluation. Lehrstuhl für Soziologie, Nürnberg (2004)

[3] Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the eighth SIAM international conference on data mining, pp. 243–254 (2008) doi: 10.1137/1.9781611972788.22

[4] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection. In: Barbará, D., Sushil, J. (eds.) Applications of data mining in computer security. Springer, Boston, pp. 77–101 (2002) doi: 10.1007/978-1-4615-0953-0_4

[5] Goodall, D.W.: A new similarity index based on probability. Biometrics **22**(4), 882–907 (1966) https://www.jstor.org/stable/2528080

[6] Hahsler, M., Buchta, C., Gruen, B., Hornik, K.: arules: Mining Association Rules and Frequent Itemsets, R package version 1.7-2 (2021) https://CRAN.R-project.org/package=arules.

[7] Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning, pp. 296–304. Morgan Kaufmann (1998)

[8] Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: Proceedings of 2010 IEEE International Conference on Data Mining, pp. 911–916 (2010)

[9] Milligan, G.W., Cooper, M.C.: Methodology review: Clustering methods. Appl. Psychol. Meas. **11**(4), 329–354 (1987) doi: 10.1177/014662168701100401

[10] R Core Team: R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2020) https://www.R-project.org/.

[11] Qiu, W., Joe, H.: clusterGeneration: Random Cluster Generation (with Specified Degree of Separation), R package version 1.3.7 (2020) https://CRAN.R-project.org/package=clusterGeneration

[12] Řezanková, H., Löster, T., Húsek, D.: Evaluation of categorical data clustering. In: Advances in Intelligent Web Mastering 3, pp. 173–182. Springer Verlag, Berlin (2011)

[13] Rousseeuw, P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20,** 53–65 (1987)

[14] Schwarz, G.: Estimating the dimension of a model, Ann. Stat. **6,** 461–464 (1978)

[15] Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. The University of Kansas Science Bulletin **28**, 1409–1438 (1958)

[16] Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (1972) doi: 10.1108/eb026526

[17] Šulc, Z., Cibulková, J., Řezanková, H.: Nomclust 2.0: an R package for hierarchical clustering of objects characterized by nominal variables. Comput. Stat. **37**(5), 2161–2184 (2022) doi: 10.1007/s00180-022-01209-4

[18] Šulc, Z., Řezanková, H.: Comparison of similarity measures for categorical data in hierarchical clustering. J. Classif. **36**(1), 58–72 (2019) doi: 10.1007/s00357-019-09317-5


DRAFT

# A multivariate hidden semi-Markov model for the analysis of multiple air pollutants

Marco Mingione[a], Pierfrancesco Alaimo Di Loro[b], Francesco Lagona[b], and Antonello Maruotti[a]

[a]Dpt. of Political Sciences, Roma Tre University; marco.mingione@uniroma3.it, francesco.lagona@uniroma3.it
[b]Dpt. GEPLI, Libera Università Maria Ss. Assunta (LUMSA); p.alaimodiloro@lumsa.it, a.maruotti@lumsa.it

DRAFT

**Abstract**

The analysis of multivariate time series of pollutant concentrations is complicated by the complex interactions between pollutants, which may not be constant over over time. We propose to approximate the joint data distribution by a parsimonious finite mixture of vector auto-regressive models, whose parameters are driven by the evolution of a latent semi-Markov process. This results in a vector auto-regressive hidden semi-Markov model that is capable to (1) describe the exposure to pollution in terms of a few latent regimes, (2) associate these regimes with specific combinations of pollutant concentration levels as well as distinct correlation structures between concentrations, and (3) provide estimates of sojourn times in each regime and transition probabilities between regimes. We apply the proposed model to the daily time-series of nine pollutant concentrations recorded at the Marylebone Road station, in London, between 2012 to 2017.

***Keywords:*** air quality, EM algorithm, hidden Markov model, semi-Markov process, penalized regression

## 1. Introduction

Environmental risks are defined as all the external physical, chemical, biological, and work related factors that affect a person's health and well-being. They include pollution, radiation, noise, land use patterns, work environment, and climate change, that are all well-recognised as important causes of disease burden for populations and National Health Systems. Among them, air pollution is classified as one of the greatest environmental risks to health by the WHO[1], causing millions premature deaths worldwide every year (7). Its extent can be measured by the concentrations of various pollutants in the air (10), making its nature intrinsically multi-dimensional.

Here, we propose a multivariate model-based approach, belonging to the class of multivariate vector auto-regressive Hidden Semi-Markov Models (HSMMs; (2)), that can capture the dynamic evolution of environmental risk factors determined by multiple time and cross-dependent components.

HSMMs are recognised as a promising tool driving the assessment of environmental risk building procedure. Specifically, they allow for policy makers to differentiate the overall air-pollution risk exposure depending on the state of the environment identified by the latent process. This synthesis resembles other existing composite metrics such as the air quality index proposed by the US EPA that provides six

[1]see https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

risk levels: “good”, “moderate”, “unhealthy for sensitive groups”, “unhealthy”, “very unhealthy”, “hazardous”. Furthermore, HSMMs are intrinsically dynamic and provide a conditional framework for the quantification of the environmental risk calculated on the resulting predictive distribution. HSMMs can be also viewed as a structured Hidden Markov Models (8; 11) and the predictive distribution is a finite mixture of the semi-Markovian emission distributions. Thus, they allow an analytical assessment of the overall current and future environmental risk conditional on the past values of the process, making them invaluable monitoring tools for regulatory agencies.

Nevertheless, the specification of our model poses two notable statistical challenges: (i) the typical model selection issue related to the order of the autoregression; (ii) the large number of parameters in the state-specific covariance matrices that generate unstable estimates and increasing computational burden. We propose a penalized likelihood approach to HSMMs to deal with these issues, following similar works in the field (3; 4). Here, we contribute to this literature by applying the LASSO regularization on both the hidden state covariance matrix and the vector autoregressive coefficients.

The remainder of the paper is organized as follows. Section 2. summarizes the modeling framework and the estimation process. Section 3. describes the application and the results on the air quality indicators (AQIs) recorded at the Marylebone Road station, in London (UK), from 2012 to 2017. Finally, Section 5. contains some final comments and further developments.

## 2. A model for multivariate times series of pollutant concentrations

The data that motivated our proposal are in the form of a multivariate time series $(\boldsymbol{y}_t, t \geq 0)$, where the vector $\boldsymbol{y}_t = [y_{t1}, \ldots, y_{tp}]$ includes the concentrations of $p$ pollutants at time $t$, $t = 1, \ldots, T$. A popular modelling approach in this setting is provided by the class of vector auto-regressive (VAR) models with lag $H$, where the conditional distribution of the process at time $t$ given the past $\mathcal{H}_t = \{\boldsymbol{y}_\tau,\ \tau = 1, \ldots, t-1\}$:

$$f_{\boldsymbol{Y}_t}(\boldsymbol{y}_t|\mathcal{H}_t) = f_{\boldsymbol{Y}_t}\left(\boldsymbol{y}_t \mid \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{Y}_{t-H} = \boldsymbol{y}_{t-H}\right), \quad \boldsymbol{y}_t \in \mathbb{R}^p,\ t = 1, \ldots, T.$$

is a multivariate normal distribution with mean

$$\boldsymbol{\mu}_t = \boldsymbol{\beta}_0' + \sum_{h=1}^{H} \boldsymbol{\alpha}_h' \boldsymbol{y}_{t-h}, \tag{1}$$

and covariance matrix $\Sigma$. Under a traditional VAR model, the effect of past values on the process is time-constant and measured by the autoregressive coefficients $\boldsymbol{\alpha}$. It is however well known that the dynamic of air quality is affected by unobserved, time-varying weather conditions. This source of latent heterogeneity can be captured by assuming that the conditional distribution of the process given the past is a mixture of $K$ VARs (5). A mixed VAR can be seen as a hierarchical VAR, whose parameters evolve according to the values taken by a sequence of independent multinomial distributions. Such an independence assumption is an obvious limitation of the model, which can be relaxed by allowing latent multinomial distributions to be temporally correlated. Such extension can be easily obtained by specifying the latent multinomial process as a homogeneous Markov chain. A mixture of VARs whose parameters evolve according to a latent Markov chain are known in the literature as multivariate hidden Markov models, and they have been already exploited for the analysis of multivariate environmental time series (9). Under a multivariate hidden Markov model, the sojourn times that the latent process spends in a specific state follow a Geometric distribution. Although this assumption can be realistic in some settings, it can be a serious shortcoming in the analysis of urban pollution, where sojourn times of specific air quality regimes may well show more complicated patterns. To avoid this source of misspecification, we propose a finite mixture of VARs whose parameters vary according to the states of a semi-Markov process. A discrete-time semi-Markov process $(S_t, t \geq 0)$ taking values in a finite state space, say $\{1, 2, \ldots K\}$, can be defined as follows. First, let $(U_t, t \geq 0)$ be a homogeneous Markov chain with a special transition probabilities matrix having all diagonal entries equal to zero (i.e the chain is not allowed to remain in the same state in two subsequent times). A realization of a semi-Markov process is

DRAFT

obtained by replacing the state $k$ of the Markov chain at time $t$ by a run of $d$ copies of state $k$, where $d$ is drawn from a state-specific dwell-time distribution $d_k$. When the dwell time-distributions are all equal to a geometric distribution, then a semi-Markov process reduces to a Markov chain. Otherwise, it extends the Markov chain framework by allowing for flexible dwell times.

Let $(S_t, t \geq 0)$ be a latent semi-Markov process, taking values in the set $\{1 \dots K\}$, whose distribution is known up to a $K \times K$ matrix of transition probabilities with zero diagonal entries and a battery of $K$ dwell time distributions that depend on a finite set of parameters. Conditionally on the value $k$ taken by the latent process at time $t$, we assume that the conditional distribution of the observed process given the past is a multivariate normal distribution with mean

$$\boldsymbol{\mu}_t = \boldsymbol{\beta}'_{k0} + \sum_{h=1}^{H} \boldsymbol{\alpha}'_{kh} \boldsymbol{y}_{t-h}, \tag{2}$$

and covariance matrix $\Sigma_k$. We remark that the resulting mixture VAR is stable, hence stationary, if all the companion matrices:

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{\alpha}_{k1} & \boldsymbol{\alpha}_{k2} & \cdots & \boldsymbol{\alpha}_{kH} \\ \boldsymbol{I}_p & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_p & \mathbf{0} \end{bmatrix}, \quad k = 1, \dots, K,$$

have all the eigen-values with modulus $< 1$ (5).

## 2.1 Estimation

The proposed hidden semi-Markov VAR model can be represented as a multivariate HMM by appropriately augmenting the state sequence, as proposed by Langrock and Zucchini (8). Such representation is exact if all dwell time distributions have finite support. It otherwise provides an approximation with the desired accuracy. An HMM representation of our proposal allows us to exploit the efficient likelihood maximization algorithms that have been developed for HMMs. These algorithms rely on a EM approach, where the maximization of a weighted (data-augmented) log-likelihood function is alternated with weights updating, up to convergence.

The maximization step is further integrated by considering a penalized approach for the estimation of the covariance matrix. This regularizes the estimation process, favoring the model identifiability and highlighting the most evident correlation patterns across the $p$ outcomes. Following (4), we use a convex combination of the maximum likelihood estimator and a scaled identity matrix with the same trace:

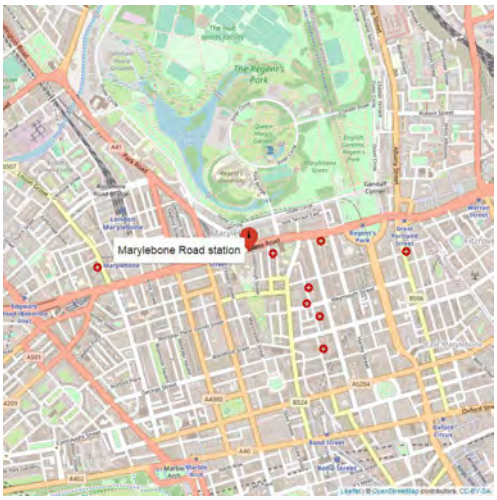
$$\boldsymbol{\Sigma}_k = \frac{1}{1+\lambda_\Sigma} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{\boldsymbol{ML}} + \frac{\lambda_\Sigma}{1+\lambda_\Sigma} c\mathbf{I}, \quad \text{with} \quad \mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{\boldsymbol{ML}}) = tr(c\mathbf{I}).$$

The coefficient $\lambda_\Sigma \geq 0$ is a shrinking parameter that controls for the strength of the penalization. Similarly, we consider a L1-norm penalty on the VAR coefficients and introduce a LASSO shrinking parameter $\lambda_\alpha$ in order to regularize their estimation and simultaneously select a smaller subset of lags that exhibits the strongest effects.

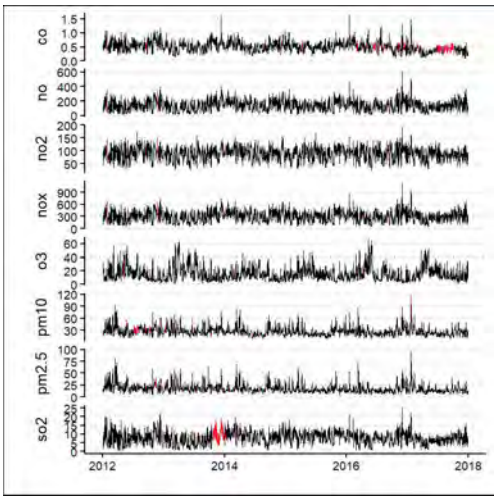
# 3. Application to urban pollution

We apply the proposed model to the daily time-series of $p = 9$ air quality indicators (AQIs) that have been recorded at the Marylebone Road station from 2012 to 2017 in London, for a total of $T = 2,192$ days of observation. This station is placed in the north-western part of the city on a very busy urban road (part of the London ring) and is close to Regent's park (see Figure 1a). Data have been downloaded from the `openair R` package (1), which provides several tools for the analysis of air pollution data, and include the concentrations of the following: carbon monoxide (co), nitrogen dioxide (no2), nitrogen

DRAFT




(a)




(b)

Figure 1: (a) Location of MY1 (Marylebone Road station); (b) daily time-series of the AQIs, including imputed missing values (red line); (c) marginal distribution of the AQIs.

DRAFT

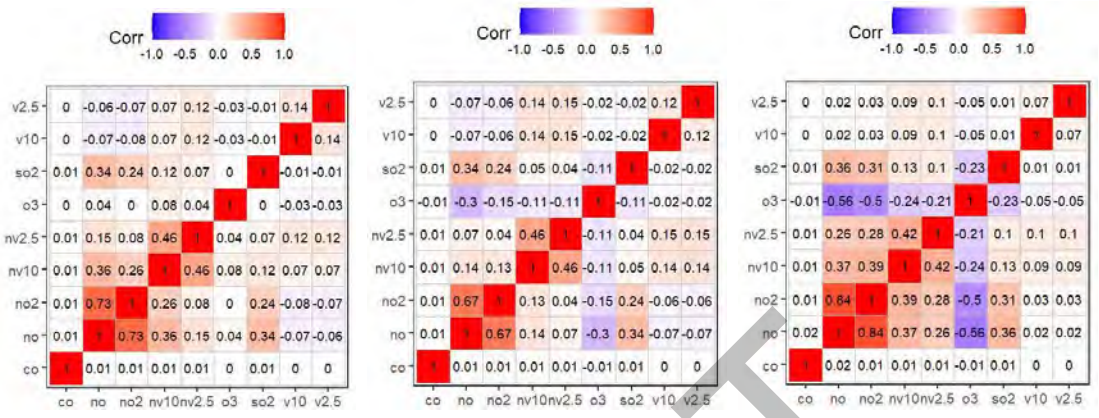
oxide (no), ozone (o3), sulfur dioxide (so2), volatile and non-volatile matter at $2.5\mu m$ (v2.5, nv2.5) and at $10\mu m$ (v10, nv10). As it is common in such applications, the raw data however contained some missing values, perhaps arising from the malfunctioning of the monitoring device. Note that missing values are not necessarily concomitant and the missing data patterns may vary across the AQIs. In our application, we assume that the missing data pattern is Missing at Random (MAR) for each AQI, and there is independence in the missing data mechanism across AQI (i.e. the probability of recording a missing value for one indicator does not depend on concomitant or previous missing values of another indicator). Overall, we counted a total of 804 missing out of $T \times p = 17,536$ values ($< 5\%$), with a percentage of missingness going from a minimum of $\approx 2\%$ for no and no2, up to $\approx 10\%$ for co. Given these low percentages, we used the `Amelia R` package (6) that combines bootstrapping and EM algorithms to provide efficient multiple imputation of incomplete set of data, including time-series data. Here, we use 50 multiple imputations and averaged all the results to get the final estimate of the missing values. A more rigorous approach would rely on integrating the missing value imputation within the proposed EM algorithm, at the price of an additional computational burden. The obtained time-series are shown in Figure 1b, where the imputed values are coloured in red. The series appear stationary through the whole considered time-window and show substantial correlation, particularly when spikes of high concentrations are recorded.

## 4. Results

We run a preliminary analysis on the data described in Section 3. We set the shrinking parameter of the covariance matrix $\lambda_\Sigma = 0.1$ and use cross-validation within the EM algorithm to adaptively select the best LASSO shrinking parameter $\lambda_\alpha$ at each iteration. We set $K = 3$ as to test the hypothesis of observing three possible patterns: pollutants below, in the mid-range, or above average. We consider only the effect of one lag, i.e. $H = 1$. The estimated conditional averages of pollutant concentrations in each state are reported in Table 1. Looking at the point estimates, state 2 can be interpreted as the high-pollution state, showing the largest pollutant average concentrations, except for ozone, which is however uncorrelated with all other pollutants (see Figure 2b). On the other hand, state 3 is the low-pollution state, state 1 is the intermediate state. We estimate 931 days in state 1, 39 days in state 2 and 1222 days in state 3. As expected, the dwell time distribution of state n. 3 is more skewed to the right than the other two distributions, as shown in Figure 3. The LASSO shuts down the effect of carbon oxide on all other pollutants.

| state | co | no | no2 | nv10 | nv2.5 | o3 | so2 | v10 | v2.5 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.79 | 289.83 | 123.56 | 42.43 | 30.36 | 6.34 | 14.57 | 6.63 | 5.93 |
| 1 | 0.6 | 186.92 | 105.78 | 27.77 | 18.49 | 10.52 | 9.94 | 3.89 | 3.16 |
| 3 | 0.42 | 95.44 | 75.38 | 18.94 | 12.04 | 19.08 | 5.97 | 2.91 | 2.19 |

Table 1: Estimated conditional averages in each of the considered latent states.




(a) (b) (c)

Figure 2: Estimated correlation matrices of the pollutants when they are in state 1 (a), 2 (b) and 3 (c).

DRAFT

## 5. Conclusions and further developments

The proposed model is able to describe complex multivariate patterns that present temporal and cross-variable correlation. The non-parametric HSMM specification makes it able to detect and estimate the lower-dimensional structure controlling the different regimes of the multivariate process. The shrinking factors control for the substantial complexity of its specification. Indeed, the full model envisions a faceted dependence structure which depends on a great number of parameters. The adoption of regularized estimation methods becomes a necessity more than an opportunity. An interesting byproduct of the LASSO penalty is its ability to perform selection together with the shrinking. This allows highlighting the most relevant auto-regressive patterns across different components of the multivariate process.

The preliminary analysis on the time-series of pollutants recorded at the Marylebone Road station in London returns promising results. In the near future we aim at implementing a specific selection proce-

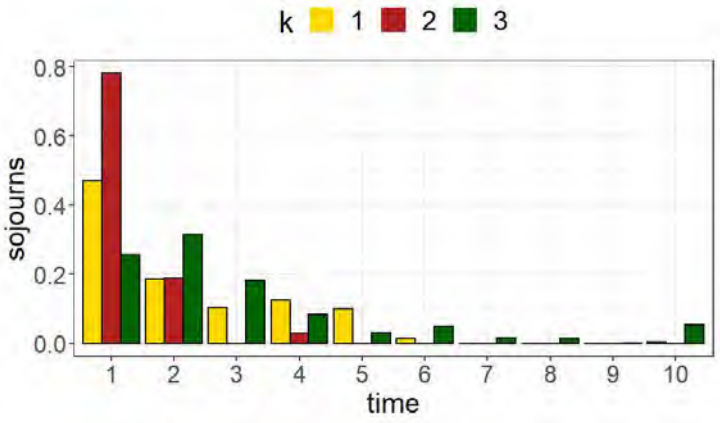



Figure 3: Estimated sojourn probabilities.

dure for the best number of states $K$. Furthermore, we can enrich the results analysis by accompanying them with risk and conditional risk measures, other than associating to each state the probabilities to exceed the risk thresholds fixed by the WHO.

## References


[1] D. C. Carslaw and K. Ropkins. Openair—an r package for air quality data analysis. *Environmental Modelling & Software*, 27:52–61, 2012.

[2] M. Ehrmann, M. Ellison, and N. Valla. Regime-dependent impulse response functions in a markov-switching vector autoregression model. *Economics Letters*, 78(3):295–299, 2003.

[3] A. Farcomeni. Penalized estimation in latent markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency. *Biometrical Journal*, 59(5):1035–1046, 2017.

[4] M. Fiecas, J. Franke, R. von Sachs, and J. Tadjuidje Kamgaing. Shrinkage estimation for multivariate hidden markov models. *Journal of the American Statistical Association*, 112(517):424–435, 2017.

[5] P. W. Fong, W. K. Li, C. Yau, and C. S. Wong. On a mixture vector autoregressive model. *Canadian Journal of Statistics*, 35(1):135–150, 2007.

[6] J. Honaker, G. King, M. Blackwell, and M. M. Blackwell. Package 'amelia'. *Version. View Article*, 2010.

[7] P. J. Landrigan. Air pollution and health. *The Lancet Public Health*, 2(1):e4–e5, 2017.

[8] R. Langrock and W. Zucchini. Hidden markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1):715–724, 2011.

[9] A. Maruotti, J. Bulla, F. Lagona, M. Picone, and F. Martella. Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *The Annals of Applied Statistics*, 11(3):1617 – 1648, 2017.

[10] W. H. Organization et al. Who global air quality guidelines: particulate matter (pm2. 5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary. 2021.

[11] J. Pohle, T. Adam, and L. T. Beumer. Flexible estimation of the state dwell-time distribution in hidden semi-markov models. *Computational Statistics & Data Analysis*, 172:107479, 2022.


DRAFT

# A smooth transition autoregressive model for matrix-variate time series

Andrea Bucci[a]

[a]Department of Economics and Law, University of Macerata; andrea.bucci@unimc.it

**Abstract**

In many applications, data are observed as matrices with temporal dependence. Matrix-variate time series modeling is a new branch of econometrics. Although stylized facts in several fields, the existing models do not account for regime switches in the dynamics of matrices that are not abrupt. In this paper, we extend linear matrix-variate autoregressive models by introducing a regime-switching model capable of accounting for smooth changes, the matrix smooth transition autoregressive (MSTAR) model. We provide a thorough examination of the estimation process and evaluate the finite-sample performance of the MSTAR model estimators with simulated data.

***Keywords:*** Matrix-valued time series; Smooth transition; Multivariate time series

DRAFT

## 1. Introduction

Time series processes are present in many fields, including econometrics, finance, biology, and ecology. The typical classification of time series analysis is between univariate and multivariate data, and both have been extensively studied in the related literature (10). There exists a third possible specification in the temporal domain which involves multidimensional datasets (2). Of particular interest are bi-dimensional data where the variables are organized in matrices that vary over time. While matrix-valued data has been studied in several papers (12), only recently the matrix-variate autoregressive (MAR) model proposed by Wang et al. (11) and Chen et al. (4) has opened up new territories for modeling the temporal dependency in matrix-variate problems with a twofold advantage with respect to the more used stacked vectorization form: the original matrix structure is preserved meaning that the coefficient matrices can be interpreted row- and column-wisely (4); the number of parameters is drastically decreased.

Most of the research in this field has focused on linear autoregressive matrix-variate models. However, economic and financial systems have often been shown to exhibit both structural and behavioral changes. Accordingly, Liu and Chen (7) have recently introduced a threshold version of the MAR model that accounts for abrupt changes in the matrices of parameters, while Billio et al. (3) propose a smooth transition model for temporal networks estimated through a Bayesian approach. In this paper, we suppose that the regime changes can be smooth and we introduce the matrix-variate smooth transition (MSTAR) model inspired by the vector smooth transition autoregressive (VSTAR) model (9). Although a typical assumption is that the transition variable is weakly stationary, we also analyse the case in which a regime changes coincides with a structural break, meaning that the temporal trend is used as a transition variable. This may help identifying common breaks in matrix-valued problems. We investigate the estimation process that allows to both estimate the coefficient matrices and the parameters of the transition function. The finite sample performance of the estimators is assessed through a simulation study.

The remainder of the paper is organized as follows. Section 2 introduces the MSTAR model. A numerical study is conducted in Section 3 to assess the finite-sample properties of the estimators, while Section 4 concludes.

## 2. Matrix smooth transition autoregressive model

Consider an $m \times n$ matrix $\mathbf{Y}_t$ observed at time $t$, for $t = 1, \ldots, T$. Let vec$(\cdot)$ be the vectorization of a matrix by stacking its columns. One way to introduce some nonlinearity in the dynamics of $\mathbf{Y}_t$ is generalising to the matrix-variate framework the threshold autoregressive model proposed by Quandt (8), as done in Liu and Chen (7). Another possible nonlinear extension may foresee the use of a smooth transition mechanism, as done for multivariate time series in Anderson and Vahid (1), and Tërasvirta and Yang (9). Supposing the absence of exogenous variables and a single lag, the zero-mean 2-regime vector smooth transition autoregressive (VSTAR) model can be defined as follows

$$\text{vec}(\mathbf{Y}_t) = \mathbf{\Phi}_0 \text{vec}(\mathbf{Y}_{t-1}) + \mathbf{G}(\gamma, \mathbf{c}; \mathbf{s}_t) \left(\mathbf{\Phi}_1 \text{vec}(\mathbf{Y}_{t-1})\right) + \text{vec}(\mathbf{E}_t) \tag{1}$$

where $\mathbf{\Phi}_k$ are $mn \times mn$ parameter matrices, for $k = 0, 1$, $\mathbf{E}_t$ is an $m \times n$ matrix of errors, and $\mathbf{G}(\gamma, \mathbf{c}; \mathbf{s}_t)$ is an $mn \times mn$ diagonal matrix, such that

$$\mathbf{G}(\gamma, \mathbf{c}; \mathbf{s}_t) = \text{diag}\left\{g_1(\gamma_1, c_1; s_{1,t}), \ldots, g_{mn}(\gamma_{mn}, c_{mn}; s_{mn,t})\right\}, \tag{2}$$

where $g(\cdot)$ is typically a standard logistic function, $g_i(\gamma_i, c_i; s_{i,t}) = [1 + \exp\{-\gamma_i (s_{i,t} - c_i)\}]^{-1}$, and $s_{i,t}$, for $i = 1, \ldots, mn$, is a weakly stationary transition variable usually chosen among the set of variables in vec$(\mathbf{Y}_{t-1})$, although exogenous variables are possible as well (5). Special cases of the model in (1) foresee a common transition variable for all the $mn$ equations or, more strictly, a unique transition function that drives the regime changes of the autoregressive parameters. Extending to the matrix-variate framework the smooth transition version the MAR model may imply introducing a regime-switching either for the parameters related to the column-wise effect, those related to the row-wise effect or both. However, the interpretation of the parameters in such models becomes extremely difficult, since the transition variables and the parameters of the function can differ for each element of the matrices in the second regime. A simpler version of the aforementioned models foresees that the transition variables, as well as the parameters of the transition functions, are the same for all the equations (6). For its simplicity, its popularity in most of the macroeconomic problems and its similarity with a threshold autoregressive model as the one introduced in Liu and Chen (7), we use this specification to extend in a matrix-variate context the VSTAR model. Supposing a two-regime model (*i.e.*, a single transition function), we extend the MAR model by introducing an additive nonlinear component, such that

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' + g_t(\gamma, c; s_t)(\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}') + \mathbf{E}_t, \tag{3}$$

where $g_t(\gamma, c, ; s_t)$ is, as before, a standard logistic function (for ease of notation we will refer to this function as $g_t$). Obviously, more than two regimes are also allowed even if in practice a number of regimes greater than 3 is of scarce interest. This matrix smooth transition autoregressive (MSTAR) model implies that the conditional mean of the entire matrix changes the regime based on the values of the unique transition variable, $s_t$. Nevertheless, finding a common transition variable that drives the dynamics of the entire matrix is not always easy, especially when the entries of the matrix are different indices. For this reason, a more interesting candidate transition variable is the normalized temporal trend, $s_t = t/T$. In this case, model (1) becomes a linear vector regression with a smooth structural break (5). It follows that, if we use $s_t = t/T$ as a transition variable in the MSTAR model proposed in this article, the estimation of the threshold parameter $c$ implies a single structural break detection for all the matrix-valued time series.

### 2.1 Estimation of the MSTAR model

In a general MAR model, assuming that the entries of $\mathbf{E}_t$ are i.i.d. normal with mean zero and a constant variance, the estimates of the parameters are the solution of the least squares problem

$$\min_{\mathbf{A},\mathbf{B}} \sum_t \| \mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' \|_F^2 . \tag{4}$$

DRAFT

Chen et al. (4) show that the estimates of the matrices of parameters can be obtained either through a projection method, maximum likelihood and iterative least squares. In the estimation process of the nonlinear extension of the MAR, we focus on the latter method. The estimation of the MSTAR model is more complicated than a simple MAR model for the presence of the threshold and the slope parameters. Moreover, the splitting algorithm used for the iterative least squares estimates used for a matrix-variate threshold autoregressive model (7) is not applicable. One way to solve these shortcomings is using the same algorithm proposed in Hubrich and Tërasvirta (6) which foresees the minimization of the objective function conditionally on $\gamma$ and $c$. Let the optimization problem be

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \sum_t \parallel \mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}' \parallel_F^2, \tag{5}$$

which leads to the following loss function to be minimized:

$$Q = \text{tr}\left(\sum_t \left(\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}'\right)'\left(\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}'\right)\right). \tag{6}$$

It is clear that if all the coefficients are zeros, than the parameters $\gamma$ and $c$ are non-identifiable. A typical assumption to solve this issue is that not all the coefficients are zeros. Therefore, conditionally on the values of the transition function, the minimization of the loss function in Eq. (6) is guaranteed to have at least one global minimum and can be simply solved by taking its first derivatives where $g_t$ is treated as a known scalar. As in the case of the MAR, the solution of (5) is found by iteratively updating one at a time the matrices $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$, and $\hat{\mathbf{D}}$, while keeping the others fixed, starting with some initial matrices.

Once obtained the following iterative least squares estimates of the autoregressive coefficient matrices

$$\begin{aligned}
\mathbf{A} &\leftarrow \left(\sum_t \mathbf{Y}_t\mathbf{B}\mathbf{Y}_{t-1}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}'\mathbf{B}\mathbf{Y}_{t-1}'\right)\left(\sum_t \mathbf{Y}_{t-1}\mathbf{B}'\mathbf{B}\mathbf{Y}_{t-1}'\right)^{-1} \\
\mathbf{B} &\leftarrow \left(\sum_t \mathbf{Y}_t'\mathbf{A}\mathbf{Y}_{t-1} - g_t\mathbf{D}\mathbf{Y}_{t-1}'\mathbf{C}'\mathbf{A}\mathbf{Y}_{t-1}\right)\left(\sum_t \mathbf{Y}_{t-1}'\mathbf{A}'\mathbf{A}\mathbf{Y}_{t-1}\right)^{-1} \\
\mathbf{C} &\leftarrow \left(\sum_t g_t\mathbf{Y}_t\mathbf{D}\mathbf{Y}_{t-1}' - g_t\mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}'\mathbf{D}\mathbf{Y}_{t-1}'\right)\left(\sum_t g_t^2\mathbf{Y}_{t-1}\mathbf{D}'\mathbf{D}\mathbf{Y}_{t-1}'\right)^{-1} \\
\mathbf{D} &\leftarrow \left(\sum_t g_t\mathbf{Y}_t'\mathbf{C}\mathbf{Y}_{t-1} - g_t\mathbf{B}\mathbf{Y}_{t-1}'\mathbf{A}'\mathbf{C}\mathbf{Y}_{t-1}\right)\left(\sum_t g_t^2\mathbf{Y}_{t-1}'\mathbf{C}'\mathbf{C}\mathbf{Y}_{t-1}\right)^{-1},
\end{aligned}$$

the optimization problem foresees the minimization of the loss in Eq. (5) with respect to $\gamma$ and $c$. Finding the optimum of such a function is performed analytically. In practice, the numerical algorithm may be slow and may converge to some local minimum. For this reason, we apply the same method proposed by (6) which implies the use of a grid search among a set of possible values for $\gamma$ and $c$. We construct a discrete grid in the parameter space of the transition function parameters and we estimate as before the autoregressive coefficients, conditionally on the pair of values on the grid. The estimation of the parameters in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ is carried out for all the values in the grid and the residuals sum of squares

$$\text{SSQ} = \text{tr}\left(\sum_t \text{vec}(\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}')'\text{vec}(\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' - g_t\mathbf{C}\mathbf{Y}_{t-1}\mathbf{D}')\right) \tag{7}$$

is collected. We then choose as starting values for the iterative least squares presented above the pair of values $(\gamma, c)$ that produces the smallest SSQ. Once obtained the estimates of the autoregressive coefficient parameters, the algorithm proceeds with the estimation of $\gamma$ and $c$ fixing all the parameters in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$. This continues until convergence.

DRAFT

## 3. Simulation study

To study the finite-sample properties of the model introduced in this article, we perform a simulation study in which we generate $R = 100$ replications using different settings in terms of dimensionality of the matrix (*i.e.*, with different $m$ and $n$), and the length of the time series, with $T = 400, 600, 1000$.

$\mathbf{Y}_t$ is generated from model (3), where $\mathbf{A}$ and $\mathbf{B}$ are diagonal matrices with 0.20 entries on their diagonal, while $\mathbf{C}$ and $\mathbf{D}$ are computed as a diagonal matrix with entries 0.75. For all the replications we use the same coefficient matrices and we let $\mathbf{E}_t$ vary. This guarantees that $\rho(\mathbf{A}) \cdot \rho(\mathbf{B}) < 1$ and $\rho(\mathbf{C}) \cdot \rho(\mathbf{D}) < 1$, then all the DGPs are stationary. Since the iterative algorithm for the estimation of the matrices foresees a set of initial values for the matrices $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$, we sample the value of $\mathbf{B}$ from a uniform distribution $\mathcal{U}(0, 0.2)$, while the entries of the initial $\mathbf{C}$ and $\mathbf{D}$ are sampled from a uniform distribution $\mathcal{U}(0, 0.5)$. In addition, we let $\gamma$ range between 1 and 50 and $c$ between 0 and 1 to compute the searching grid algorithm. The innovation matrix, $\mathbf{E}_t$, is sampled from a matrix normal distribution, with the $mn \times mn$ covariance matrix equal to $\mathbf{\Sigma} = \mathbf{I}_m \otimes \mathbf{I}_n$. We choose to use as a transition variable the scaled temporal trend, *i.e.*, $s_t = t/T$, we set $c = 0.65$ and $\gamma = 10$.

We also estimate the stacked version of the model, this means estimating a VLSTAR model on $\text{vec}(\mathbf{Y}_t)$, and for both the models we compute the box-plot among the 100 replications of the following loss functions

$$\| \hat{\mathbf{B}} \otimes \hat{\mathbf{A}} - \mathbf{B} \otimes \mathbf{A} \|_F^2 \text{ and } \| \hat{\mathbf{D}} \otimes \hat{\mathbf{C}} - \mathbf{D} \otimes \mathbf{C} \|_F^2 . \tag{8}$$

Figure 1 shows the simulation results in terms of the measures in (8) for an increasing number of observations (from left to right) and an increasing dimension (from top to bottom) taking values in $(m, n) = (2, 3)$, and $(4, 6)$. For each simulation setting, we report in Figure 2 and 3 the box-plot of the difference of the estimated and observed threshold and slope parameters in the 100 replications.

From Figure 1 it emerges that, for each simulation setting, the MSTAR model strongly outperforms the VLSTAR model in the estimation of the coefficient matrices for both the regimes when the dimension of the matrix is $(4, 6)$. Moreover, it always overperforms the VLSTAR model in the estimation of the coefficient matrix in the first regime with mixed evidence in terms of the coefficient matrix in the second regime. In Figure 2, it can be observed that for $(m, n) = (2, 3)$, the median difference between the estimated and the observed threshold is extremely close to zero regardless of the sample size (last row in the Figure), meaning that the threshold is correctly estimated. Overall, the estimated thresholds are in line with the observed value for both the methods, and the difference between the two methods is often negligible. Still, for larger matrices and larger samples, the MSTAR model seems to outperform the VLSTAR one also in terms of threshold estimation. Figure 3 reports the box-plots among the $R$ replicates of the slope parameter $\gamma$. It can be observed that the two models perform similarly well in the estimation of the slope parameter when the dimension of the matrix is $(2, 3)$. When the matrix dimension and the sample size increase, the difference between the estimated and the observed slope parameter is strictly close to zero for the MSTAR model, while it diverges for the VLSTAR.

## 4. Conclusions

In this article, we have proposed a smooth transition autoregressive model for matrix-variate time series which extends the VLSTAR model. In a numerical study, we show that the MSTAR model is able to outperform the stacked vectorization form through a VLSTAR in terms of estimated coefficients, threshold and slope parameters. This is mostly true for larger samples and larger matrices.

There are several open questions that could be solved in future works. For instance, a linearity test of the matrix-variate time series could avoid the use of a nonlinear specification when a linear one is needed. Moreover, a proper procedure for the detection of the regimes in a smooth transition model would be helpful.

DRAFT

Figure 1: Comparison of the estimated coefficient matrices for the MSTAR and VLSTAR models. The rows identify an increasing matrix dimension of $(m, n) = (2, 3), (4, 6)$, while the number of observations in the sample grows by column. For each subfigure, the plot on the left denotes the first regime, while the plot on the right reports the measure for the second regime.

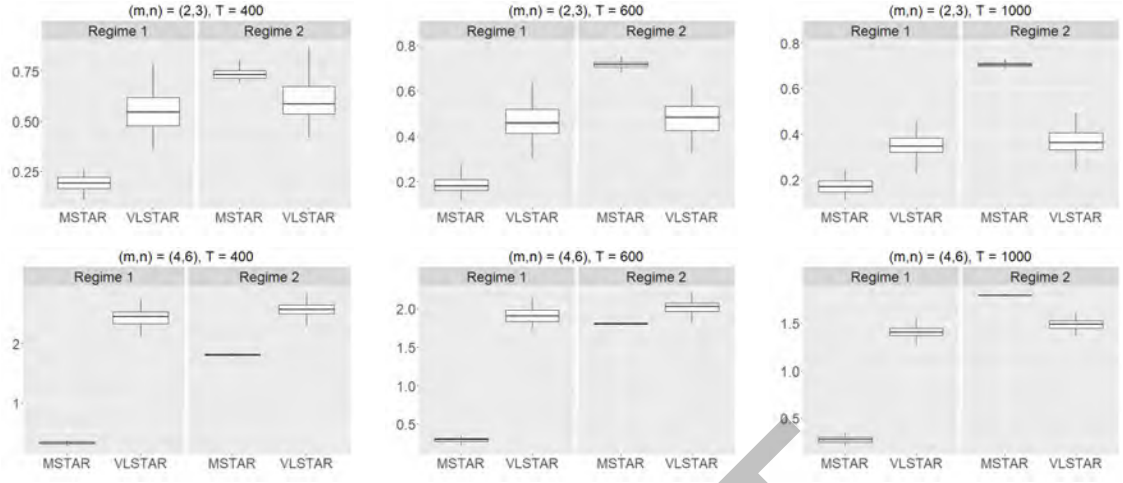
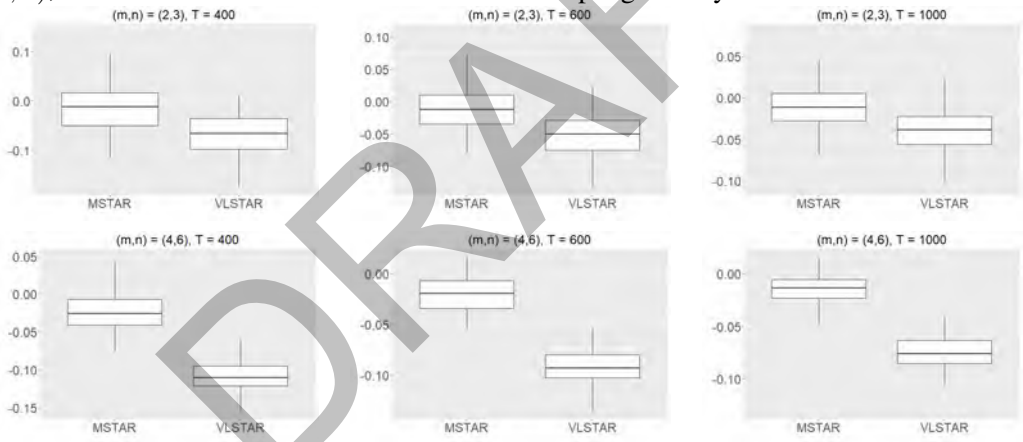
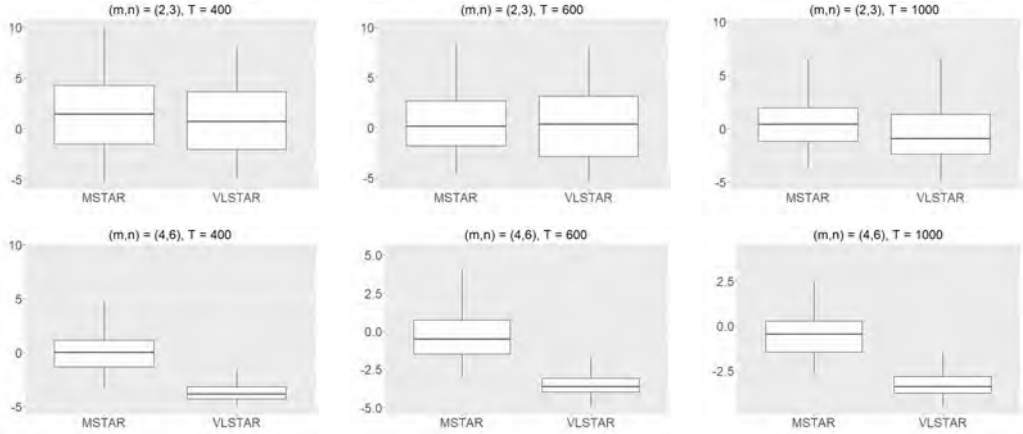



Figure 2: Comparison of the differences between the estimated and real ($c = 0.65$) threshold parameters for the MSTAR and VLSTAR models. The rows identify an increasing matrix dimension of $(m, n) = (2, 3), (4, 6)$, while the number of observations in the sample grows by column.




DRAFT

Figure 3: Comparison of the differences between the estimated and real ($\gamma = 10$) slope parameters for the MSTAR and VLSTAR models. The rows identify an increasing matrix of $(m, n) = (2, 3), (4, 6)$, while the number of observations in the sample grows by column.



## References


[1] Anderson, H., Vahid, F.: Testing multiple equation systems for common nonlinear components. Journal of Econometrics. **84**, 1–36 (1998)

[2] Billio, M., Casarin, R., Costola, M., Iacopini, M.: Matrix-variate Smooth Transition Models for Temporal Networks. In: Bekker, A., Ferreira, J.T., Arashi, M., Chen, D.G., Innovations in Multivariate Statistical Modeling, pp. 137-167. Springer (2022)

[3] Billio, M., Casarin, R., Iacopini, M., Kaufmann, S.: Bayesian Dynamic Tensor Regression. Journal of Business and Economic Statistics. **41**, 1–11 (2022)

[4] Chen, R., Xiao, H., Yang, D.: Autoregressive models for matrix-valued time series. Journal of Econometrics. **222**, 539–560 (2021)

[5] He, C., Tërasvirta, T., González, A.: Testing Parameter Constancy in Stationary Vector Autoregressive Models Against Continuous Change. Econometric Reviews. **28**, 225–245 (2008)

[6] Hubrich, K., Tërasvirta, T.: Thresholds and Smooth Transitions in Vector Autoregressive Models. In: Avdances in Econometrics, pp. 273-326. Emerald Group Publishing Limited (2013)

[7] Liu, X., Chen., Y.: Identification and estimation of threshold matrix-variate factor models. Scandinavian Journal of Statistics. **49**, 1383–1417 (2022)

[8] Quandt., R.E.: The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. Journal of the American Statistical Association. **53**, 873–880 (1958)

[9] Tërasvirta, T., Yang, Y.: Specification, estimation and evaluation of vector smooth transition autoregressive models with applications. CREATES Research Paper, Aarhus University (2014)

[10] Tsay, R.S.: Multivariate Time Series Analysis: With R and Financial Applications. Wiley Series in Probability and Statistics (2014)

[11] Wang, D., Liu, X., Chen, R.: Factor models for matrix-valued high-dimensional time series. Journal of Econometrics. **208**, 231–248 (2019)

[12] Zhou, H., Li, L., Zhu, H.: Tensor Regression with Applications in Neuroimaging Data Analysis. Journal of the American Statistical Association. **108**, 540–552 (2013)


DRAFT

# Dynamic network models with time-varying nodes

Luca Gherardini[a], Mauro Bernardi[b], and Monia Lupparelli[a]

[a]Department of Statistics, Computer Science, Applications "G. Parenti" University of Florence, viale Morgagni 59, 50133 Firenze, Italy;
luca.gherardini@unifi.it, monia.lupparelli@unifi.it
[b]Department of Statistical Sciences, University of Padova, via Cesare Battisti 241, 35121 Padova, Italy;
mauro.bernardi@unipd.it

**Abstract**

We develop a class of random effect models for dynamic networks aiming to account for a time-varying network topology. Ignoring this aspect may lead to a bias in the parameters estimate of the model for the edges, since, in principle, an edge is not observable when the related couple of nodes does not belong to the network. A Bayesian conjugate approach is proposed for the inference of this class of models based on the Pólya-Gamma latent variable method. A preliminary simulation study has been implemented to investigate the empirical performances of the proposed approach.

***Keywords:*** Bayesian inference, Mixture model, Zero-inflation.

DRAFT

## 1. Introduction

Networks are used to represent complex data structures, that arise in different fields of science, including economics, biology, and sociology, among others. These data encode the relationships between different units or actors, stimulating interesting research questions and new methodological approaches. For a general introduction to network science, see Newman (2018). Networks are often not static objects, since they can evolve over time and this dynamic behaviour adds a further level of complexity. Most existing methods only consider a single snapshot of the network at a given time, making it difficult to study the network evolution. In this context, developing advanced statistical models for dynamic network analysis becomes relevant. Main approaches available in the literature belong to the broad class of latent variable models, which are latent space models and stochastic blockmodels. For a general overview of both frameworks, see Kim et al. (2018). Another popular class of models used for temporal networks are continuous-time Markov processes and temporal exponential random graph, we refer to Goldenberg et al. (2009) and references therein for an introduction to the topic. However, most of the approaches mentioned above model the dynamic behaviour of edges without considering that the network's topology may vary over time. We expect that ignoring this issue can lead to distortion for the parameter estimates of the network model since the model is not able to distinguish between observed missing edges and lacks of edges for pairs of nodes which do not belong to the network topology at a given time.
To address this relevant issue, we develop a fully dynamic modelling framework that takes into account both the node and edge temporal behaviour. Inspired by the dynamic version of latent factor model for undirected binary networks (Durante and Dunson, 2014), we propose a class of zero-inflated Bernoulli models for the network edges, which discriminates between structural zeros for missing edges and those produced by observing a lack of edges between pairs of observed nodes. The inference approach for this class of models is developed within the Bayesian paradigm and relies on a Gibbs sampling algorithm

with Pólya-Gamma the data augmentation scheme, introduced by Polson et al. (2013). The performance of our approach is explored through a preliminary simulation study. The remainder of this contribution is organized as follows. Section 2. describes the model formulation in detail. In Section 3, we develop a Gibbs sampling algorithm for inference. Section 4. provides some preliminary results of our simulation study. Finally, Section 5. contains concluding remarks and further extensions for the proposed approach.

## 2. Dynamic network model

A time-varying graph $\{\mathcal{G}_t\}_{t=1}^T = \{(\mathcal{N}_t, \mathcal{E}_t)\}_{t=1}^T$, is represented by a set of *nodes* or *vertices* $\mathcal{N}_t = \{1, \ldots, V_t\}$ and an *edge*'s set $\mathcal{E}_t \subseteq \{(i,j) : i,j \in \mathcal{N}_t, i > j\}$, for $t = 1, \ldots, T$. The entire graph structure is encoded by a sequence of adjacency matrices $\boldsymbol{\mathcal{Y}} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$ that provide a full representation of the *dynamic network*. Each $\mathbf{Y}_t = \{Y_{ijt}\}$ denotes an $V_t \times V_t$ symmetric matrix, characterizing the *undirected* binary network connectivity with no self-loops. The realization of $Y_{ijt}$ is denoted by $Y_{ijt} = y_{jit}$, which is equal to 1 if nodes $i$ and $j$ share a connection at time $t$ and is equal to 0 otherwise. In order to take into account the vertex dynamics, we define the maximum network size $V_{\max} = |\mathcal{N}_{\max}|$, where $\mathcal{N}_{\max}$ is a set of all nodes until $T$, such that $\mathcal{N}_t \subseteq \mathcal{N}_{\max}$, for $t = 1, \ldots, T$. Therefore, our network at time $t$ is represented by an $V_{\max} \times V_{\max}$ adjacency matrix $\mathbf{Y}_t$, where the set of links $\mathcal{E}_t \subseteq \mathcal{N}_{\max} \times \mathcal{N}_{\max}$ and entries are defined as before.

### 2.1 Model formulation

Let $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iT}) \in \mathbb{R}^T$ denotes the $T$- dimensional vector, such that $Z_{it} = 1$ if $i \in \mathcal{N}_t$, and 0 otherwise. Assuming independence between $Z_{it}$ and $Z_{jt}$ with $i \neq j$, the joint presence/absence of pair of nodes in the network at time $t$, can be described by $\boldsymbol{Z}_{ij}^\star = \boldsymbol{Z}_i \circ \boldsymbol{Z}_j, \in \mathbb{R}^T$, where $\circ$ denotes the element-wise vector product operator. This new variable follows a Bernoulli distribution with probability equal to $\lambda_{ijt}$, i.e. $Z_{ijt}^\star \sim \mathcal{B}(\lambda_{ijt})$. The dynamic network's topology is fully characterized by the sequence $\boldsymbol{\mathcal{Z}}^\star = \{\boldsymbol{Z}_1^\star, \boldsymbol{Z}_2^\star, \ldots, \boldsymbol{Z}_{V_{\max}(V_{\max}-1)/2}^\star\}$, which means that $Y_{ijt} = 0$ if $Z_{ijt}^\star = 0$ at time $t$, otherwise $Y_{ijt}$ follows a Bernoulli distribution, i.e., $Y_{ijt} \sim \mathcal{B}(\pi_{ijt})$. As a consequence, we can define the following zero-inflated distribution:

$$Y_{ijt} | Z_{ijt}^\star, \lambda_{ijt}, \pi_{ijt} \sim \begin{cases} 0\,, & \text{with probability} \quad 1 - \lambda_{ijt} \\ \mathcal{B}(\pi_{ijt})\,, & \text{with probability} \quad \lambda_{ijt}. \end{cases} \tag{1}$$

where $\pi_{ijt}$ the denotes dynamic edge probability, for $t = 1, \ldots, T$ and $i > j, \in \mathcal{N}_{\max}$. This model can be reformulated as a two-component mixture model where the first part addresses the structural zero, due to the absence of a pair of nodes at time $t$ and the second part is related to the edge probability through a Bernoulli distribution, "adjusted"for the presence of that pair, at given time $t$. Therefore, we can write (1) as a two-component finite mixture model:

$$Y_{ijt} \sim (1 - \lambda_{ijt})\, \mathbb{1}_{\left(z_{ijt}^\star = 0\,,\, y_{ijt} = 0\right)} + \lambda_{ijt}\, \mathcal{B}\left(\pi_{ijt}\right) \mathbb{1}_{(z_{ijt}^\star = 1)}, \quad \begin{array}{l} t = 1, \ldots, T, \\ i = 2, \ldots, V_{\max}, j = 1, \ldots, i-1 \end{array} \tag{2}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function and $z_{ijt}^\star$ is an observable binary variable such that with probability $(1 - \lambda_{ijt})$, $z_{ijt}^\star = y_{ijt} = 0$ implying the absence of a pair of nodes and edge connecting them.

DRAFT

Moreover, with probability $\lambda_{ijt}$, we model the joint presence of node's pair, i.e $z^{\star}_{ijt} = 1$ and $y_{ijt}$ is in turn drawn from a Bernoulli distribution with probability $\pi_{ijt}$. This representation is very convenient as $Z^{\star}_{ijt}$ is an observable process and in this manner, we are able to discriminate structural zeros due to the topology of the network from those related to the absence of an edge between a pair of nodes. Furthermore, we assume that both edges and nodes probabilities are modelled via logistic regression models:

(*i*) $\text{logit}(\pi_{ijt}) = \mathbf{x}_{ijt}^{\intercal}\boldsymbol{\beta}$, where $\mathbf{x}_{ijt} = (x_{ijt1}, x_{ijt2}, \ldots, x_{ijtp})^{\intercal}$ is a set of $p$ dyad edges covariates and $\boldsymbol{\beta} = (\beta, \ldots, \beta_p)^{\intercal}$ is the $p \times 1$ vector.

(*ii*) $\text{logit}(\lambda_{ijt}) = \mathbf{w}_{ijt}^{\intercal}\boldsymbol{\gamma}$, where $\mathbf{w}_{ijt} = (w_{ijt1}, w_{ijt2}, \ldots, w_{ijtq})^{\intercal}$ is a set of $q$ dyad nodes covariates and $\boldsymbol{\gamma} = (\gamma, \ldots, \gamma_q)^{\intercal}$ is the $q \times 1$ vector.

Following the model assumptions outlined above, the complete-data likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}}^{\star}) = \prod_{t=1}^{T} \prod_{i>j \in \mathcal{N}_{\max}} \left(\lambda_{ijt}^{z^{\star}_{ijt}} (1-\lambda_{ijt})^{1-z^{\star}_{ijt}}\right) \left(\pi_{ijt}^{y_{ijt}} (1-\pi_{ijt})^{1-y_{ijt}}\right)^{z^{\star}_{ijt}} \tag{3}$$

This model can be generalized with the inclusion of further components, including for instance a spatial/temporal latent factor or a latent space component, on both edges and nodes processes. In the next Section, we propose a first attempt of the dynamic model with random effects.

## 3. Inference

Let us define two independent random effects denoted by $\phi_{1,ij}|\xi_1^{(ij)} \sim \mathcal{N}(0, \xi_1^{(ij)})$ and $\phi_{2,ij}|\xi_2^{(ij)} \sim \mathcal{N}(0, \xi_2^{(ij)})$, with $\xi_1^{(ij)}, \xi_2^{(ij)} \in \mathbb{R}^{+}$. These components capture the heterogeneity of the nodes interaction and edges connectivity patterns. As before, we assume:

$$\begin{aligned} \text{logit}(\pi_{ijt}) &= \mathbf{x}_{ijt}^{\intercal}\boldsymbol{\beta} + \phi_{1,ij} \\ \text{logit}(\lambda_{ijt}) &= \mathbf{w}_{ijt}^{\intercal}\boldsymbol{\gamma} + \phi_{2,ij}\,, \end{aligned} \tag{4}$$

where the lagged observation of $z^{\star}_{ij\,t-1}$ is included in the covariate set $\mathbf{w}_{ijt}$ to account for the dynamic trend of the nodes, for $t = 1, \ldots, T$ and $i = 2, \ldots, V_{\max}, j = 1, \ldots, i-1$. For estimation purposes, we follow the Bayesian approach. The Bayesian inferential procedure requires the specification of the prior distribution for the unknown vector of parameters $\boldsymbol{\vartheta} = (\boldsymbol{\gamma}^{\intercal}, \boldsymbol{\beta}^{\intercal}, \xi_1^{(ij)}, \xi_2^{(ij)})^{\intercal}$. Posterior sampling is performed with a Gibbs sampler via Pólya-Gamma data augmentation scheme, that allows a conjugated inference in this specific framework. Therefore, we define $\boldsymbol{\Omega}_t = \text{blockdiag}(\boldsymbol{\Omega}_t^{(z^{\star})}, \boldsymbol{\Omega}_t^{(y)})$, where $\boldsymbol{\Omega}_t^{(\cdot)} = \text{diag}(\boldsymbol{\omega}_t^{(\cdot)}) = (\omega_{12,t}, \omega_{13,t}, \ldots, \omega_{(V_{\max}-1)V_{\max}/2,t})$ and each $\omega_{ijt}^{(\cdot)} \sim \text{PG}(1, 0)$. The sequence of latent variables is denoted by $\boldsymbol{\Omega}^{\star} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_T)$.

| Edges process | Nodes process |
|---|---|
| $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\beta,0}, \boldsymbol{\Sigma}_{\beta,0})$ | $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_{\gamma,0}, \boldsymbol{\Sigma}_{\gamma,0})$ |
| $\xi_1^{(ij)} \sim \mathcal{IG}(a_{ij}^{(\phi_1)}, b_{ij}^{(\phi_1)})$ | $\xi_2^{(ij)} \sim \mathcal{IG}(c_{ij}^{(\phi_2)}, d_{ij}^{(\phi_2)})$ |

Table 1: Prior probability distributions.

DRAFT

The augmented likelihood can be written as (see Polson et al., 2013):

$$
\begin{aligned}
p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}}^\star, \boldsymbol{\Omega}^\star | \mathbf{X}, \mathbf{W}, \boldsymbol{\vartheta}) = & \prod_{t=1}^{T} \prod_{i>j} \left[ \left( \frac{\mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}}}{1 + \mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}}} \right)^{y_{ijt}} \left( 1 - \frac{\mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}}}{1 + \mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}}} \right)^{1 - y_{ijt}} p\left(\omega_{ijt}^{(y)}\right) \right]^{z_{ijt}^\star} \\
& \times \left[ \left( \frac{\mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}}}{1 + \mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}}} \right)^{z_{ijt}^\star} \left( 1 - \frac{\mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}}}{1 + \mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}}} \right)^{1 - z_{ijt}^\star} \right] p\left(\omega_{ijt}^{(z^\star)}\right) \\
= & \prod_{t=1}^{T} \prod_{i>j} \left[ \frac{(\mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}})^{y_{ijt}}}{1 + \mathrm{e}^{\mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij}}} p\left(\omega_{ijt}^{(y)}\right) \right]^{z_{ijt}^\star} \\
& \times \left[ \frac{(\mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}})^{z_{ijt}^\star}}{1 + \mathrm{e}^{\mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij}}} \right] p\left(\omega_{ijt}^{(z^\star)}\right) , \qquad (5)
\end{aligned}
$$

which can be easily exploited to find the full conditional distributions of the parameters and latent factors. The resulting Gibbs sampling algorithm is made of the following steps:

(*a*) sample $\omega_{ijt}^{(z^\star)} \sim \mathrm{PG}(1, \mathbf{w}_{ijt}^\intercal \boldsymbol{\gamma} + \phi_{2,ij})$ and update $\boldsymbol{\gamma}$ and $\phi_{2,ij}$.

(*b*) conditional on $z_{ijt}^\star = 1$, draw $\omega_{ijt}^{(y)} \sim \mathrm{PG}(1, \mathbf{x}_{ijt}^\intercal \boldsymbol{\beta} + \phi_{1,ij})$ and update $\boldsymbol{\beta}$ and $\phi_{1,ij}$.

These steps are updates of the Bayesian logistic regression model, with a slight modification when $z_{ijt}^\star = 1$. In particular, we apply the standard update results for $\boldsymbol{\beta}$ and $\phi_{1,ij}$, if $z_{ijt}^\star = 1$ for each pair $(i, j)$, at time $t$. Regarding the random effect $\phi_{1,ij}$, we simulate from its prior when a specific pair of nodes $(i, j)$ does not belong to the network.

DRAFT

## 4. Preliminary simulation study

We perform a simulation study to evaluate the empirical performance of the proposed algorithm. In particular, we are interested in the full conditional distributions of the parameters for the edges model when the topology of the nodes changes over time. The main goal of this simulation study is to compare our method, which takes into account the dynamic network's topology with a standard approach that ignores this aspect. The detailed setting of the simulation study is reported below.

### 4.1 Simulation setup

We perform a simulation study by drawing a sample of $T = 100$ observations for $V_{\max} = 20$ nodes, according to the model defined in equation (4), in which the lagged variable $z_{ij\,t-1}^\star$ was included among the covariates of the nodes. In particular, we have fixed $\boldsymbol{\gamma} = (-0.8, 0.2, 0.8)^\intercal$, $\boldsymbol{\beta} = (2, 1)^\intercal$ and $\xi_1^{(ij)} = \xi_2^{(ij)} = 1$ for the random effect variances. These values produce around $51\%$ of zeros on the network topology.

Table 2: Summaries of the posterior distribution for the parameters

| | Mean | Std. Dev | $\mathrm{CI}_{95\%}$ |
|---|---|---|---|
| $\beta_0$ | 1.93 | 0.13 | (1.68,2.18) |
| $\beta_1$ | 0.97 | 0.06 | (0.86,1.08) |
| $\gamma_0$ | -0.80 | 0.08 | (-0.97,-0.64) |
| $\gamma_1$ | 0.18 | 0.04 | (0.14,0.21) |
| $\gamma_2$ | 0.81 | 0.04 | (0.75,0.89) |
| $\xi_1^{(ij)}$ | 1.00 | 0.11 | (0.80,1.24) |
| $\xi_2^{(ij)}$ | 0.98 | 0.19 | (0.64, 1.38) |

(a) Toplogy correction.

| | Mean | Std. Dev | $\mathrm{CI}_{95\%}$ |
|---|---|---|---|
| $\beta_0$ | -0.32 | 0.10 | (-0.50,-0.13) |
| $\beta_1$ | 0.08 | 0.02 | (0.05, 0.11) |
| $\gamma_0$ | -0.79 | 0.09 | (-0.97,-0.61) |
| $\gamma_1$ | 0.18 | 0.02 | (0.14,0.21) |
| $\gamma_2$ | 0.79 | 0.04 | (0.72,0.86) |
| $\xi_1^{(ij)}$ | 1.07 | 0.12 | (0.84, 1.31) |
| $\xi_2^{(ij)}$ | 1.29 | 0.14 | (1.02, 1.58) |

(b) Without correction.

For posterior inference, we ran the Gibbs sampler for $R = 30000$ iterations, discarding the first 5000 as burn-in and thinned the iterations by keeping every $10^{th}$ sample. Table (2) shows some statistics for our MCMC posterior sample, both for our method which takes into account the topology structure of the network and for the same one, which ignores this aspect.

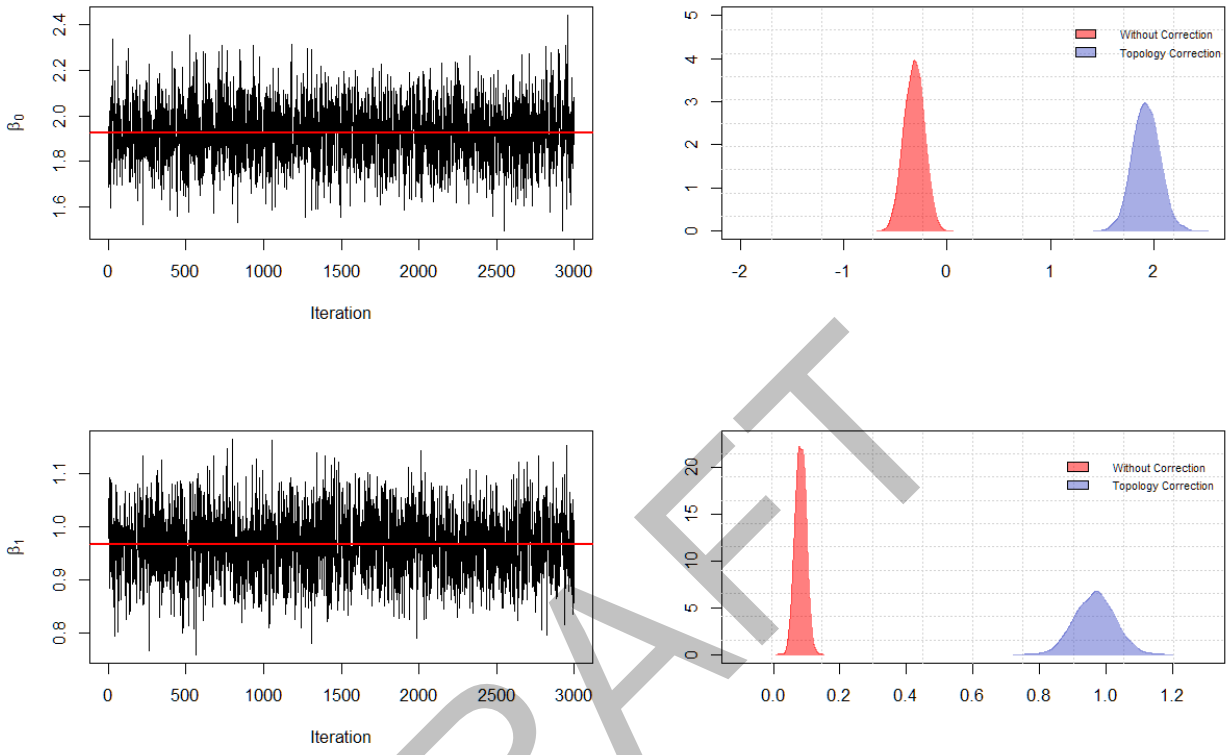



Figure 1: Trace plots for $(\beta_0, \beta_1)$ (topology correction), with a true value in red (left panel) and density plots of full conditionals distribution for $(\beta_0, \beta_1)$, where the red one is related to a model without correction and the blue one is referred to a model with a topology correction.

Figure (1) (right panel) compares the full conditionals distribution for both methods, showing that our methods provide more concentrated distributions around the true values with slight additional variability. These results suggest that even for moderate sample sizes with a large percentage of zeros, the proposed Bayesian approach provides unbiased estimates for $\boldsymbol{\beta}$, emphasising that the network topology is an important aspect to take into account. Clearly, if $\lambda_{ijt} \to 1$ this bias vanishes.

## 5. Concluding remarks

We propose a very general approach for modelling a fully dynamic network characterized by vertex and edge dynamics. In particular, using a zero-inflated Bernoulli representation we are able to capture the network connectivity patterns, taking into account the entire topology of the dynamic network. We have performed a preliminary simulation study to assess our model specification and proposed inference. The results show that if we do not consider the node's behaviour, the parameters associated with link probabilities results are completely underestimated. This novel class of methods can be extended by adding a temporal latent factor on the nodes and edges dynamics, including for example a latent space component, which embeds nodes in a lower dimensional space. Finally, we'll apply our method to real temporal network datasets in order to study their dynamic behaviour over time.

DRAFT

## References


[1] Durante, D. and Dunson, D.B.: Nonparametric Bayes Dynamic Modelling of Relational Data. Biometrika, **101**, 4, 883–98 (2014)

[2] Goldenberg, A., Fienberg, S. E., Zheng, A. X. and Airoldi, E. M.: A survey of statistical network models. Foundations and Trends in Machine Learning, **2**, 129–233 (2009)

[3] Kim B, Lee KH, Xue L, Niu X.: A review of dynamic network models with latent variables. Statistics surveys, **12**, 105-135 (2018).

[4] Newman, M. E. J.: Networks: an introduction. Oxford University Press, New York (2018)

[5] Polson, N. G., Scott, J. G., and Windle, J.: Bayesian inference for logistic models using Pólya–Gamma latent variables. Journal of the American Statistical Association, **108** (504), 1339–1349 (2013)


DRAFT

# Time lapse analysis of nuclear calcium spiking in plant cells during symbiotic signaling

Ivan Sciascia[a], Andrea Crosino[a], and Andrea Genre[a]

[a]Università di Torino Department of Life Sciences and Systems biology;
ivan.sciascia@unito.it

**Abstract**

The analysis of the time lapse series of nuclear calcium concentration obtained with FRET-based confocal microscopy imaging was performed with spectral analysis, to look for periodicities in the calcium peaks. The results demonstrate that for our dataset it is not possible to distinguish a periodicity in the calcium peaks. Further studies with larger datasets will better investigate this result for confocal microscopy time lapse analysis.

***Keywords:*** time-lapse, microscopy, spectral analysis

DRAFT

## 1. Introduction

The confocal microscope is a fluorescence microscope that can perform tomographic scans of biological samples, focusing on individual optical planes (optical sections). As a result, the confocal microscope outputs a three-dimensional high resolution image even from relatively thick samples, including living organisms. As such, confocal imaging has been widely used not only to obtain 3D reconstructions of biological tissues and organs, but also for the live imaging of cellular and physiological processes as they develop in living cells. This is the case for calcium-mediated signaling, a very common mechanism of signal transduction across eukaryotes. In plants, a remarkable example of such calcium-mediated signals has long been studied in root symbiotic interactions with nitrogen-fixing bacteria and arbuscular mycorrhizal fungi [9]. In both cases, the early perception of microbe-derived molecules by the root epidermal cells triggers prolonged oscillations in nuclear calcium concentration, known as calcium spiking and required for the appropriate regulation of gene expression and cellular as well as systemic responses that allow symbiosis establishment [1], [3]. In this study we have analyzed nuclear calcium spiking traces recorded in root epidermal cells after their stimulation with short-chain chito-oligosaccharides, the so-called Myc-factors that arbuscular mycorrhizal fungi release to alert host plants of their vicinity. Such calcium spiking traces are characterized by a series of peaks in calcium concentration that start a few minutes after Myc-factor application and continue for over 30 minutes [12]. Intriguingly, the peak sequence is markedly irregular, which has previously been proposed as a characteristic signature of Myc-factor induced signaling, compared to the more regulat spiking described in legumes in response to rhizobial Nod-factors.

## 2. Materials and Methods

Signal molecules included exudates from germinated arbuscular mycorrhizal (AM) fungal spores (GSE), generally referred to as Myc factors, activate a symbiotic signalling pathway that includesnucleus-

associated Ca2+ signals and trigger plant symbiotic responses [3], such as transcriptional regulation, lateral root formation and starch accumulation in roots ([11], [8], [10]). These pre-symbiotic responses [3] have been defined as part of an anticipation program, which prepares the host for a successful association [5] .In our case, we use the confocal microscope to detect variations in Ca2+ ions concentration within nuclei of Medicago truncatula root cells after chito-oligosaccharides treatment[2]. Recent studies on AM fungal exudates characterized chito-oligosasccharides as bioactive molecules responsible for the activation of AM symbiotic responses in the host plant. These molecules trigger responses [9] in the host roots, like Ca2+ spiking and time-series analyzes have been performed in several studies [12], [6], [7], [4].

We used transformed root organ cultures (ROCs) of M. truncatula expressing the NupYC construct. The YC construct is a chimeric protein obtained from fusion between a fluorescent protein with blue light emission spectrum (ECFP), a Ca2+ sensitive calmodulin domain (CaM), the M13 spacer peptide and a second fluorescent protein with yellow light emission spectrum (EYFP). YC construct can detect variation of Ca2+ concentration in a range from 10-8 to 10-2 M by a phenomenon called FRET (Förster Resonance Energy Transfer): resonance energy transfer between fluorophores, that occurs when donor fluorophore emission wavelength and acceptor fluorophore absorption wavelength substantially overlap. When fluorophores are a few dozen Å far from each other, the donor excitement causes resonance energy transfer to the acceptor, which is therefore the only molecule to emit fluorescence. In the YC construct, this process is used to visualize variations in cellular Ca2+ concentration: in Ca2+ lacking, the fluorophores are not such close to give rise to FRET phenomenon, they thus absorb and emit light independently; but in presence of cellular ca2+, it binds to CaM, which folds up moving close the 2 fluorophores ECFP and EYFP and allowing FRET; a downturn in the blue fluorescence and an increase in the yellow one will be observed. For our analyses we used a nucleoplasmin-tagged cameleon (NupYC2.1), which permits to observe the variations of ca2+ within the cellular nuclei. Confocal microscope analyses were conducted treating excised 1-2cm-long lateral roots with 10-6 M chito-oligosaccharides (COs) solution, sterile distilled water was used as negative control. Each root was excised and placed in a silicon micro-chamber (Invitrogen) to avoid the root damage caused by compression against the coverslip glass. A sterile distilled water drop was added for preventing root tissues desiccation. A Leica TCS SP2 AOBS confocal laser-scanning microscope, equipped with a 40x water-immersion objective, was used for these experiments. Roots were first observed in bright field to identify epidermal cells devoid of root hairs (atrichoblasts), focusing on their nuclei. Images were recorded setting the pinhole diameter at 6 Airy units, to obtain high-thickness optical sections in order to entirely include the diameter of focused nuclei. After balancing the intensity level of YFP and CFP, water on microchamber was then removed employing filter paper and substituted with 200 $\mu$l of CO solution using a micro-pipette. Images were scanned at a resolution of 512 X 512 pixels and collected every 5 seconds over a period of 30 min after treatment. The average intensity of yellow and cyan emitted fluorescence was recorded for each nucleus. Such values were exported in a .txt file, which was then elaborated in Microsoft Excel to calculate the ratio between yellow and cyan fluorescence: this ratio corresponds to FRET intensity within the YC protein and was plotted over time to visualize variations in ca2+ concentration for each nucleus during the 30 minutes acquisition.

DRAFT

## 3. Results

The results of the analyzes can be broken down into disaggregated analyzes and aggregate analyses. In the disaggregate analyses, described in Figure 1 the nuclei that make spike peaks are analyzed by single portion of root and treatment, while the aggragated data allow the analysis of the Ca2+ peaks by sum of roots for each single treatment. The data were considered for the WT and DMI2 experimental sets by proceeding first with the autocorrelation analysis (autocorrelation function, ACF) and then with the cross correlation analysis (cross correlation function, CCF). The results demonstrate autocorrelation within the same series as in Figure 3, decreased between series of different experimental sets as described in Figure 4. The periodogram useful for identifying periodicity demonstrates the presence of an initial

peak indicating that the series do not appear to have a periodicity as described in Figure 2.

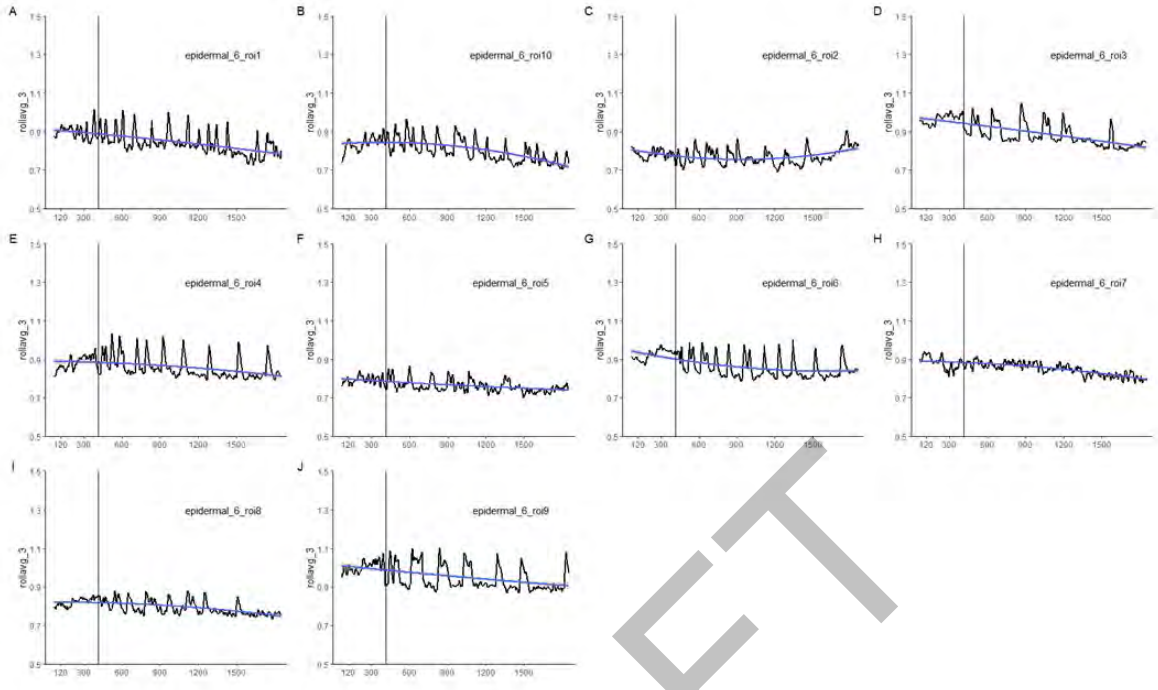



Figure 1: Calcium spike time series

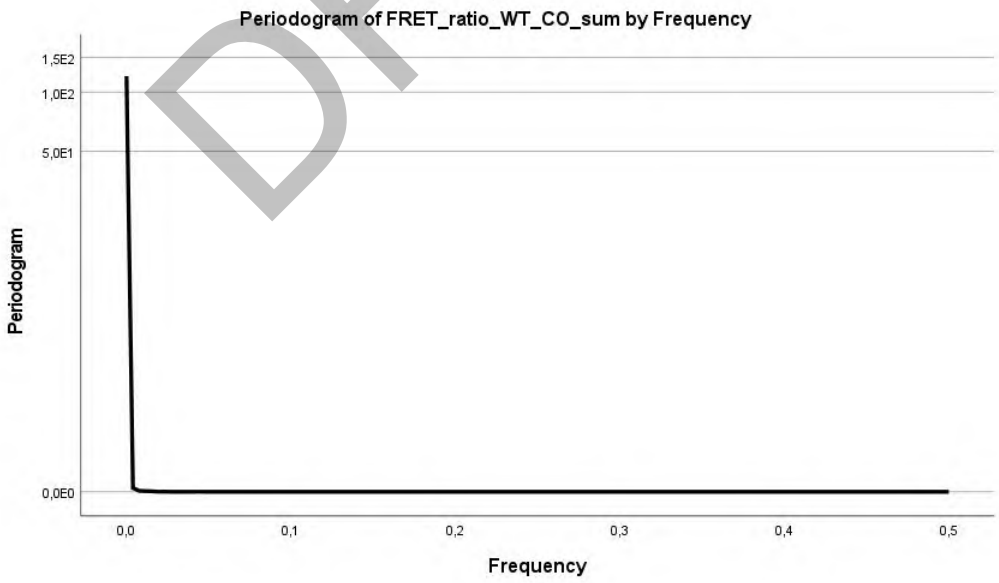



Figure 2: Periodogram

DRAFT

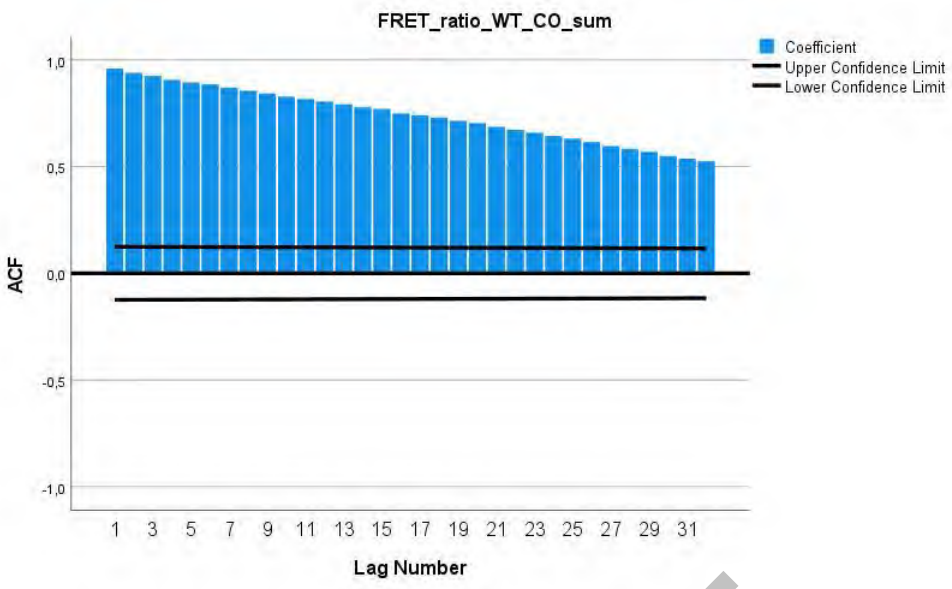



Figure 3: Autocorrelation function WT CO

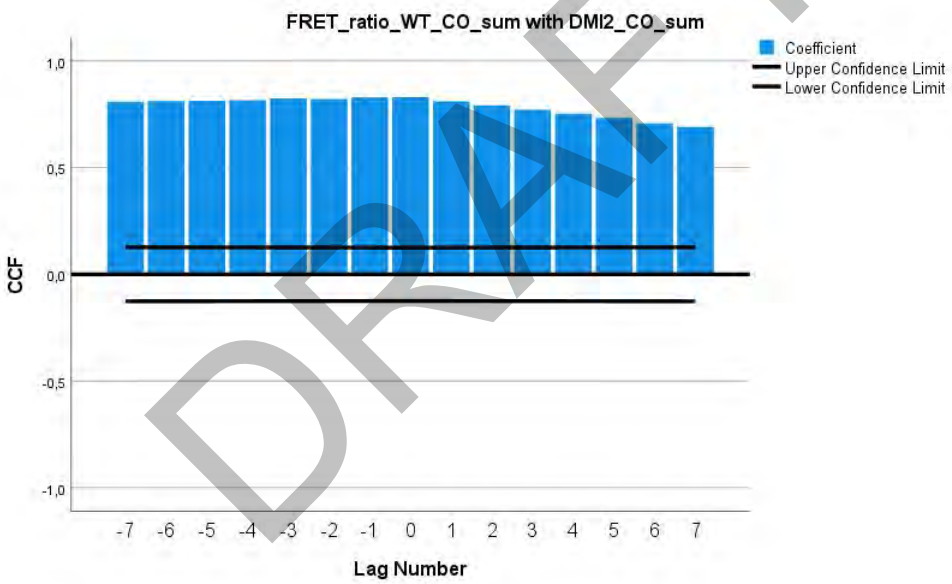



Figure 4: Cross correlation function WT CO and DMI 2 CO

DRAFT

## References

[1] Bonfante P., Requena N. Dating in the dark: how roots respond to fungal signals to establish arbuscular mycorrhizal symbiosis. Curr. Opin. Plant Biol. 14, 451-457 (2011)

[2] Chabaud M., Genre A., Sieberer B. J., Faccio A., Fournier J., Novero M., Barker D. G., Bonfante P. Arbuscular mycorrhizal hyphopodia and germinated spore exudates trigger Ca 2+ spiking in the legume and nonlegume root epidermis. New Phytol. 189: 347-355 (2011)

[3] Genre A, Chabaud M, Balzergue C, Puech Pages V, Novero M, Rey T, Fournier J, Rochange S, Becard G, Bonfante P, Barker DG. Short chain chitin oligomers from arbuscular mycorrhizal fungi trigger nuclear Ca2+ spiking in Medicago truncatula roots and their production is enhanced by strigolactone. New Phytol 198: 190-202 (2013)

[4] Granqvist, E., Oldroyd, G.E., Morris, R.J. Automated Bayesian model development for frequency detection in biological time series. BMC Syst Biol 5, 97 (2011) https://doi.org/10.1186/1752-0509-5-97
[5] Gutjahr C. and Parniske M. Cell and Developmental Biology of Arbuscular Mycorrhiza Symbiosis. Anna. Rev. Cell Dev. Biol. 29: 593-617 (2013)
[6] Kosuta S, Chabaud M, Gough C, De J, Barker DG, Bécard G . A diffusible factor from arbuscular mycorrhizal fungi induces symbiosis specific MtENOD11 expression in roots of Medicago truncatula. Plant Physiol 131: 952-962 (2003)
[7] Kosuta S, Hazledine S, Sun J, Miwa H, Morris RJ, et al. Differential and chaotic calcium signatures in the symbiosis signaling pathway of legumes. Proc. Natl. Acad. Sci. USA 105: 9823-9828 (2008)
[8] Kuhn H, Küster H, Requena N . Membrane steroid binding protein 1 induced by a diffusible fungal signal is critical for mycorrhization in Medicago truncatula. New Phytol 185: 716-733 (2010)
[9] Maillet F., Poinsot V., André O., Puech- Pagès V., Haouy A., Gueurnier M., Cromer L., Giraudet D., Formey D., Niebel A., Andres Martinez E., Driguez H., Bécard G., Dénarié J. . Fungal lipochitooligosaccharide symbiotic signals in arbuscular mycorrhiza. Nature. 469: 58-64 (2011)
[10] Mukherjee A., Ane J-M. Germinating spore exudates from arbuscular mycorrhizal fungi: molecular and developmental responses in plants and their regulation by ethylene. Mol. Plant-Microbe Interact. 24: 260-270 (2011)
[11] Olah B., Brière C., Bécard G., Dénarié J., Gough C. Nod factors and a diffusible factor from arbuscular mycorrhizal fungi stimulate lateral root formation in Medicago truncatula via the DMI1/DMI2 signalling pathway. The Plant Journal 44: 195-207 (2005)
[12] Russo, G., Spinella, S., Sciacca, E., Bonfante, P., Genre, A. Automated analysis of calcium spiking profiles with CaSA software: two case studies from root-microbe symbioses. BMC Plant Biol 13, 224 (2013)

DRAFT

# Two-stage weighted least squares estimator of multivariate conditional mean observation-driven time series models

Mirko Armillotta[a]

[a]Department of Econometrics and Data Science, Vrije Universiteit Amsterdam;
m.armillotta@vu.nl

**Abstract**

We introduce a general parametric multivariate model where the first two conditional moments are assumed to be multivariate time series. The focus of the estimation is the conditional mean parameter vector. Quasi-Maximum Likelihood Estimators (QMLEs) based on the linear exponential family are typically employed for such estimation problems when the true multivariate conditional probability distribution is unknown or too complex. Although QMLEs provide consistent estimates they may be inefficient. In this paper we study a two-stage Multivariate Weighted Least Square Estimators (MWLSEs), which enjoy the same consistency property as the QMLEs and provides efficiency gain with suitable choice of the covariance matrix. We compare the estimation performance of the QMLEs and MWLSEs through simulation experiments.

***Keywords:*** Quasi-likelihood, Multivariate Integer-valued GARCH, Multivariate Poisson.

DRAFT

## 1. Inefficiency of QMLE

Consider a multivariate stationary and ergodic sequence of real-valued processes $\{\mathbf{Y}_t\}_{t\in\mathbf{Z}}$ with first conditional moment given by

$$\mathrm{E}(\mathbf{Y}_t|\mathcal{F}_{t-1}) = \lambda(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots; \boldsymbol{\theta}_0) = \boldsymbol{\lambda}_t(\boldsymbol{\theta}_0)\,, \quad t \in \mathbf{Z} \tag{1}$$

where $\mathbf{Z}$ is the set of integers, $\mathcal{F}_t$ denotes the $\sigma$-field generated by $\{\mathbf{Y}_s\,,\ s \le t\}$, $\lambda : \mathbf{R}^\infty \times \Theta \to \mathbf{R}^d$ is a known measurable function, with $\boldsymbol{\theta}_0$ unknown parameter vector belonging to some compact parameter space $\Theta \subset \mathbf{R}^m$. The vector $\boldsymbol{\theta}_0$ is quite general as it is not constrained to have a specific structure; Sometimes the specification of model (1) can be completed by setting a function of the data $\mathbf{Y}_t = f(\boldsymbol{\lambda}_t, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}_0)$ where $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbf{Z}}$ is an iid sequence. It is the case of the Vector Autoregressive Moving-Average models (VARMA), see Lütkepohl (8, Ch. 11), or also the Multivariate Generalized Autoregressive Conditional Heteroskedasticity model (MGARCH); see Francq and Zakoian (5, Ch. 10). However, although the Maximum Likelihood Estimator (MLE) is sometimes readily computable when the joint conditional distribution function (cdf) can be specified (as in the case of VARMA and MGARCH models) this may not always be the case. For example, when count time series models are analyzed, for models encompassed in eq. (1) such as the Multivariate Integer-valued GARCH (MINGARCH), Fokianos et al. (4), a Poisson distribution can be specified for the marginal probability mass functions (pmf) of the univariate process $Y_{i,t}$, for $i = 1, \ldots, d$. However, to obtain a complete joint cdf a particular joint pmf needs to be specified as well. The functional form of such joint pmf can be unknown or too complex to be employed

for the MLE. Indeed, the development of a multivariate count time series model would be based on specification of a joint pmf distribution, so that the standard likelihood inference and testing procedures can be developed. Although several alternatives have been proposed in the literature, see the review in Fokianos (3, Sec. 2), the choice of a suitable multivariate version of the Poisson pmf is a challenging problem. In fact, multivariate Poisson-type pmfs have usually complicated closed form and the associated likelihood inference is theoretically and computationally cumbersome. Furthermore, in many cases, the available multivariate Poisson-type pmfs implicitly imply restricted models, which are of limited use in applications (e.g. covariances always positive, constant pairwise covariances). See Inouye et al. (6) and Fokianos (3) for a discussion on the choice of multivariate count distributions and several alternatives. Other approaches try to avoid the problem of specifying a multivariate Poisson distribution, like in Fokianos et al. (4), where the joint distribution of the vector $\{\mathbf{Y}_t\}$ is constructed by following a copula approach where all marginal distributions of $Y_{i,t}$ are univariate Poisson, conditionally to the past, with mean process $\lambda_{i,t}$, for $i = 1, \dots, d$. However, this approach does not allow to perform a ML estimation but requires appealing the class of Quasi-Maximum Likelihood Estimators (QMLEs) based on the linear exponential family that assumes independence between the univariate processes $Y_{i,t}$, for $i = 1, \dots, d$, see eq. (10) in Fokianos et al. (4). The QML estimator $\hat{\boldsymbol{\theta}}_Q$ can be seen as the vector which solves the first order conditions that typically take the form

$$\mathbf{S}_T^Q(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \mathbf{D}_t^{-1}(\boldsymbol{\theta})\Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big) = 0\,, \tag{2}$$

where $\mathbf{S}_T^Q(\boldsymbol{\theta})$ is the score of the QMLEs, $\mathbf{D}_t(\boldsymbol{\theta})$ is the $d \times d$ diagonal matrix with diagonal elements equal to $\nu_{i,t}(\boldsymbol{\theta})$ for $i = 1, \dots, d$ and $\nu_{i,t}(\boldsymbol{\theta}_0)$ is defined in

$$\mathbf{V}(Y_t|\mathcal{F}_{t-1}) = \boldsymbol{\nu}(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots; \boldsymbol{\theta}_0) = \boldsymbol{\nu}_t(\boldsymbol{\theta}_0)\,, \quad t \in \mathbf{Z} \tag{3}$$

which is the vector of conditional variances of the marginal process $Y_{i,t}$ with $\nu : \mathbf{R}^\infty \times \Theta \to [0, +\infty)^d$ being a measurable function. From the property of the linear exponential family we often have that the variance model depends on the mean model, i.e. $\nu_{i,t}(\boldsymbol{\theta}_0) = g(\lambda_{i,t}(\boldsymbol{\theta}_0))$ where $g(\cdot)$ is some measurable function. The possible value of $\nu_{i,t}$ is however restricted by the fact that it must match the conditional variance of an exponential family distribution. For example, in the QMLE employed in Fokianos et al. (4) all marginal distributions of $Y_{i,t}$ are univariate Poisson, conditionally to the past, so the variances processes are $\nu_{i,t}(\boldsymbol{\theta}_0) = \lambda_{i,t}(\boldsymbol{\theta}_0)$, for $i = 1, \dots, d$. In practice, one can easily conceive that the conditional variance may have other forms. Obviously, the choice of a wrong marginal pmf, or probability density function (pdf), may affect the efficiency of the QMLEs. Moreover, it is clear that the score of the QMLE defined in (2) does not directly include any joint dependence between the univariate processes because is computed upon a quasi-likelihood which assumes contemporaneous independence among $Y_{i,t}$, for $i = 1, \dots, d$. Then the QMLEs suffer from two main sources of inefficiency: constrained variance specification and exclusion of joint dependence parameters.

DRAFT

## 2. Two-stage Multivariate Weighted Least Square Estimators

The present contribution proposes a novel class of two-stage Multivariate Weighted Least Square Estimators (MWLSEs) of the mean parameters $\boldsymbol{\theta}_0$. Given a theoretical matrix weight function $\mathbf{W}_t = \mathbf{W}(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots)$, where $\mathbf{W}$ is a measurable function from $\mathbf{R}^\infty \to \mathbf{R}^{d\times d}$ a first-stage MWLSE is defined by $\hat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}\in\Theta} l_T(\boldsymbol{\theta}, \mathbf{W})$ where

$$l_T(\boldsymbol{\theta}, \mathbf{W}) = \frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta}, \mathbf{W}_t)\,, \qquad l_t(\boldsymbol{\theta}, \mathbf{W}_t) = \Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big)' \mathbf{W}_t \Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big)\,. \tag{4}$$

When $\mathbf{W}_t = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix of suitable dimension, (4) corresponds to the multivariate version of the Conditional Least Squares (CLS) estimator of Klimko and Nelson (7). It is well-known

that the optimal choice of the weights $\mathbf{W}_t$ is inverse of the true conditional covariance matrix of the process, $(\tilde{\mathbf{V}}_t^{-1})_{t\geq 1}$. In practice, the value of $\tilde{\mathbf{V}}_t^{-1}$ is generally unknown. We assume for the conditional covariance matrix a parametric specification of the form $\mathbf{V}_t = \mathbf{V}(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots; \boldsymbol{\tau}_0)$, where $\mathbf{V}$ is a measurable function defined in $\mathbf{R}^{\infty} \to \mathbf{R}^{d\times d}$ and the true parameter vector $\boldsymbol{\tau}_0$ may contain $\boldsymbol{\theta}_0$ or some of its elements. Then, the optimal sequence of weights is estimated by

$$\hat{\mathbf{W}}_t = \mathbf{V}_t(\hat{\boldsymbol{\tau}})^{-1} = \mathbf{D}_t(\hat{\boldsymbol{\tau}})^{-1/2}\mathbf{P}(\hat{\boldsymbol{\tau}})^{-1}\mathbf{D}_t(\hat{\boldsymbol{\tau}})^{-1/2} \tag{5}$$

where, analogously to Section 1, $\mathbf{D}_t(\hat{\boldsymbol{\tau}})$ is the $d \times d$ diagonal matrix with diagonal elements equal to $\nu_{i,t}(\hat{\boldsymbol{\tau}})$ for $i = 1, \ldots, d$ and $\nu_{i,t}(\hat{\boldsymbol{\tau}})$ is defined as in (3) with obvious rearrangement of the notation and $\mathbf{P}(\hat{\boldsymbol{\tau}})$ is a $d \times d$ correlation matrix whose structure can be defined by the researcher and whose single element has a functional form of the type $r_{i,j}(\hat{\boldsymbol{\tau}}) = r(\mathbf{Y}_1, \ldots, \mathbf{Y}_T; \hat{\boldsymbol{\tau}})$ for $i, j = 1, \ldots, d$ where $r : \mathbf{R}^T \to (-1, 1)$. The estimator $\hat{\boldsymbol{\tau}}$ will be a function of the estimator $\hat{\boldsymbol{\theta}}_1$ of $\boldsymbol{\theta}_0$, and possibly of estimates of some extra parameters $\boldsymbol{\gamma}_0$. In this way $\hat{\boldsymbol{\tau}}$ represents a first-step estimator of $\boldsymbol{\tau}_0$ and the two-stage WLSE is given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\, l_T(\boldsymbol{\theta}, \hat{\mathbf{W}})\,, \qquad l_T(\boldsymbol{\theta}, \hat{\mathbf{W}}) = \frac{1}{T}\sum_{t=1}^{T} l_t(\boldsymbol{\theta}, \hat{\mathbf{W}}_t) \tag{6}$$

where $l_t(\boldsymbol{\theta}, \hat{\mathbf{W}}_t)$ is defined analogously to (4) and $\hat{\mathbf{W}}_t$ has the form (5). We can see that the MWLSE defined in (6) is the solution of the first order condition system

$$\mathbf{S}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\mathbf{V}_t(\hat{\boldsymbol{\tau}})^{-1}\Big(\mathbf{Y}_t - \boldsymbol{\lambda}_t(\boldsymbol{\theta})\Big) = 0\,. \tag{7}$$

The MWLSEs do require only the correct specification of the first conditional moment of the process instead of entirely specifying its conditional distribution. In this sense, they are semi-parametric estimators because, except for the first moment, they are totally agnostic about the distribution of the observations. In this way the estimation is robust against possible misspecification of the conditional distribution and simplifies the inference when the full MLE is undoable or too complex.

The first order conditions (2) and (7) allow for a direct and intuitive comparison between QMLEs and MWLSEs. The first main difference is that in (7) there is no particular constraint on the shape of the univariate conditional variances, $\nu_{i,t}(\hat{\boldsymbol{\tau}})$ whereas in (2) the form of the variances are constrained by the univariate conditional distribution specified in the quasi-likelihood. Consequently, the MWLSE adds more flexibility in the choice of the variances form and this can lead to an improved efficiency estimator. A second difference lies in the possibility of including correlation effects in the estimation system (7) whereas the QMLE simply implies no correlation in the score (2), i.e. $\mathbf{P}^{-1} = \mathbf{I}$. Intuitively, the MWLSE is expected to explain a larger part of the variability connected to the process of study. We will show with a simulated experiment that when a moderate or high degree of contemporaneous dependence is indeed present among the variables $Y_{i,t}$ for $i = 1, \ldots, d$ the MWLSE entails a relevant gain in efficiency with respect to the QMLE.

## 3. An example with Poisson data

We apply the methodology discussed in the previous section to the Multivariate INGARCH model with Poisson data

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \mathbf{c} + \mathbf{A}\boldsymbol{\lambda}_{t-1} + \mathbf{B}\mathbf{Y}_{t-1}\,, \tag{8}$$

where $\mathbf{c}$ is a vectorial intercept of positive unknown parameters and $\mathbf{A}$ and $\mathbf{B}$ are $d \times d$ unknown parameter matrices assumed to be non-negative such that $\lambda_{i,t} > 0$ for all $i$ and $t$. $\mathbf{N}_t(\boldsymbol{\lambda}_t)$ is a sequence of $d$-dimensional iid marginally Poisson count processes, with intensity 1, counting the number of events in the interval of time $[0, \lambda_{1,t}] \times \cdots \times [0, \lambda_{d,t}]$ and whose structure of dependence is modelled through

DRAFT

a copula construction $C(\dots)$ on their associated exponential waiting times random variables. The complete structure of the data generating process (dgp) is described in Fokianos et al. (4, p. 474). For the scopes of our example it is just important to state that the joint conditional pmf resulting from (8) does not have a closed form (so the MLE is unfeasible) but the univariate conditional pmfs $q(\cdot)$ are Poisson, i.e. $Y_{i,t}|\mathcal{F}_{t-1} \sim q(\lambda_{i,t}(\boldsymbol{\theta}_0))$, for $i = 1, \dots, d$ and the contemporaneous dependence between the variables $Y_{i,t}$ is modelled through a copula $C(\dots; \rho)$ which depends on one unknown parameter, say $\rho$, capturing the contemporaneous correlation among the variables. Since the copula construction in the dgp is imposed on latent variables (exponential waiting times) it cannot be included in the likelihood estimation and a Poisson QMLE of the type $l_T^P(\theta) = T^{-1} \sum_{t=1}^{T} \sum_{i=1}^{d} q(\lambda_{i,t}(\boldsymbol{\theta}_0))$ has been carried out in the previous literature; see Fokianos et al. (4, eq. 10), Armillotta and Fokianos (1) and Armillotta and Fokianos (2), among others. The associated QML estimator $\hat{\boldsymbol{\theta}}_P$ is the vector which maximizes $l_T^P(\theta)$ and clearly solves the first order system (2) where the elements on the diagonal matrix $\mathbf{D}_t(\hat{\boldsymbol{\theta}}_P)$ have the form $\lambda_{i,t}(\hat{\boldsymbol{\theta}}_P)$.

We summarize the steps required in order to carry out the two-stage MWLSE.

1. Estimate $\hat{\boldsymbol{\theta}}_1$, the first-stage MWLSE (4) of model (8) by setting $\mathbf{W}_t = \mathbf{I}$.
2. Choose a functional form for $\nu_{i,t}(\boldsymbol{\tau})$.
3. Choose a structure for $\mathbf{P}(\boldsymbol{\tau})$ and a functional form for $r_{i,j}(\boldsymbol{\tau})$.
4. Estimate $\hat{\boldsymbol{\theta}}$ the second-stage MWLSE (6) of model (8) by setting $\hat{\mathbf{W}}_t = \mathbf{V}_t(\hat{\boldsymbol{\theta}}_1)^{-1}$.

For the first-stage estimation a reasonable weighting scheme in absence of any information will be $\mathbf{W}_t = \mathbf{I}$ so that $\hat{\boldsymbol{\theta}}_1$ represents the CLS estimator. For the accomplishment of the second-stage estimation we need to define a suitable weight matrix (5), which is translated in steps 2-3 of the algorithm above. Since the data are marginally Poisson-distributed we set $\nu_{i,t}(\boldsymbol{\tau}) = \lambda_{i,t}(\boldsymbol{\tau})$. In this case no additional parameters $\boldsymbol{\gamma}$ are employed so we have $\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\theta}}_1$.

Reminding that the copula construction in the dgp is imposed on latent variables and since, in practice, the dependence structure among the integer-valued random variables is completely determined by the copula parameter $\rho$, the true structure of the conditional correlation matrix $\mathbf{P}(\tau)$ is unknown. In this cases, similarly to the theory of Generalized Estimating Equation (GEE) Zeger and Liang (12), the researcher may choose a working correlation structure for the estimation. There exist several suitable alternatives in the literature; see for example Pan (9, p. 121). Sometimes the inversion of the $d \times d$ matrix $\mathbf{P}(\cdot)$ may be unfeasible for numerical reasons. Furthermore, since only the inverse of such matrix would be required by the proposed estimating procedure, we suggest to pick a correlation structure where an analytical form for the inverse is known. For example, in this work we decide to impose a simple structure like the equicorrelation structure (EQC), $\mathbf{P}(\boldsymbol{\tau}) = (1 - r(\boldsymbol{\tau}))\mathbf{I} + r(\boldsymbol{\tau})\mathbf{J}$, where $\mathbf{J}$ is a $d \times d$ matrix of ones and an equal pairwise correlation $r_{i,j}(\boldsymbol{\tau}) = r(\boldsymbol{\tau})$ is assigned to all the possible couples of the multivariate time series. Such working correlation matrix is appealing since it has an analytical form of the inverse, i.e. $\mathbf{P}^{-1}(\boldsymbol{\tau}) = (a - b)\mathbf{I} + b\mathbf{J}$, with $a = [1 + (d-2)r] / \{(1-r)[1 + (d-1)r]\}$ and $b = -r / \{(1-r)[1 + (d-1)r]\}$, where we set for simplicity $r(\boldsymbol{\tau}) = r$. For a proof see Rao (10, p. 67)

Borrowing the idea of GEE, the form of the the correlation parameter can be tackled using moment estimators types functions, see for example Zeger and Liang (12, Sec. 4). We consider two alternatives, $\hat{r}_1 = 2 \sum_{i=1}^{d} \sum_{j>i} \hat{r}_{ij}(\hat{\boldsymbol{\theta}}_1)$ and $\hat{r}_2 = \max_{i,j=1,\dots,d,\, j>i} \hat{r}_{ij}(\hat{\boldsymbol{\theta}}_1)$, where

$$\hat{r}_{ij}(\hat{\boldsymbol{\theta}}_1) = \frac{\sum_{t=1}^{T} [Y_{i,t} - \lambda_{i,t}(\hat{\boldsymbol{\theta}}_1)][Y_{j,t} - \lambda_{j,t}(\hat{\boldsymbol{\theta}}_1)]}{\sqrt{\sum_{t=1}^{T} [Y_{i,t} - \lambda_{i,t}(\hat{\boldsymbol{\theta}}_1)]^2} \sqrt{\sum_{t=1}^{T} [Y_{j,t} - \lambda_{j,t}(\hat{\boldsymbol{\theta}}_1)]^2}}$$

are the empirical Pearson correlations from the first-stage MWLSE estimation.

To examine the performance of the QMLEs and MWLSEs we run a simulation study, by generating data from model (8), with different dimension $d$ and sample size $T$, using the algorithm depicted in Fokianos et al. (4, p. 474) and the Gaussian copula for $C(\dots; \rho)$; data are produced by employing a copula parameter $\rho$ selected by an equidistant grid of values in the interval $[0.3, 0.9]$. Then, we compare the estimation performances of the QMLE versus the MWLSE, by measuring their relative efficiency, with the relative Mean Square Error, $e(\hat{\boldsymbol{\theta}}_P, \hat{\boldsymbol{\theta}}) = \sum_{s=1}^{S} ||\hat{\boldsymbol{\theta}}_{P,s} - \boldsymbol{\theta}||^2 / \sum_{s=1}^{S} ||\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}||^2$, where $S = 500$

DRAFT

is the number of simulations performed and $\hat{\boldsymbol{\theta}}_s$ is the estimator associated with the replication $s$. Clearly, $e(\hat{\boldsymbol{\theta}}_P, \hat{\boldsymbol{\theta}}) > 1$ shows improved efficiency of $\hat{\boldsymbol{\theta}}$ when compared to $\hat{\boldsymbol{\theta}}_P$. The same comparison can be done marginally for each parameter of the model $e(\hat{\boldsymbol{\theta}}_{P,h}, \hat{\boldsymbol{\theta}}_h) = \sum_{s=1}^{S}(\hat{\boldsymbol{\theta}}_{P,h,s} - \boldsymbol{\theta}_h)^2 / \sum_{s=1}^{S}(\hat{\boldsymbol{\theta}}_{h,s} - \boldsymbol{\theta}_h)^2$, for $h = 1, \ldots, m$. The results of the Monte Carlo simulations are summarized in Figure 1. When the

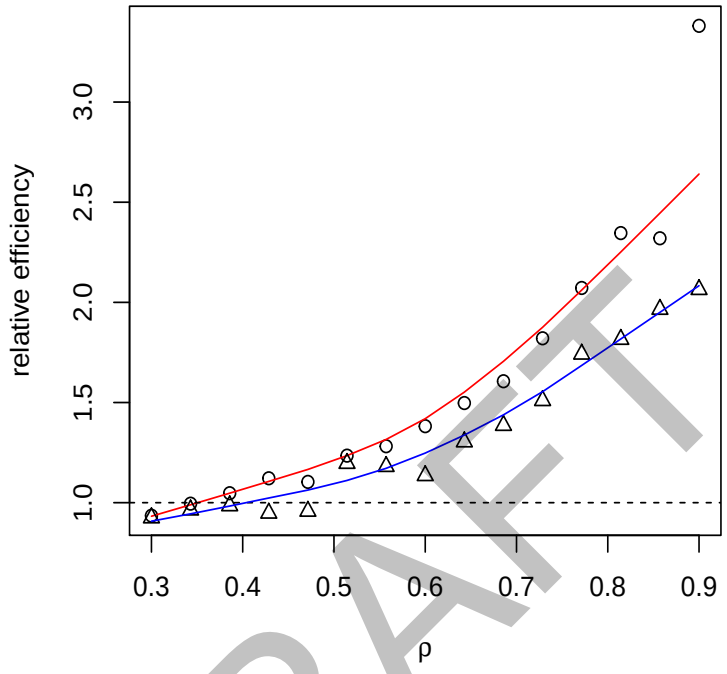



Figure 1: Plot for $e(\hat{\boldsymbol{\theta}}_P, \hat{\boldsymbol{\theta}})$ versus values of copula parameter $\rho$, with EQC working correlation matrix and $r(\hat{\boldsymbol{\theta}}_1) = \hat{r}_2$. Data generated by model (8) with 500 simulations. Triangles: $d = 2, T = 100$. Points: $d = 4, T = 200$. Blue line: LOWESS smoother at $d = 2, T = 100$. Red line: LOWESS smoother at $d = 4, T = 200$. Dashed line: horizontal line $e(\hat{\boldsymbol{\theta}}_P, \hat{\boldsymbol{\theta}}) = 1$.

DRAFT

data are generated with a low correlation ($\rho < 0.4$) the two estimator show similar performances. However, when the data are generated with a moderate or strong correlation (copula parameter $\rho > 0.4$) the MWLSE is relatively more efficient than the QMLE. The improvement in efficiency is larger as the correlation becomes stronger and tends to grow even by increasing dimension and time size $(d, T)$. This would be expected since the QMLE becomes a poor approximation of the true likelihood as the dependence structure of the multivariate count process $\mathbf{Y}_t$ becomes stronger, whereas the specified MWLSE methodology appears to be able to account for a significant part of the correlations among the counts, even though the working correlation does not reflect the true correlation structure of the data. Similar results are obtained by comparing the marginal efficiencies $e(\hat{\boldsymbol{\theta}}_{P,h}, \hat{\boldsymbol{\theta}}_h)$, therefore are omitted. The employment of the estimator $\hat{r}_1$ gave analogous results but with gain in efficiency less than the gain obtained by used $\hat{r}_2$.

Although the result of the present simulation study are encouraging, the problem of improving the efficiency of the QMLE is only at the initial stage and further studies are required. For example, alternative working correlation structures may be used, like the AR-1 correlation structure $\mathbf{P}(\boldsymbol{\tau}) = (r_{ij})$ where $r_{ij} = r^{|i-j|}$, for $i, j = 1, \ldots, d$ and $i \neq j$. This matrix also have analytical inverse, $\mathbf{P}^{-1}(\boldsymbol{\tau}) = 1/(1-r^2)\mathbf{T}(\boldsymbol{\tau})$, where $\mathbf{T}(\boldsymbol{\tau})$ is a tridiagonal matrix, with the main diagonal consisting of the elements of the $d \times 1$ vector $(1, 1+r^2 \ldots, 1+r^2, 1)$, and the remaining two diagonals are the elements of the

$(d-1) \times 1$ vector of $(-r, \ldots, -r)$; see Sutradhar and Kumar (11, eq. 1.1). Moreover, more complex correlation structures could be considered. In addition, further estimators for the correlation parameter $r$ may be available. Finally, the development of an asymptotic theory for the proposed MWLSE estimator would be of interest.

## References


[1] Armillotta, M., Fokianos, K.: Poisson network autoregression. arXiv:2104.06296 (2021)

[2] Armillotta, M., Fokianos, K.: Testing linearity for network autoregressive models. arXiv:2202.03852 (2022)

[3] Fokianos, K.: Multivariate count time series modelling. Econ. Stat. (2022, In press)

[4] Fokianos, K., Støve, B., Tjøstheim, D., Doukhan, P.: Multivariate count autoregression. Bernoulli **26**, 471–499 (2020)

[5] Francq, C., Zakoian, J.-M. GARCH Models: Structure, Statistical Inference and Financial Applications. John Wiley & Sons, Chichester (2019)

[6] Inouye, D. I., Yang, E., Allen, G. I. , Ravikumar, P.: A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdiscip. Rev. Comput. Stat. **9**, 1–25 (2017)

[7] Klimko, L. A., Nelson, P. I.: On conditional least squares estimation for stochastic processes. Ann. Stat. **6**, 629–642 (1978)

[8] Lütkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin (2005)

[9] Pan, W.: Akaike's information criterion in generalized estimating equations. Biometrics **57**, 120–125 (2001)

[10] Rao, C. R.: Linear statistical inference and its applications, Volume 2. Wiley, New York (2002)

[11] Sutradhar, B.,Kumar, P.: The inversion of correlation matrix for MA (1) process. Appl. Math. Lett. **16**, 317–321 (2003)

[12] Zeger, S. L., Liang, K.-Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics **73**, 121–130 (1986)


DRAFT

# Assessing the performance of nuclear norm-based matrix completion methods on $CO_2$ emissions data

Rodolfo Metulini[a], Francesco Biancalani[b], Giorgio Gnecco[b], and Massimo Riccaboni[b]

[a]Department of Economics, University of Bergamo; rodolfo.metulini@unibg.it
[b]Laboratory for the Analysis of CompleX Economics Systems (AXES), IMT School for Advanced Studies; francesco.biancalani@imtlucca.it; giorgio.gnecco@imtlucca.it; massimo.riccaboni@imtlucca.it

DRAFT

**Abstract**

With the aim of performing a counterfactual analysis using the statistical learning method of Matrix Completion (MC) to evaluate the impact of the Emissions Trading System (ETS) - an instrument launched by the European Union in 2005 to reduce $CO_2$ emissions and mitigate global warming - in this paper we study the performance of recently proposed nuclear norm regularized MC methods for panel data when applied to $CO_2$ emission data. Results show that the inclusion of individual and time fixed effects in the MC optimization problem, and the pre-processing of the original data, increase the performance of the method.

***Keywords:*** emissions trading system, counterfactual analysis, statistical learning, missing data, causal inference

## 1. Introduction

$CO_2$ emissions represent a rising concern in relation to pollution and climate change (7). Economic systems produce large amounts of $CO_2$ by the use of fossil energy, thereby governments are trying to address the production to new systems aimed at minimizing emissions. In this regard, the European Union implemented a market of emission rights called the Emissions Trading System (ETS) that was launched in 2005.

A counterfactual analysis for policy evaluation permits quantifying the reduction of $CO_2$ emissions due to ETS is in place. In our ongoing works we adopt an approach based on Matrix Completion (MC) (4), which is a supervised statistical learning method to reconstruct a partially incomplete matrix and whose idea relies on the minimization of a trade-off between the approximation error on a set of observed entries and a proxy of the rank of the reconstructed matrix, e.g., its nuclear norm. We adopt an MC approach because, in the absence of the ETS policy, counterfactual $CO_2$ emission levels are not known for the treated countries (namely EU countries) in the years of treatment. Hence, based on the method introduced by Athey et al. (1), we use MC to generate estimates of such counterfactual values, and compare them with the actual $CO_2$ emission levels, with the final aim to estimate the effect of treatment on $CO_2$ emission levels through the ETS policy.

The MC optimization problem proposed and theoretically investigated by Mazumder et al. (5), which belongs to the class of "nuclear norm regularized" MC methods (because the error minimization is penalized by the nuclear norm of the reconstructed matrix) and which is generally solved using the so-called Soft Impute algorithm, has been widely adopted (e.g., in (6) for the reconstruction of Input-Output tables, or in (3) to analyse economic complexity). When the matrix to reconstruct displays the

countries in rows and the years in columns (as is the case of the $CO_2$ emissions matrix considered in this work), we are in a panel data set-up. More recently, specific nuclear norm regularized MC methods solved with Soft Impute algorithm and that explicitly take into account for individual and time effects in the MC optimization problem have been proposed for panel data by Athey et al. (1).

A necessary condition for obtaining credible counterfactual results regards a satisfying performance of the MC method in reconstructing the original matrix in absence of any treatment. The aim of this paper is that of assessing and comparing the performance of the above-mentioned MC methods in reconstructing $CO_2$ emissions matrix. We develop a simulation study where different amounts of unknown entries are allowed and we evaluate the performance in terms of Mean Absolute Percentage Error (MAPE). Because a criterion for an efficient reconstruction lies in the maintenance of the original decomposition of the total deviance in within and between country deviance, an analysis based on such a decomposition in the reconstructed matrix is also conducted.

Results show that, for our simulation study, the MC method proposed by Athey et al. generally outperforms the MC method proposed by Mazumder. When the first method is applied to a $l_1$ row-normalized matrix, the performance is extremely good even for a large amount of unknown entries. The between and within deviance decomposition is also preserved when the method proposed by Athey et al. is applied to the original matrix (i.e., without $l_1$ row-normalization).

## 2. Methods

MC is a statistical method used to predict unobserved entries of a matrix in terms of the set of the remaining observed entries. Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, MC works by finding a suitable low-rank approximation of $\mathbf{M}$, by assuming the model $\mathbf{M} = \mathbf{C}\mathbf{G}^T + \mathbf{E}$, where $\mathbf{C} \in \mathbb{R}^{m \times r}$, $\mathbf{G} \in \mathbb{R}^{n \times r}$, whereas $\mathbf{E} \in \mathbb{R}^{m \times n}$ is a matrix of errors. According to (5), the rank-$r$ approximating matrix $\mathbf{C}\mathbf{G}^T$ is found by solving the following MC optimization problem based on a LASSO-like nuclear norm penalty:

$$\underset{\hat{\mathbf{M}} \in \mathbb{R}^{m \times n}}{\text{minimize}} \left( \frac{1}{|\Omega^{\text{tr}}|} \sum_{(i,j) \in \Omega^{\text{tr}}} \left( M_{i,j} - \hat{M}_{i,j} \right)^2 + \lambda \|\hat{\mathbf{M}}\|_* \right) , \tag{1}$$

where $\Omega^{\text{tr}}$ is a subset of pairs of indices $(i, j)$ corresponding to positions of known entries of $\mathbf{M}$, $\hat{\mathbf{M}}$ is the matrix to be optimized, $\lambda \geq 0$ is a regularization constant, and $\|\hat{\mathbf{M}}\|_*$ is the nuclear norm of the matrix $\hat{\mathbf{M}}$. The regularization constant $\lambda$ controls the trade-off between fitting the known entries of the matrix $\mathbf{M}$ and achieving a small nuclear norm. The "Soft Impute" algorithm has been proposed in (5) to solve equation 1. Penalization parameter $\lambda$ is chosen by first randomly dividing the set of positions of unobserved entries of the matrix $\mathbf{M}$ into a validation set $\Omega^{\text{val}}$ and a test set $\Omega^{\text{test}}$. Then, the optimization problem is solved for several choices of $\lambda$ and the optimal $\lambda$ is that corresponding to the smallest average RMSE on the validation set. Herein after, we call this method MC Baseline (`MCB`).

Athey et al. (1) propose a methodological advancement of Mazumder nuclear norm regularized MC consisting of explicitly including individual and time fixed effects in the optimization problem, specifically intended to use in panel data. The optimization problem to solve is:

$$\begin{aligned} \underset{\hat{\mathbf{L}} \in \mathbb{R}^{m \times n}, \hat{\mathbf{\Gamma}} \in \mathbb{R}^{m \times 1}, \hat{\mathbf{\Delta}} \in \mathbb{R}^{n \times 1}}{\text{minimize}} \quad & \left( \frac{1}{|\Omega^{\text{tr}}|} \sum_{(i,j) \in \Omega^{\text{tr}}} \left( M_{i,j} - \hat{M}_{i,j} \right)^2 + \lambda \|\hat{\mathbf{L}}\|_* \right) , \\ \text{subject to} \quad & \hat{\mathbf{M}} = \hat{\mathbf{L}} + \hat{\mathbf{\Gamma}} \mathbf{1}_{\boldsymbol{n}}^\top + \mathbf{1}_{\boldsymbol{m}} \hat{\Delta}^\top , \end{aligned} \tag{2}$$

where $\mathbf{1}_{\boldsymbol{n}}$ and $\mathbf{1}_{\boldsymbol{m}}$ are column vectors made respectively of $n$ entries and $m$ entries, all equal to 1; $\hat{\mathbf{M}}$ is decomposed as $\hat{\mathbf{M}} = \hat{\mathbf{L}} + \hat{\mathbf{\Gamma}} \mathbf{1}_{\boldsymbol{n}}^\top + \mathbf{1}_{\boldsymbol{m}} \hat{\Delta}^\top$ (where $\hat{\mathbf{L}}$, $\hat{\mathbf{\Gamma}}$ and $\hat{\Delta}$ have to be chosen in such a way to solve the optimization problem above); $\|\hat{\mathbf{L}}\|_*$ is the nuclear norm of the matrix $\hat{\mathbf{L}}$. $\hat{\mathbf{\Gamma}} \mathbf{1}_{\boldsymbol{n}}^\top$ and $\mathbf{1}_{\boldsymbol{m}} \hat{\Delta}^\top$ model, respectively, row fixed effects (and can be interpreted as individual fixed effects) and column fixed effects

DRAFT

(time fixed effects) in the reconstruction $\hat{\mathbf{M}}$ of $\mathbf{M}$. Differently from MC optimization problem by (5), the nuclear norm $\|\hat{\mathbf{L}}\|_*$ is used instead of $\|\hat{\mathbf{M}}\|_*$ (i.e., the fixed effects $\hat{\mathbf{\Gamma}}\mathbf{1}_n^\top$ and $\mathbf{1}_m\hat{\Delta}^\top$ are not regularized). Herein after, we refer to this method as MC Fixed Effects, or simply `MCFE`. When row fixed effects $\hat{\mathbf{\Gamma}}\mathbf{1}_n^\top$ are constrained to zero, the model includes time fixed effects only. We refer to this case as MC Time Fixed Effects (`MCTFE`).

## 3. Application to $CO_2$ Emissions

Adopted data on $CO_2$ emissions at the national level (2) are freely available at https://op.europa.eu/en/publication-detail/-/publication/df9c194b-81ba-11e9-9f05-01aa75ed71a1/language-en for the period 2000 – 2016 and for 43 countries (30 from the European Union (EU) and 13 extra-EU). In this application, considered time period is from 2000 to 2005, in order to avoid possible treatment effects coming from the ETS. Moreover, we restrict the analysis to 26 countries (14 from the European Union (EU) and 12 extra-EU, which is the result of having dropped out all small countries and all extra-EU countries having special agreements with the EU).

Specifically, in this section we compare the performance of `MCB`, `MCTFE` and `MCFE`. The performance of the methods is assessed both with respect to the original matrix (`raw`) and to a suitably pre-processed matrix ($l_1$ row-normalization by country). This pre-processing is made with the aim of making comparable the orders of magnitude of all the elements (i.e., the countries) of that matrix. To avoid overfitting, for each row the computation of the $l_1$ norm takes into account only the elements in the training set and their number. According to the design of the experiments, for any specific percentage of unknown entries (from 0 to 50%, at intervals of 1), 200 replications have been generated, where the unknown entries (test set) are chosen at random according to the desired percentage.

Methods `MCTFE` and `MCFE` have been performed with `mcnnm_cv` function in `MCPanel` R package. The function automatically chooses the optimal value for $\lambda$ by dividing the observed entries into training and validation sets and by minimizing the average RMSE on the validation set[1]. Method `MCB` has been performed with the same function by constraining both time and individual fixed effects to zero.

Figure 1 displays the optimal $\lambda$ at increasing percentages of unknown entries (average over 200 replications). The increase in the optimal $\lambda$ is less pronounced if we include both individual and time effects (`MCFE`). We note that, on the $l_1$ row-normalized matrix, optimal $\lambda$ is always low with method `MCFE`, meaning that the penalization by the nuclear norm plays a minor role in cases when time and country fixed effects are taken into account.

To study and compare the performance of the methods, MAPE[2] has been evaluated on the test set as:

$$100 \times \frac{1}{|\Omega^{\text{test}}|} \sum_{i=1}^{|\Omega^{\text{test}}|} \frac{\mid y_i - \hat{y}_i \mid}{y_i},$$

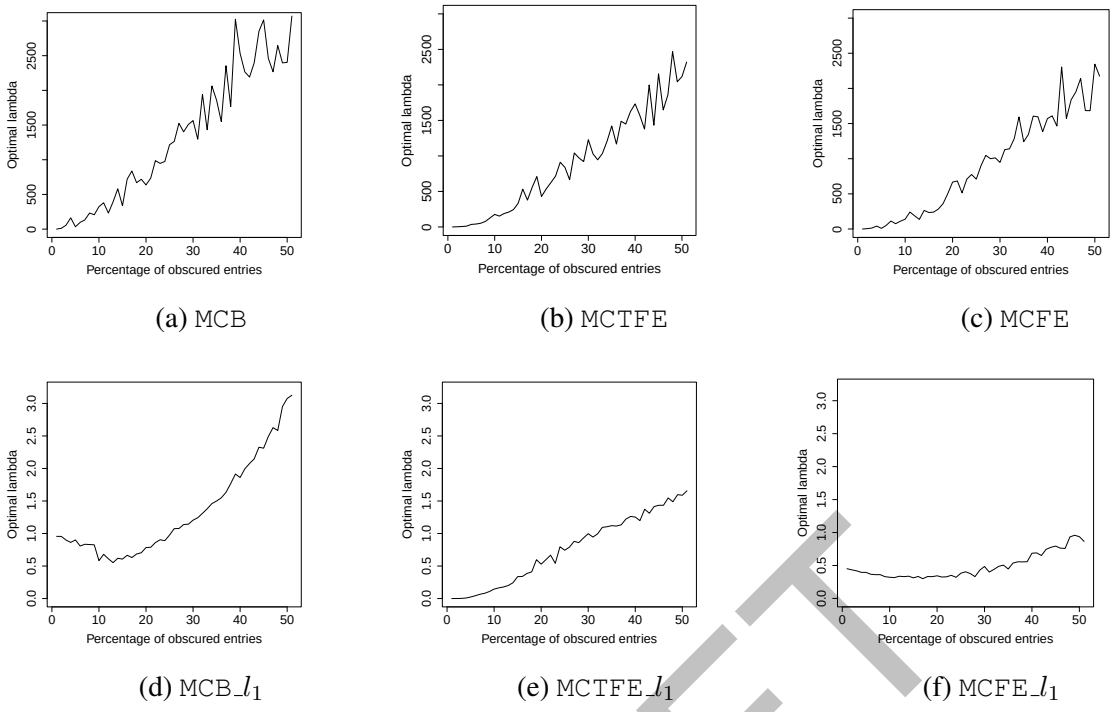
where $|\Omega^{\text{test}}|$ is the dimension of the test set, $y$ and $\hat{y}$ are, respectively, true and estimated values. According to Figure 2, MAPE is low, even for a large amount of unknown entries, on the $l_1$ row-normalized matrix. In general, method `MCFE` performs slightly better than method `MCB`.

The performance of these methods has been evaluated also with respect to their capability of preserving the original ratio among between and within country deviance. Decomposition of the total deviance in between and within country deviance has been evaluated on the reconstructed matrix according to the

[1] By considering the average RMSE, the comparison of RMSE on validation sets of different dimension is fair.

[2] MAPE in this case has to be preferred over RMSE in order to allow a comparison of the performance of the method on the original and on the normalized matrix.

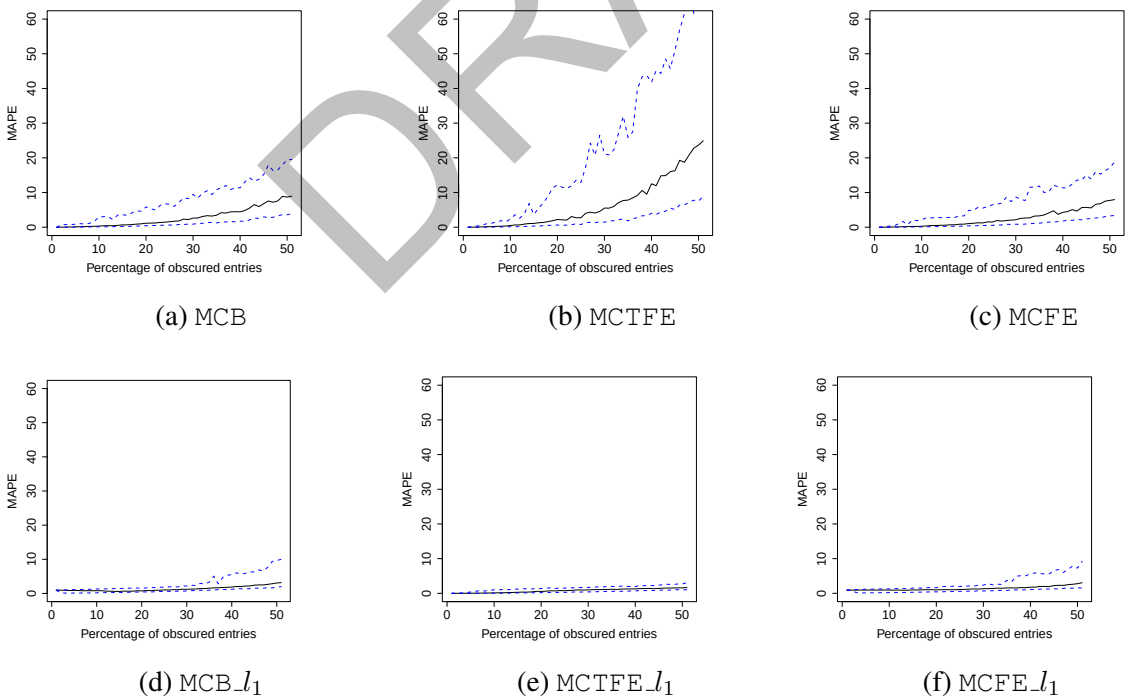
DRAFT




(a) MCB (b) MCTFE (c) MCFE

(d) MCB_$l_1$ (e) MCTFE_$l_1$ (f) MCFE_$l_1$

Figure 1: Optimal $\lambda$ at increasing percentages of unknown entries. Mean over the 200 replications. Top: raw matrix. Bottom: $l_1$ row-normalization by country.




(a) MCB (b) MCTFE (c) MCFE

(d) MCB_$l_1$ (e) MCTFE_$l_1$ (f) MCFE_$l_1$

Figure 2: MAPE at increasing percentages of unknown entries. Median over the 200 replications (solid lines), 95% confidence bands (blue dashed lines). Top: raw matrix. Bottom: $l_1$ row-normalization by country.

DRAFT

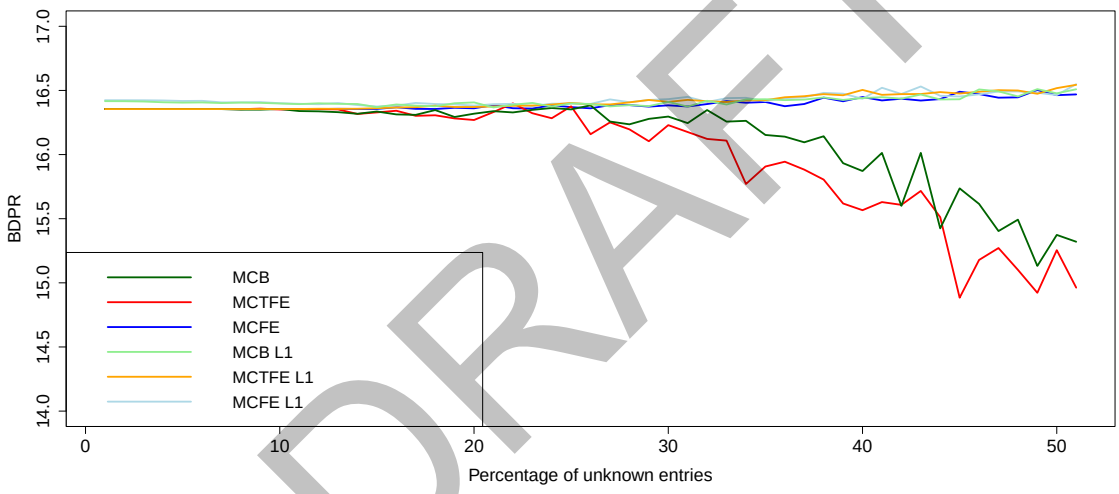
equation:

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{t}(\hat{y}_{ij}-\hat{y}_{..})^2}_{\text{Total deviance}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_{i.}-\hat{y}_{..})^2}_{\text{Between deviance}}+\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{t}(\hat{y}_{ij}-\hat{y}_{i.})^2}_{\text{Within deviance}},$$

where $n$ and $t$ stay, respectively, for the number of countries and the number of years, $\hat{y}_{ij}$ is the $CO_2$ emission for the country $i$ at time $t$, $\hat{y}_{..}$ is the matrix average, $\hat{y}_{i.}$ is the country $i$ average. The same equation holds for the original matrix when estimated values $\hat{y}$ are replaced with the original values $y$. The Between Deviance Percentage Ratio is evaluated as:

$$\text{BDPR}=\frac{\text{Between deviance}}{\text{Total deviance}}*100\%,$$

in order to understand whether such a ratio departs from that of the original matrix by increasing the amount of unknown entries. In the original matrix, the between deviance stands to 16.4% of the total deviance. By applying MC to the original (i.e., non normalized) matrix, the percentage of 16.4 is maintained for large amounts of unknown entries only using method `MCFE` (Figure 3). When considering the $l_1$ row-normalized matrix, all methods preserve the original BDPR.




DRAFT

Figure 3: Between deviance percentage ratio (BDPR) at increasing percentages of unknown entries. Median over the 200 replications.

## 4. Discussion and conclusions

To perform a robust counterfactual analysis on the emission of $CO_2$ after the Emission Trading System launched by the European Union in 2005, in this paper we assessed the performance of some nuclear norm regularized Matrix Completion (MC) methods in terms of Mean Absolute Percentage Error (MAPE) and in terms of decomposition in within and between country deviance. A study of the optimal $\lambda$ penalization parameter shows that, at increasing percentages of unknown entries, the increase in the optimal $\lambda$ is less pronounced for the MC method with both individual and time fixed effects. This means that by better modelling the error minimization component of the MC optimization problem, the component that depends on the nuclear norm decreases in relevance. This is particularly true when applying MC to the $l_1$ row-normalized matrix. MAPE generally increases by increasing the percentage of unknown entries. However, this increase is less pronounced when using the MC method with both fixed

effects instead of the one that constraint all effects to zero, or if we apply MC to the $l_1$ row-normalized matrix. The deviance decomposition of the original matrix is generally preserved for low amounts of unknown entries. By adopting the MC method with both fixed effects or by normalizing the rows of the matrix, the decomposition in within and between countries is preserved even for a larger amount of unknown entries. Overall, it can be concluded that the MC method without fixed effects presents good performance when applied to a row-normalized matrix. By using the MC method with both fixed effects, the performance is good even when applying the MC to the original matrix.

## References


[1] Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K.: Matrix completion methods for causal panel data models. Journal of the American Statistical Association **116(536)**, 1716-1730 (2021)

[2] Corsatea T.D., Lindner S., Arto, I., Roman, M.V., Rueda-Cantuche J.M., Velazquez Afonso A., Amores A.F., Neuwahl F.: World Input-Output Database Environmental Accounts. Update 2000-2016, EUR 29727 EN, Publications Office of the European Union, Luxembourg, (2019)

[3] Gnecco, G., Nutarelli, F., & Riccaboni, M.: A machine learning approach to economic complexity based on matrix completion. Scientific Reports, **12(1)**, 9639 (2022)

[4] Hastie T, Tibshirani R, Wainwright M, Statistical Learning with Sparsity: The Lasso and its Generalizations. CRC Press, New York (2015).

[5] Mazumder R, Hastie T, Tibshirani R,: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. Journal of Machine Learning Research **11**, 2287-2322 (2010)

[6] Metulini, R., Gnecco, G., Biancalani, F., & Riccaboni, M.: Hierarchical clustering and matrix completion for the reconstruction of world input-output tables. AStA Advances in Statistical Analysis, 1-46 (2022)

[7] Yoro, K. O., & Daramola, M. O., CO2 emission sources, greenhouse gases, and the global warming effect. In: Advances in carbon capture, pp. 3-28. Woodhead Publishing (2020).


DRAFT

# Deep Learning for smart and sustainable agriculture

Amalia Vanacore[a], Armando Ciardiello[a,b], Annalisa Izzo[b], Pierdomenico Zaffino[b], Carolina Vecchio[b], Gennaro Pio Auricchio[b] and Luigi Uccello[b]

[a] Dept. of Industrial Engineering, University of Naples Federico II, p.le Tecchio 80, Naples, Italy;
amalia.vanacore@unina.it,armando.ciardiello@unina.it
[b] Deloitte Consulting SRL SB, Via Tortona 25, Milan, Italy

**Abstract**

In recent years, machine learning and deep learning techniques have emerged as effective tools to address the challenges related to smart and sustainable agriculture thanks to their ability to automate the task of processing various types of data (*e.g.*, soil condition data and meteorological information) and transform this data into knowledge that will serve in the decision-making process. The aim of this paper is to showcase through a real case study and an exemplificative application how relevant issues related to the sustainable management of farming activities can be addressed in an effective way by employing deep learning algorithms. The real case study concerns the development of a vineyard yield prediction model using time series of satellite images whereas the exemplificative application showcases how deep learning techniques based on object detection can be usefully exploited to build a plant disease detection system.

**Keywords:** Smart and Sustainable Agriculture, Yield Prediction, Disease detection, Deep Learning

DRAFT

## 1. Introduction

Agriculture is a key component in the 2030 Agenda for Sustainable Development to achieve the Sustainable Development Goals (SDGs), since it is the common thread which holds all the SDGs together. The growing population and the complexity of climate change have forced the agriculture sector to data-driven management and automation to increase production while using fewer resources. To achieve the objectives of both the SDGs and the Farm to Fork Strategy, presented in the first quarter of 2020 [11], European farmers need to transform their production processes in a more sustainable way, making the best use of nature-based, technological, digital solutions to properly allocate resources (*e.g.*, water, pesticides, fertilizers), deliver better environmental results and make all productive systems resilient to climate change. Smart Agriculture (SA) supports a data-driven management system for sustainable farming by precisely determining the steps that need to be practiced at its due season through the adoption of modern information technologies, software tools, and smart embedded devices. In such scenario, Machine Learning (ML) and Deep Learning (DL) algorithms can support farmers in making faster decisions on crop management. A main task in crop management is to identify which features (*e.g.*, topography, temperature, soil type, soil quality, rainfall, soil moisture, pH, electrical conductivity, soil nutrients) affect crop yield and the specific requirements to monitor for different phases of crop phenology such as: seeding, growing and maturity, so as to build for example an accurate crop yield prediction model which can help farmers to identify the type of crop to grow and when to grow [25]. ML and DL algorithms can also support farmers to assess agriculture land suitability [21]; measure fruit maturity to identify the suitable harvest moment [5]; detect nutrient status for precision fertilizers management [4]; predict the irrigation schedule and the amount of water to be applied to reduce high water consumption by adapting supplies to the crop needs and avoiding losses [1]. DL algorithms are the most prominent choice for building disease identification systems [15]. They can offer accurate

detection of diseases and weeds to minimize the excessive use of fungicides and pesticides and the need for field scouts [2]; help classify types of plant diseases, since most disease infections are too subtle for the naked eye to be able to perceive [19]; pinpoint which weed infestation areas are most critical, thus allowing farmers to target actions based on environmental customization [3]. This paper aims at showcasing how the adoption of predictive analytics based on DL algorithms can support a knowledge-based crop management system. The paper is organized as follows: Section 2 is devoted to vineyard yield prediction modelling using time series of satellite images; Section 3 illustrates an exemplificative application of DL algorithms for plant disease detection; Conclusion are drawn in Section 4.

## 2. Vineyard yield prediction

Vineyards for wine production are complex farming systems where aspects related to crop performance and grape and wine compositions are worthy of consideration. In general, with a given climate, vineyard capacity (*i.e.*, grape yield per hectare) is regulated and limited to specific quantities to not exceed certain levels of berry sugar concentrations and dry matter accumulation. An early prediction of the final yield level can be used to establish, and later undertake, farming practices capable of obtaining desired levels of vineyard yield.

### 2.1 Methodologies

Conventional methods for vineyard yield prediction [14], [23] are based on manual sampling strategies for the estimation of the average number of productive vines (APV), the average number of clusters per productive vine (ACV) and the average cluster weight (ACW). Sampling can be performed on vineyard representative sites at various phenological stages, and yield potential can be estimated using historical data (e.g., ACW at harvest). However, the dependency on historical data regarding cluster weight at harvest is a disadvantage because of climate change environment. An alternative manual method consists of weighing a sample of clusters at the lag phase (just before veraison) while considering a growth factor (generally 2) until full maturation. The lag phase method is based on destructive sampling; its dependency on a berry growth factor, makes it sensitive to growth conditions, cultivar, and management practices. Both methods are labor intensive and time consuming, especially in highly variable vineyards where a higher number of samples is required for an accurate yield estimation; moreover, being dependent on sampling in the field and manual counting, these methods are highly sensitive to spatial variability.

Recently, the adoption of ML and DL techniques tries to overcome the limitations of traditional yield estimation methods. Two alternative approaches are dominant in the specialized literature: (1) combination of object detection algorithms and computer vision systems to build models that try to do an exhaustive enumeration of grape clusters to estimate final vineyard yield and (2) combination of predictive analytics methods and remote sensing techniques that exploit environmental and vegetative status data to predict final yield. In the following we focus on the latter approach since it has many advantages over the former: it is less intrusive in the daily operational context, moreover, remote sensing provides a higher coverage capability and significant low-cost for data acquisition. Long short-term memory (LSTM) neural networks have been successfully applied to crop yield estimation and prediction [22] as they are suitable for learning the relationship between time series data (*e.g.*, temperature and rainfall) and a continuous scalar variable (*e.g.*, final yield). LSTM is a special kind of recurrent neural network (RNN) developed by [9] to tackle the vanishing-gradient-problem that occurs with conventional RNN and leads to the inability to learn long-term dependencies. LSTM is developed to control the whole information flow within neurons, for this purpose, a gating mechanism controls the process of adding and deleting information from an iteratively propagated cell state. Thus, the process of forgetting can be controlled, and a defined memory behavior is realized to model short-term as well as long-term dependencies.

### 2.2 Data acquisition

The dataset was built collecting vineyard data and satellite time series of atmospheric and vegetation status data considering four different vintages (viz. 2018, 2019, 2020, 2021) of 264 vineyards located in Southern Italy. Vineyard data were obtained from wine Production Declarations of producers which contain records that allow the traceability of the vineyard plot location, grape variety, and final yield. Time series of temperature, precipitation, atmospheric pressure, and Leaf Area Index (LAI) were

DRAFT

obtained using ERA5-Land a database with a spatial resolution of 9 km and a temporal resolution of 1 day. Time series of vegetation indices [26] (i.e., Normal Difference Vegetation Index, NDVI; Modified Soil-Adjusted Vegetation Index, MSAVI; Normalized Difference Moisture Index, NDMI; Enhanced Vegetation Index, EVI) were calculated using a spectral imaging transformation of multispectral images with a spatial resolution of 10 m and a temporal resolution of 10 days obtained from Sentinel-2 a multispectral imaging mission covering 13 spectral bands that range from the visible range to the shortwave infrared (SWIR). Vegetation indices summarize relevant ground information to monitor vineyards, vine growth cycles, nutrient and water stress, thus they are widely employed for better vineyard management in precision agriculture applications [8].

Satellite time series data acquisition was articulated into four phases: obtaining geographical coordinates of vineyards using Google Earth Engine (GEE) code editor; extracting satellite image of the region of interest using GEE API in Python; cleaning satellite image via S2 cloudless cloud detection algorithm [20]; extracting time series of predictors by cutting out the polygon representing the region of vineyard and calculating for each satellite image: average, maximum, minimum of pixels for Temperature; average of pixels for Rainfall and Pressure; average, median, maximum, minimum and standard deviation of pixels for NDVI, EVI, NDMI and MSAVI; and average of pixels for LAI. Climatic and vegetation status time series data were used as predictors to develop the vineyard yield prediction model; each predictor consists of 1056 time series (related to the 264 vineyards observed from 2018 to 2021) extracted at 30 equally spaced time steps with a time span of 10 days. Extracting time series of predictors instead of directly using satellite images as input data stands as the optimal option considering two specificities of the application context: (1) the small number of involved vineyards and available vintages put a constrain on the dataset size and thus limit the effective training of sophisticated prediction models (*e.g.* [12]) able to process satellite images by accounting for both temporal and spatial dependencies; (2) the need to limit costs forced the choice of the satellite Era5 for the acquisition of climatic data; however, the data resolution (9 km grid spacing) allows to obtain a single pixel rather than a pixelated image of the vineyard since the area of this latter is smaller than the pixel size.

### 2.3 Model Architecture

The model architecture has been defined using optimized parameters obtained via the grid search hyperparameter optimization technique by keeping the value of six parameters as follows: number of layers specified as [2, 4, 6]; number of neurons in the hidden layers specified as [20, 50, 80, 100, 150]; learning rate specified as [0.001, 0.005, 0.01, 0.03, 0.05]; batch size specified as [16, 32, 64, 128]; activation functions in hidden layers specified as sigmoid, hyperbolic tangent (tanh) and rectified linear unit (ReLU); optimizers specified as Stochastic Gradient Descent (SGD), Root Mean Square Propagation (RMSProp) and ADAptive Moment estimation (ADAM).

During the training, early stopping with patience for stagnant progress of 50 epochs has been used. Each training iteration has been allowed to continue for a maximum of 500 epochs. Hyper-parameters tuning and model selection have been carried out using the 5-fold cross-validation. Significantly better results have been obtained with three hidden layers with 80 neurons, batch size of 16, learning rate of 0.01, nonlinear activation function ReLu and ADAM optimizer. The model has been implemented in Python programming language using Keras library, a high-level Neural Network API and the back-end processing TensorFlow library.

The proposed model has been compared to conventional sampling strategy in performing a prediction of the final yield approximately 30 days before harvest. The test sets were based on a sample of 30 vineyards (located in the same region) and performances have been assessed via Mean Absolute Percentage Error (MAPE) which results equal to 15,6 % and 21,6% for the LSTM strategy and the traditional one, respectively. However, besides higher predictive performance, the advantages of the strategy based on LSTM are evident when considering the reduction in product waste, operating costs and time needed to obtain yield estimation. There is still room for performance improvement of the proposed strategy via high resolution satellite images and/or additional key variables as predictors.

## 3. Grapevine Disease Detection

Plants suffer from different diseases caused by different agents like bacteria, fungi, viruses etc. depending on climate and environmental conditions. The rate of disease spread depends on current crop

DRAFT

conditions and its susceptibility to infection. When plants are affected by diseases, they show a range of symptoms such as colored spots, or streaks that can occur on the leaves, stems, and seeds of the plant. These visual symptoms continuously change in color, shape, and size as per disease progress.

### 3.1 Methodologies

Plant disease detection is traditionally performed by continuously monitoring crop conditions via visual observation (*i.e.*, direct approach). This strategy is ineffective since growers may not have adequate disease identification knowledge and on the other hand the availability of experts for continuous monitoring is not affordable. In the last decade, scientific literature has shown a substantial increase of indirect approaches for disease detection and classification [6]. Indirect approach uses image processing techniques to identify and classify disease symptoms for different crops. These techniques lead to faster disease detection and diagnosis and allow to determine the exact quantity of pesticides required for a particular disease. The first applications of indirect approach for disease detection adopted more traditional ML algorithms such as support vector machines (SVM) and K-means clustering, however, these methods have shown low detection efficiency because of the complexity of image preprocessing and feature extraction. In recent years, disease identification strategies have been developed using DL algorithms mostly based on Convolutional Neural Networks (CNN), the most frequently used are AlexNet, GoogLeNet, VGG-16, and ResNet-50. These techniques suffer some limitation and drawbacks when used as automatic disease recognition system in field conditions, for this reason DL-based object detection algorithms (*e.g.*, Region-Based Convolutional Neural Network (R-CNN), Fast R-CNN, Faster R-CNN, Yolo and SSD-MobilNet) have been explored as a solution for successful plant disease detection in realistic conditions characterized by images with complex backgrounds or with multiple leaves (diseased or healthy) from different plant species along with different classes.

### 3.2 Dataset

A total number of 365 images of diseased grapevines was taken from the PlantVillage dataset [10] and EdenLibrary Online Repository [17]. All these images were collected under uncontrolled conditions and without focusing on a particular organ (*e.g.*, clusters or leaves). Table 1 reports for each disease the number of images divided between train and test sets.

Table 1: Number of images per disease

| Disease | Train | Test |
|---|---|---|
| Esca | 108 | 48 |
| Downy Mildew (DW) | 35 | 24 |
| Powdery Mildew (PW) | 66 | 22 |
| Eriophyes vitis (EV) | 45 | 17 |
| Total | 254 | 111 |

The total number of diseased leaves annotations amounts to 7475. To increase dimensionality, the training set was enlarged using augmentation techniques that do not negatively influence the detection of diseases: 90°-180°-270° image rotation; brightness variation; image magnification [13]. Finally, a total number of 1157 images and 23421 annotations was obtained.

### 3.3 Model Architecture

The architecture employed for developing the DL-based object detection model is YOLOv7 [24], one of the best available solutions in terms of speed and accuracy, characterized by high ability to detect small objects like leaf spot diseases (*e.g.*, [18], [7] and [16]).

In the training process, the pre-trained weights of the MS COCO dataset have been used as the initial weights and imported into YOLO architecture.

To generate a robust YOLO model for disease detection, hyper-parameters, such as optimizer (SGD, ADAM), learning rate (0.01, 0.001, 0.0001, 0.00001, 0.0000001) and momentum (0.5, 0.9, 0.99) have been tuned as they can significantly reduce training time and improve the model performance. During training, early stopping with patience for stagnant progress of 50 epochs has been used. The training and test evaluation of each trained model has been performed with hold-out sets. The following

DRAFT

hyperparameters have been selected after tuning: Batch size=16; Image size=1280; Optimizer= ADAM; Momentum=0.937; Learning rate=0.01; Epochs=500.

The model has been implemented in Python using PyTorch framework. For testing purpose, 132 images of healthy grapevines collected from EdenLibrary Repository have been considered.

Table 2: Confusion matrix and performance results

| | | Predicted | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Esca | DW | PW | EV | Healthy | Precision | Sensitivity | F1 score |
| Actual | Esca | 48 | 0 | 0 | 0 | 0 | 100.00% | 100.00% | 100.00% |
| | DW | 0 | 22 | 0 | 0 | 2 | 100.00% | 91.67% | 95.65% |
| | PW | 0 | 0 | 9 | 0 | 13 | 52.41% | 40.90% | 45.95% |
| | EV | 0 | 0 | 0 | 14 | 3 | 70.00% | 82.35% | 75.67% |
| | Healthy | 0 | 0 | 8 | 6 | 118 | 86.13% | 89.39% | 87.73% |

The results reported in Table 2 show good performance for detection of Esca, DW and EV diseases. The model has difficulties in distinguishing PW from Healthy leaves, in fact PW only partially reduces the leaf aesthetic, unlike the other diseases, which considerably change the color of leaves.

From an operational point of view, it is not important that the model detects all the diseased leaves, but it is sufficient that the model identifies at least one diseased leaf to produce an alert; thus, to assess model performance all bounding boxes related to a diseased leaf have been labeled into a single class. Results in Table 3 show a satisfactory model performance.

Table 3: Performance of grapevine disease detection system

| | |
|---|---|
| Accuracy | 87.00% |
| Precision | 88.00% |
| Sensitivity | 84.00% |

## 4. Conclusion

The use of DL techniques in SA offers a high level of accuracy and outperforms existing traditional ML algorithms, providing flexible tools that can be deployed as an aid for the sustainable management of farming activities. The proposed vineyard yield prediction model based on satellite remote sensed meteorological and vegetation data shows better performance results with respect to conventional strategy based on destructive sampling, and, in addition, it has a positive impact on sustainability in terms of reduction of product waste, costs, labor and time needed for the yield estimation. The grapevine disease detection strategy described in the exemplificative application can be used in real field conditions where images are characterized by complex backgrounds, this is the case for example, of images captured by a drone flying over crops. The integration with Unmanned Aerial Vehicles (UAV) will make possible to automate the detection of diseases in the vineyard in real time to take necessary countermeasures.

**Acknowledgments** The research has been funded by project SCAE – Smart and Connected Agrifood Ecosystem – F/200119/01/X45.

### References

[1] Ahansal, Y., Bouziani, M., Yaagoubi, R., Sebari, I., Sebari, K., & Kenny, L.: Towards smart irrigation: A literature review on the use of geospatial technologies and machine learning in the management of water resources in arboriculture. Agronomy, 12(2), 297 (2022).

[2] Ahmad, A., Saraswat, D., & El Gamal, A.: A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. Smart Agric. Technol., 100083 (2022).

[3] Bah, M. D., Hafiane, A., & Canals, R.: Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. Remote sens., 10(11), 1690 (2018).

[4] Durai, S. K. S., & Shamili, M. D.: Smart farming using machine learning and deep learning techniques. Decis. Anal. J., 3, 100041 (2022).

DRAFT

[5] El-Bendary, N., El Hariri, E., Hassanien, A. E., & Badr, A.: Using machine learning techniques for evaluating tomato ripeness. Expert Syst. Appl., 42(4), 1892-1905 (2015).
[6] Ferentinos, K. P.: Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric., 145, 311-318 (2018).
[7] Giakoumoglou, N., Pechlivani, E. M., Sakelliou, A., Klaridopoulos, C., Frangakis, N., & Tzovaras, D.: Deep Learning-Based Multi-Spectral Identification of Botrytis cinerea. Smart Agric. Technol., 100174 (2023).
[8] Giovos, R., Tassopoulos, D., Kalivas, D., Lougkos, N., & Priovolou, A.: Remote sensing vegetation indices in viticulture: A critical review. Agriculture, 11(5), 457, (2021).
[9] Hochreiter, S., & Schmidhuber, J.: Long short-term memory. Neural Comput., 9(8), 1735-1780 (1997).
[10] Hughes D, Marcel S.: An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint, arXiv:1511.08060 (2015).
[11] Hurduzeu, G., Pânzaru, R. L., Medelete, D. M., Ciobanu, A., & Enea, C.: The Development of Sustainable Agriculture in EU Countries and the Potential Achievement of Sustainable Development Goals Specific Targets. (SDG 2). Sustainability, 14(23), 15798 (2022).
[12] Khaki, S., Wang, L., & Archontoulis, S. V.: A cnn-rnn framework for crop yield prediction. Front. Plant Sci., 10, 1750 (2020)
[13] Kobayashi, K., Tsuji, J., & Noto, M.: Evaluation of data augmentation for image-based plant-disease detection. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2206-221 (2018).
[14] Laurent, C., Oger, B., Taylor, J. A., Scholasch, T., Metay, A., & Tisseyre, B.: A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture. Eur. J. Agron., 130, 126339 (2021).
[15] Liu, J., & Wang, X.: Plant diseases and pests detection based on deep learning: a review. Plant Methods, 17, 1-18, (2021).
[16] Morbekar, A., Parihar, A., & Jadhav, R.: Crop disease detection using YOLO. In: 2020 international conference for emerging technology (INCET), pp. 1-5 (2020).
[17] Mylonas, N., Malounas, I., Mouseti, S., Vali, E., Espejo-Garcia, B., & Fountas, S.: Eden library: A long-term database for storing agricultural multi-sensor datasets from uav and proximal platform. Smart Agric. Technol., 2, 100028 (2022).
[18] Nayar, P., Chhibber, S., & Dubey, A. K: An Efficient Algorithm for Plant Disease Detection Using Deep Convolutional Networks. In: 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 156-160 (2022).
[19] Saleem, M. H., Potgieter, J., & Arif, K. M.: Plant disease detection and classification by deep learning. Plants, 8(11), 468 (2019).
[20] Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., ... & Žust, L.: Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. Remote Sens. Environ, 274, 112990 (2022).
[21] Taghizadeh-Mehrjardi, R., Nabiollahi, K., Rasoli, L., Kerry, R., & Scholten, T.: Land suitability assessment and agricultural production sustainability using machine learning models. Agronomy, 10(4), 573 (2020).
[22] Van Klompenburg, T., Kassahun, A., & Catal, C.: Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric., 177, 105709 (2020).
[23] Victorino, G., Braga, R. P., Santos-Victor, J., & Lopes, C. M.: Comparing a New Non-Invasive Vineyard Yield Estimation Approach Based on Image Analysis with Manual Sample-Based Methods. Agronomy, 12(6), 1464 (2022).
[24] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint, arXiv:2207.02696 (2022).
[25] Wang, J., Si, H., Gao, Z., & Shi, L.: Winter wheat yield prediction using an LSTM model from MODIS LAI products. Agriculture, 12(10), 1707 (2022).
[26] Xue, J., & Su, B.: Significant remote sensing vegetation indices: A review of developments and applications. J. Sens., 1353691 (2017).

DRAFT

# Do green transition, environmental regulations and renewable energy promote ecological sustainability in G7 countries? Evidence from panel quantile regression

Aamir Javed[a], Agnese Rapposelli[b], Asif Javed[c]

[a] Department of Management and Business Administration; aamir.javed@studenti.unich.it
[b] Department of Economic Studies; agnese.rapposelli@unich.it
[c] Department of Advanced Technologies in Medicine & Dentistry; asif.javed@studenti.unich.it

DRAFT

**Abstract**

This study investigates the impact of green technology innovation, environmental taxes, renewable energy consumption, along with economic growth, trade openness and urban population, on the ecological sustainability of the G7 countries within the framework of Environmental Kuznets Curve (EKC) hypothesis. To this end, we utilized a combination of second-generation methodologies including quantile regression (QR), augmented mean group (AMG), fully modified ordinary least square (FMOLS), dynamic ordinary least square (DOLS) and fixed effect ordinary least square (FE-OLS). The empirical findings show that green technological innovation, environmental taxes and urban population reduce ecological footprint and enhance environmental sustainability, whereas renewable energy consumption shows a negative and significant relationship with ecological footprint results also confirm the existence of EKC hypothesis in G7 economies. These results indicate that policy makers should encourage investments in green innovation and renewable energy industries.

**Keywords:** Green technology innovation, Environmental taxes, Renewable energy consumption, Ecological sustainability, Panel quantile regression.

## 1. Introduction

Many countries are working for mitigating the adverse effects of climate change, land degradation, desertification and other associated human and environmental degradation. After the increasing awareness of the dangerous effects of greenhouse gases (GHGs) on humans and ecosystems, the introduction of carbon tax, innovations and energy-saving and efficient technologies are some of the specific strategies targeted at reducing carbon dioxide ($CO_2$) emissions and other GHGs [20]. Despite the efforts done, the Intergovernmental Panel on Climate Change reports an increase in $CO_2$ emissions from 9434.4 to 34,649.4 million tons between 1961 and 2011 [12]. The Energy International Agency also reports that $CO_2$ emissions increased from 29,714.2 to 33,444 million tons between 1999 and 2017. The Paris Agreement, signed in 2015, reveals that the world's $CO_2$ emissions grew by 1.3% between 2006 and 2016, and 1.6% in 2017, thus constituting about 81% of greenhouse gases.

Environmental pollution has a multi-dimensional effect on the ecological system; hence, the proxy used for environmental quality has remained mixed. $CO_2$ emissions are extensively investigated in the literature and remain the core of international climate change agreements; besides, there are some other factors, such as deterioration in the quality of soil, forest and water, which are facing ecological threats [2]. These factors are of great importance and represent an integral part of the ecosystem. Hence, an ecological footprint indicator based on the concept of carrying the capacity of ecosystem represents an important issue in the

ecological system [2]. Moreover, green innovation provides benefits to consumers and organizations and helps reducing the severe environmental effects [18]. It comprises various techniques, including waste recycling, energy-saving, pollution prevention and environmental management [34]. Green technologies are capturing the continuous attention of researchers, resulting from the growing concern of people regarding environmental conditions [18]. The consumption of renewable energy sources is also one of the major determinants in breaking the strong relationship between fuel pollution, $CO_2$ emissions, and the economy growth. Energy utilization is directly linked with economic growth, which indicates that energy consumption is also responsible for environmental quality [23, 25]. The conventional sources of renewable energy, including hydropower and biomass, reduce a substantial level of $CO_2$ emissions while contributing to approximately 17% of the overall world demand for energy. Further, environmental taxes represent the most effective policy tool for reducing GHGs [3]. It aims at the taxation of carbon emissions to increase energy efficiency, reduce environmental problems and contribute to environment protection, by internalizing negative externalities in the form of environmental pollution ([21], [4]).

Based on the above discussion, the aim of this study is to investigate the impact of green innovation, environmental taxes and renewable energy consumption, along with economic growth, trade openness and urban population, on the ecological footprint of the G7 economies for the period 1994-2018, in presence of Environmental Kuznets Curve (EKC) hypothesis. This research contributes to the literature in several ways. Firstly, the analysis is focused on the G7 economies. Since the USA is the second-largest emitter of GHGs in the world and accounts for around 13% of global emissions, pollutant emissions from the G7 countries are still high [30]. Moreover, as of 2017, 9% of total GHGs emissions come from the European Union [30], with G7 members including Germany, France, Italy and UK playing a significant role. Therefore, there is still room for discussion on the issue of reducing environmental harms in the G7 countries. Secondly, we use the ecological footprint, which is a more comprehensive indicator of environmental accounting, as a proxy for environmental quality, whereas most of the existing studies have used $CO_2$ emission. Lastly, the impact of green innovation, environmental taxes and renewable energy consumption on ecological footprint is analysed by utilizing a technique - quantile regression - that considers the conditional distribution of ecological footprint. To our best knowledge, no study has employed the quantile regression approach with fixed effect in the ecological footprint framework.

The paper is organized as follows. Section 2 reviews the relevant literature, Section 3 introduces the methodology used and presents the data used. Section 4 presents the results obtained and Section 5 concludes.

DRAFT

## 2. Literature review

Many researchers have examined the link between environmental damage and technological advancement. For instance, Chu (2022) has investigated the relationship between green technological innovation and ecological footprint for OECD countries. His findings suggest that promoting the purchase of eco-friendly technologies can help reducing the ecological footprint. Further, Feng et al. (2022) for China, Liu et al. (2022) for ten Asian economies, Sherif et al. (2022) for N-11 countries, Shan et al. (2021) for Turkey, and Ke et al. (2022) for China have concluded that green technology enhances the quality of ecological footprint.

The attainment of environmental sustainability within an economy is predicated on the enforcement of environmental regulations. Thus, several previous studies have explored the dynamic impact of environmental regulations on environmental sustainability. For instance, Dogan et al. (2022) found that environmental tax reduces carbon emissions and increase environmental sustainability for G7 countries. Similarly, other studies found that environmental tax hinders the environmental degradation ([10], [27], [28], [29]). Moreover, research on the connection between renewable energy consumption and the environment has exploded in recent years. Sahoo and Sethi (2021) examined the dynamic impact of renewable energy consumption on ecological footprint for developing countries and found that clean energy increases ecological footprint efficiency. Recent studies obtained similar findings for other countries ([1], [8], [14], [31], [32], [33]).

## 3. Data and method

This study aims at examining the impact of green technology innovation, environmental taxes, renewable energy consumption, economic growth, trade openness and urban population on the ecological footprint-growth-environmental nexus of G7 countries for the period 1994-2018. A detailed description of the above

variables is given in Table 1.

Table 1: Description of variables

| Variables | Description | Sources |
|---|---|---|
| EFP | Ecological footprint (global hectares per person) | GFN |
| GTI | Green technology innovation (No. of patents related to environmental technologies on total patents - %) | OECD |
| ET | Environmental taxes (% of GDP) | OECD |
| REN | Renewable energy consumption (% of total final energy consumption) | WDI |
| GDP | Gross Domestic Product (constant 2015 US$) per capita | WDI |
| TO | Ratio of exports plus imports over GDP (%) | WDI |
| UP | Urban population (% of total population) | WDI |

All the variables are expressed in their logarithmic forms to avoid the issue of heteroscedasticity and to ensure the robust vector findings. The empirical model we use in this study is specified as follows:

$$lnEFP_{it} = \beta_o + \beta_1 lnGTI_{i,t} + \beta_2 lnET_{i,t} + \beta_3 lnREN_{i,t} + \beta_4 lnGDP_{i,t} + \beta_5 lnGDP_{it}^2 + \beta_6 lnTO_{i,t} + \beta_7 lnUP_{i,t} + \varepsilon_{it} \quad (1)$$

where EFP, GTI, ET, REN, GDP, GDP$^2$, TO and UP denote the ecological footprint, green technology innovation, environmental taxes, renewable energy consumption, per capita GDP, the square term of per capita GDP, trade openness and urban population, respectively. To examine the impact of the explanatory variables on the ecological footprint in their normal log equation at the selected quantile levels we performed the Panel Quantile regression (PQR) approach. This approach provides the heterogeneous results as compared to the long-run mean estimates such as (FMOLS, DOLS, and FE-LS) which only provides the mean estimators. However, in the real world the data does not fulfil the normality assumption. So, these long-run mean estimates may produce the inconsistent outcomes. The following is the equation of PQR:

$$Q\tau(lnEFP) = \vartheta\tau + \delta_{1\tau} lnGTI_{i,t} + \delta_{2\tau} lnET_{i,t} + \delta_{3\tau} lnREN_{i,t} + \delta_{4\tau} lnGDP_{i,t} + \delta_{5\tau} lnGDP_{it}^2 + \delta_{6\tau} lnTO_{i,t} + \delta_{7\tau} lnUP_{i,t} + \varepsilon_{it} \quad (2)$$

The locus idea $\tau$ is based on the explicative factors. $Q\tau$ parallels to the $\tau$th distributional stage regression evaluation which can be obtained by applying the method in Eq. 3:

$$Q_\tau = \arg\min \sum_{k=1}^{q} \sum_{t=1}^{T} \sum_{i=1}^{N} (|y_{it} - \alpha_i - x'_{it} Q_\tau| w_{it}) \quad (3)$$

where q, T, N and $w_{it}$ denote the number of quantiles, years, cross sections and weight of the i-th country in the i-th year, respectively.

## 4. Empirical results

Table 2 lists the results obtained by testing for slope heterogeneity. The findings indicate that both the parameters $\tilde{\Delta}$ and $\tilde{\Delta}$ adjusted are significant, thus providing evidence that slope coefficients are heterogenous in the considered panel. Table 3 presents the results of the cross-sectional dependence analysis obtained by means of Breusch and Pagan LM test [5], Pesaran scaled LM test [17] and Pesaran CD test [16]. They provide evidence against the null hypothesis of no cross-sectional dependence. We utilize the 2nd-generation unit root test, which is more powerful in dealing with slope heterogeneity and panel cross-section dependence. The results show that GTI is significant at level, while the other variables become stationary at the first difference (Table 4). Hence, the stationarity of all variables allows us to examine the long-run elasticities of the variables under analysis. Further, the results of the cointegration test, listed in Table 5, led us to offer the most appropriate assessment techniques, such as quantile regression (QR), augmented mean group (AMG), fully modified ordinary least square (FMOLS), dynamic ordinary least square (DOLS) and fixed effect ordinary least square (FE-OLS).

DRAFT

Table 2: Slope heterogeneity test results

| Test | Statistic | Prob. |
|---|---|---|
| $\tilde{\Delta}$ | 6.811*** | 0.000 |
| $\tilde{\Delta}$ adjusted | 8.514*** | 0.000 |

Note: *** indicates significance at 1% level.

Table 3: Cross-sectional dependence test results

| Test | Statistic | Prob. |
|---|---|---|
| Breusch-Pagan LM | 110.581*** | 0.000 |
| Pesaran scaled LM | 13.823*** | 0.000 |
| Pesaran CD | 6.227*** | 0.000 |

Note: *** indicates significance at 1% level.

Table 4: Unit root test results (Pesaran, 2007)

| | CIPS | | CADF | |
|---|---|---|---|---|
| Variables | I(0) | I(1) | I(0) | I(1) |
| LNEFP | -2.792 | -3.591*** | -2.586 | -3.686*** |
| LNGTI | -3.303*** | - | -3.503*** | - |
| LNET | -1.594 | -2.853** | -1.090 | -2.494** |
| LNREN | -2.489 | -4.569*** | -1.834 | -3.946*** |
| LNGDP | -2.229 | -2.352** | -2.007 | -2.352*** |
| LNTO | -1.1947 | -3.407*** | -1.882 | -2.969*** |
| LNUP | -2.34 | -3.196*** | -0.734 | -1.509*** |

Note: *** and ** indicate significance at 1% and 5% level, respectively.

Table 5: Cointegration test results (Westerlund, 2007)

| Statistic | Value | Z-value | Robust P-value |
|---|---|---|---|
| Gt | -3.945*** | -3.054 | 0.000 |
| Ga | -4.691* | 3.681 | 0.080 |
| Pt | -9.302*** | -2.309 | 0.000 |
| Pa | -5.044* | 2.432 | 0.072 |

Note: *** and * indicate significance at 1% and 10% level, respectively.

Table 6 presents the results obtained by applying AMG, FMOLS, DOLS and FE-OLS methods to examine the long-run impact of the considered regressors on ecological footprint. For all the estimation techniques applied, the results confirm the existence of an inverted U-shape curve for the G7 countries (the relationship between economic growth and pollution is positive, while the relationship between the square term of per capita GDP and pollution is negative). Moreover, regarding the trade openness, we observe a positive significant relationship with environmental degradation. On the contrary, green technology innovation, environmental taxes and renewable energy show a negative relationship with pollution. The results obtained by applying the quantile regression method are shown in Table 7. They indicate that green technology innovation and environmental taxes reduce the level of ecological footprint for all quantiles, suggesting that any increase in these variables promotes the ecological sustainability in G7 economies. These results are in line with previous literature ([9], [11], [15]. Regarding renewable energy consumption, even if the relationship is not significant at the early stages (Q10 and Q30), it negatively affects the environmental degradation at the later stages. This implies that the increase of the use of renewable energy would mitigate the environmental deterioration in the G7 countries. These findings are also in line with the results of previous research ([1], [33]). In addition, the results show that economic growth positively influences the pollution for all quantiles. This imply that higher income in the G7 economies is an important factor of the regions' economic degeneration. Besides, the square term of per capita GDP shows a negative and significant relationship with the ecological footprint across all quantiles. This indicates an inverse U-shaped relationship

DRAFT

between these variables and confirms also in this case the EKC hypothesis, as suggested by other studies ([6], [26]), given that the coefficients of GDP and $GDP^2$ are statistically significant with positive and negative signs, respectively. Finally, trade openness shows a positive and significant effect on the ecological footprint (except for Q10), thus increasing environmental degradation in the G7 economies.

Table 6: AMG, FMOLS, DOLS and FE-OLS estimates

| Variables | AMG | FMOLS | DOLS | FE-OLS |
|---|---|---|---|---|
| LNGTI | -0.056*** | -0.049*** | -0.067*** | -0.322*** |
| LNET | 0.338** | -0.196*** | -0.254*** | -0.768*** |
| LNREN | -0.036* | -0.087** | -0.087** | -0.018*** |
| LNGDP | 33.339*** | 22.532** | 20.555*** | 21.937*** |
| LNGDP2 | -1.639** | -1.065*** | -0.976*** | -1.035*** |
| LNTO | 0.049* | 0.077*** | 0.078*** | 0.188*** |
| LNUP | -6.516** | -1.589*** | -1.435** | -1.622*** |

Note: ***, ** and * indicate significance at 1%, 5% and 10% level, respectively.

Table 7: Panel quantile regression results

| Dependent variable: LEFP | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variables | Quantile | | | | | | | | |
| | Q10 | Q20 | Q30 | Q40 | Q50 | Q60 | Q70 | Q80 | Q90 |
| LNGTI | -0.345*** | -0.310*** | -0.316*** | -0.284*** | -0.301*** | -0.274*** | -0.277*** | -0.254*** | -0.280*** |
| LNET | -0.652*** | -0.748*** | -0.746*** | -0.729*** | -0.714*** | -0.765*** | -0.797*** | -0.861*** | -0.878*** |
| LNREN | 0.009 | -0.018*** | -0.013 | -0.010*** | -0.021** | -0.019*** | -0.020*** | -0.034** | -0.058*** |
| LNGDP | 29.988** | 28.124*** | 30.419*** | 32.068*** | 30.705*** | 25.206*** | 20.396*** | 18.131*** | 16.267*** |
| LNGDP2 | -1.404** | -1.325*** | -1.436*** | -1.511*** | -1.445*** | -1.187*** | -0.960*** | -0.858*** | -0.771*** |
| LNTO | 0.047 | 0.154*** | 0.151*** | 0.142*** | 0.137*** | 0.175*** | 0.210*** | 0.258*** | 0.302*** |
| LNUP | -1.818*** | -1.839 | -1.653*** | -1.655*** | -1.651*** | -1.651*** | -1.649*** | -1.746*** | 1.732*** |
| Obs. | 175 | 175 | 175 | 175 | 175 | 175 | 175 | 175 | 175 |

Note: *** and ** indicate significance at 1% and 5% level, respectively.

DRAFT

## 5. Conclusion

This study addresses the ecological sustainability in the G7 economies (Canada, France, Germany, Italy, Japan, United Kingdom, USA) by using a combination of second-generation empirical methodologies. Within the framework of EKC hypothesis, we examined the environmental impact of green technology innovation, environmental taxes and renewable energy consumption on the ecological footprint for the period 1994-2018. The results reveal that green innovation, environmental taxes, renewable energy consumption and urban population are the most impacting factors of ecological sustainability within the G7 countries. Moreover, both income levels and trade openness significantly harm the environmental quality.

The empirical findings provide some important policy implications for the ecological sustainability in the G7 countries. First of all, they indicate that policy makers should encourage investments in green innovation and renewable energy industries. G7 authorities must plan more carefully the implementation of fiscal incentives for green energy investment while putting more effort into ensuring favourable conditions for clean energy investors. These actions will significantly boost the actualization of crucial goals for the development of renewable energy usage and targets for a sustainable environment via a lower carbon emission for G7 economies. This will not only motivate new investments in clean energy, but it will also encourage existing investors in non-renewable energy sources, such as coal and fossil fuel sources, to gradually shift their attention toward green energy investments.

## References

[1] Adekoya, O.B., Oliyide, J.A., Fasanya, I.O.: Renewable and non-renewable energy consumption–Ecological footprint nexus in net-oil exporting and net-oil importing countries: Policy implications for a sustainable environment. Renew. Energy 189, 524--534 (2022)

[2] Ansari, M.A., Khan, N.A.: Decomposing the trade-environment nexus for high income, upper and lower middle income countries: What do the composition, scale, and technique effect indicate? Ecol. Indic. 121, 107122 (2021)
[3] Bashir, M.F., Ma, B., Shahbaz, M., Jiao, Z.: The nexus between environmental tax and carbon emissions with the roles of environmental technology and financial development. Plos One, 15(11), 1--20 (2020)
[4] Bashir, M.F., Ma, B., Komal, B., Bashir, M.A.: Analysis of environmental taxes publications: a bibliometric and systematic literature review. Environ. Sci. Pollut. Res. 28, 20700--20716 (2021)
[5] Breusch, T.S., Pagan, A.R.: The Lagrange multiplier test and its applications to model specification in econometrics. Rev. Econ. Stud. 47(1), 239--253 (1980)
[6] Chen, Y., Wang, Z., Zhong, Z.: $CO_2$ emissions, economic growth, renewable and non-renewable energy production and foreign trade in China. Renew. Energy 131, 208--216 (2019)
[7] Chu, L.K.: The impact of informal economy on technological innovation – ecological footprint nexus in OECD countries: new evidence from panel quantile regression. J. Environ. Stud. Sci., 12, 515--533 (2022)
[8] Destek, M.A., Sinha, A.: Renewable, non-renewable energy consumption, economic growth, trade openness and ecological footprint: Evidence from organisation for economic Co-operation and development countries. J. Clean. Prod. 242, 118537 (2020)
[9] Doğan, B., Chu, L.K., Ghosh, S., Truong, H.H.D., Balsalobre-Lorente, D.: How environmental taxes and carbon emissions are related in the G7 economies? Renew. Energy 187, 645--656 (2022)
[10] Fang, G., Yang, K., Chen, G., Tian, L.: Environmental protection tax superseded pollution fees, does China effectively abate ecological footprints? J. Clean. Prod. 135846 (2023)
[11] Feng, S., Chong, Y., Yu, H., Ye, X., Li, G.: Digital financial development and ecological footprint: Evidence from green-biased technology innovation and environmental inclusion. J. Clean. Prod. 380, 135069 (2022)
[12] IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland (2014)
[13] Ke, H., Dai, S., Yu, H. Effect of green innovation efficiency on ecological footprint in 283 Chinese Cities from 2008 to 2018. Environ. Dev. Sustain. 24(2), 2841--2860 (2022)
[14] Li, R., Wang, X., Wang, Q.: Does renewable energy reduce ecological footprint at the expense of economic growth? An empirical analysis of 120 countries. J. Clean. Prod. 346, 131207 (2022).
[15] Liu, C., Ni, C., Sharma, P., Jain, V., Chawla, C., Shabbir, M.S., Tabash, M.I.: Does green environmental innovation really matter for carbon-free economy? Nexus among green technological innovation, green international trade, and green power generation. Environ. Sci. Pollut. Res., 29(45), 67504--67512 (2022)
[16] Pesaran, M.H.: A simple panel unit root test in the presence of cross-section dependence. J. Appl. Econom. 22(2), 265--312 (2007)
[17] Pesaran, M.H.: Testing weak cross-sectional dependence in large panels. Econom. Rev. 34(6-10), 1089--1117 (2015)
[18] Razzaq, A., Fatima, T., Murshed, M.: Asymmetric effects of tourism development and green innovation on economic growth and carbon emissions in Top 10 GDP Countries. J. Environ. Plan. Manag. 66(3), 471--500 (2023)
[19] Sahoo, M., Sethi, N.: The intermittent effects of renewable energy on ecological footprint: evidence from developing countries. Environ. Sci. Pollut. Res. 28(40), 56401--56417 (2021)
[20] Saint Akadiri, S., Alola, A.A., Olasehinde-Williams, G., Etokakpan, M.U.: The role of electricity consumption, globalization and economic growth in carbon dioxide emissions and its implications for environmental sustainability targets. Sci. Total. Environ. 708, 134653 (2020)
[21] Shahzad, U.: Environmental taxes, energy consumption, and environmental quality: Theoretical survey with policy implications. Environ. Sci. Pollut. Res. 27(20), 24848--2486 (2020)
[22] Shan, S., Genç, S.Y., Kamran, H.W., Dinca, G.: Role of green technology innovation and renewable energy in carbon neutrality: A sustainable investigation from Turkey. J Environ. Manag. 294, 113004 (2021)
[23] Sharif, A., Mishra, S., Sinha, A., Jiao, Z., Shahbaz, M., Afshan, S.: The renewable energy consumption-environmental degradation nexus in Top-10 polluted countries: Fresh insights from quantile-on-quantile regression approach. Renew. Energy 150, 670--690 (2020)
[24] Sherif, M., Ibrahiem, D.M., El-Aasar, K.M.: Investigating the potential role of innovation and clean energy in mitigating the ecological footprint in N11 countries. Environ. Sci. Pollut. Res. 1--19 (2022)
[25] Sinha, A., Shahbaz, M.: Estimation of environmental Kuznets curve for $CO_2$ emission: role of renewable energy generation in India. Renew. Energy 119, 703--711 (2018)
[26] Solarin, S.A., Al-Mulali, U., Musah, I., Ozturk, I.: Investigating the pollution haven hypothesis in Ghana: an empirical investigation. Energy 124, 706--719 (2017)
[27] Tao, R., Umar, M., Naseer, A., Razi, U.: The dynamic effect of eco-innovation and environmental taxes on carbon neutrality target in emerging seven (E7) economies. J Environ. Manag. 299, 113525 (2021)

DRAFT

[28] Telatar, O.M., Birinci, N.: The effects of environmental tax on Ecological Footprint and Carbon dioxide emissions: a nonlinear cointegration analysis on Turkey. Environ. Sci. Pollut. Res. 29(29), 44335--44347 (2022)
[29] Ullah, S., Luo, R., Adebayo, T.S., Kartal, M.T.: Dynamics between environmental taxes and ecological sustainability: Evidence from top-seven green economies by novel quantile approaches. Sustain. Dev. 4, 12-26 (2022)
[30] UNEP: Emissions Gap Report 2018. United Nations Environment Programme, Nairobi, Kenya (2018)
[31] Usman, M., Hammar, N.: Dynamic relationship between technological innovations, financial development, renewable energy, and ecological footprint: fresh insights based on the STIRPAT model for Asia Pacific Economic Cooperation countries. Environ. Sci. Pollut. Res. 28(12), 15519--15536 (2021)
[32] Usman, O., Akadiri, S.S., Adeshola, I.: Role of renewable energy and globalization on ecological footprint in the USA: implications for environmental sustainability. Environ. Sci. Pollut. Res. 27(24), 30681-30693 (2020)
[33] Zhang, Q., Shah, S.A.R., Yang, L.: Modeling the effect of disaggregated renewable energies on ecological footprint in E5 economies: Do economic growth and R&D matter? Appl. Energy 310, 118522 (2022)
[34] Zhehao, H., Gaoke, L., Zhenghui, L.: Loaning scale and government subsidy for promoting green innovation. Technolog. Forecast. Soc. Ch. 144, 148--156 (2019)

DRAFT

# Doubly Robust DID for National Parks evaluation: "just" environmental benefits, or socioeconomics impacts as well?

Riccardo D'Alberto[a], Francesco Pagliacci[b], and Matteo Zavalloni[c]

[a] Dept. of Statistical Sciences "P. Fortunati", Alma Mater Studiorum University of Bologna, Via Delle Belle Arti 41, 40126 Bologna (BO) – Italy; riccardo.dalberto@unibo.it

[b] Dept. of Land, Environment, Agriculture and Forestry, University of Padova, Viale dell'Università 16, 35020 Legnaro (PD) – Italy; francesco.pagliacci@unipd.it

[c] Dept. of Economics, Society, Politics, University of Urbino Carlo Bo, Via Saffi 42, 61029 Urbino (PU) – Italy. matteo.zavalloni@uniurb.it

**Abstract**

National Parks (NPs) and protected areas are supposed to preserve the environment and prevent the loss of biodiversity. However, having substantially incremented worldwide, they now also include many areas that are important for economic development. Also, the literature on the subject has expanded, but targeting mainly the environmental benefits. This work investigates both the environmental and socioeconomic impacts of the Italian NPs of the 90s, by applying at the municipality level a Doubly Robust Difference-In-Differences estimator combined with Propensity Score Matching. The results suggest a positive effect on the environment on both the municipalities in NPs and the neighbouring ones, both in the short (2001) and medium run (2011). There are also socioeconomic effects in terms of the increase of incoming work-commuters and the number of workers employed in the tourism sector establishments.

**Keywords:** difference-in-differences, protected areas, socioeconomic impacts, spillover effects

DRAFT

## 1. Introduction

Being the backbone of the conservation policy, protected areas (PAs) are supposed to preserve the environment and prevent the loss of biodiversity. To date, PAs have substantially incremented worldwide, being this incrementation supported by both scholars and institutions [28,29]. The literature on PAs has increased and expanded as well, not only by targeting the (expected) environmental benefits but, also, by looking for potential socioeconomic benefits that the establishment of, e.g., a National Park (NP) can bring to local communities. The state of the art in the subject presents several works focusing on the Global South [3,9,10,19,25], finding a win-win situation in terms of both poverty and negative environmental externalities alleviation. Concerning high-income countries, the focus was mainly on the US [5,11,30], while the European PAs were seldom considered [7,26], with a specific focus on tourism and management [18,20], or by adopting a merely qualitative approach [21].

This research aims at reducing the gap by analysing the Italian NPs of the 90s, assessing both their environmental and socioeconomic impacts. The novelty of the works is threefold. First, it applies an innovative methodology, i.e., the Doubly Robust Difference-In-Differences (DR DID). Second, it considers two different post-treatment periods (compared to the single post-treatment instant usually considered [5,9,10]). Third, the analysis (carried out at the municipality level) targets both the municipalities covered by the NPs and the "neighbouring" municipalities, i.e., those not covered by NPs, but contiguous to the covered municipalities, where potential spillover effects could have been generated.

## 2. Methods

When a NP is established, it is not possible to observe what would have happened in that territory if the area had not been protected. Hence, the assessment of the NP impact represents a challenge of the "observational studies" framework, where a counterfactual estimation approach is applied to mimic the experimental design. Indeed, a sample of "untreated" municipalities (not included in a NP) is selected

to form the "counterfactual group" that must be compared with the "treated" one (the municipalities in NPs). The selection is made according to the similarity between the untreated and treated municipalities in terms of the relevant characteristics that are observable [4,9,10]. These help in identifying the most similar observations to be "matched". This sub-sample of matched municipalities is then considered for the estimation of the average treatment effect on the treated (ATT) with the new DR DID estimator [24].

The similarity between treated and counterfactual municipalities is based on a set of covariates, i.e., all the relevant, observable characteristics that are 1) statistically significant in terms of treatment adoption, 2) observed *before* the treatment, or not directly associated with it, 3) predictive of the outcome(s) but not influencing the treatment status. The covariates selection must undergo the "ignorability condition" [22], or the assumption of "selection on observable" [12].

The matching procedure selected for the extraction of the sub-sample on which to estimate the ATT must guarantee the optimal balancing of the covariates and the highest level of reduction of the differences originally observed between the treated and untreated units [14,15]. The same rationale is applied to the analysis of the potential geographical spillover effects.

Let $t$ be the time ($t = 0$ is the baseline year; $t = 1$ is the follow-up year). Let $Y_{mt}$ be the outcome of interest for the $m$-th municipality at time $t$. Whereas the $m$-th municipality hosts a NP before time $t$ (i.e., it is treated), $P_{mt} = 1$. Otherwise, $P_{mt} = 0$. Easing the notation, $P_m = P_{m1}$. Following [23], let $Y_{mt}(0)$ be the outcome of municipality $m$ at time $t$ if it is not part of a NP, while $Y_{mt}(1)$ indicates the outcome of the same municipality if the NP has been established. The outcome will be $Y_{mt} = P_{\mathrm{m}} Y_{mt}(1) + (1 - P_m) Y_{mt}(0)$. Let $X_m$ be the observed set of covariates. Assume that:

**A1.** $\{Y_{m0}, Y_{m1}, P_m, X_m\}$, for $m = 1, \ldots, n$, observations are independent and identically distributed (i.i.d.). The parameter of interest is $\tau = E[Y_{m1}(1) - Y_{m1}(0)|P_m = 1]$. The ATT can be rewritten as $\tau = E[P_m = 1] - E[P_m = 1] = E[P_m = 1] - E[P_m = 1]$.

**A2.** $E[P_m = 1, X_m] = E[P_m = 0, X_m]$, meaning that the average conditional outcome of the municipalities within a NP and the municipalities out of a NP would have been the same if the NP were not established (the "parallel trend assumption").

**A3.** For some $\xi > 0$, $P(P_m = 1) > \xi$ and $P(P_m = 1|X_i) \leq 1 - \xi$, meaning that there is at least a small portion of the municipalities that are included in a NP and, for every value of the covariates, there is at least a small probability for the municipality not being part of a NP (the "overlap condition").

From these assumptions it follows $E[P_m = 1] = E[P_m = 1, X_m] + E[P_m = 0, X_m] - E[P_m = 0, X_m|P_m = 1] = E[P_m = 1] + E[P_m = 0, X_m] - E[P_m = 0, X_m|P_m = 1]$. For estimating the ATT, [1] proposed to use $\tau = \frac{1}{E[P]} E[\frac{P - ps(X)}{1 - ps(X)}(Y_1 - Y_0)]$, where $ps(X) = P(P = 1|X)$ (the propensity score). Consequently, the estimator for the ATT is $\hat{\tau}^{ps} = \frac{1}{E_n[P]} - E_n\left[\frac{P - \hat{\pi}(X)}{1 - \hat{\pi}(X)}(Y_1 - Y_0)\right]$, with $\hat{\pi}(x)$ being the estimator for the true but unknown $ps(x)$.

The DR DID estimator combines the consistency property of the ordinary DID model by [13] and the properties of the $\hat{\pi}(\bullet)$ estimator for $ps(x)$ proposed by [1]. The result is an estimator that shows robustness even if either the ordinary DID model or the model for the propensity score are misspecified. Consequently, let it be that $\Delta Y = Y_1 - Y_0$ and let it be that $\mu^{ps}_{p,\Delta}(X) = \mu^{ps}_{p,1}(X) - \mu^{ps}_{p,0}(X)$, where $\mu^{ps}_{p,t}(x)$ is the model for the true but unknown outcome regression $\gamma^{ps}_{p,t}(x) = E[Y_t|P = p, X = x]$ with $p, t = 0,1$, where the true, but unknown, $\gamma_{p,t}(x) = E[Y_t|P = p, X = x]$. The DR DID estimator for the ATT results to be $\hat{\tau}^{DRDID} = E\left[\left(w^{ps}_1(P) - w^{ps}_0(P, X, \pi)\right)\left(\Delta Y - \mu^{ps}_{0,\Delta}(X)\right)\right]$, where, for a generic function $g$, we have that $w^{ps}_1(P) = \frac{P}{E(P)}$ and $w^{ps}_0(P, X; g) = \frac{g(X)(1-P)}{1-g(X)} / E\left[\frac{g(X)(1-P)}{1-g(X)}\right]$.

## 3. Data

The NPs under consideration here are 8, depicted in Figure 1. They are classified as:

- "northern NPs", for a total number of 39 involved municipalities, including the Parco Nazionale delle Dolomiti Bellunesi (1990), the Parco Nazionale della Val Grande (1992), the Parco Nazionale delle Foreste Casentinesi, Monte Falterona e Campigna (1993);
- "central NPs", with 98 municipalities covered, including the Parco Nazionale dei Monti Sibillini (1993), the Parco Nazionale del Gran Sasso e Monti della Laga (1995), the Parco Nazionale della Maiella (1995);

DRAFT

- "southern NPs", with 68 municipalities covered, including the Parco Nazionale del Pollino (1993) and the Parco Nazionale del Vesuvio (1995).

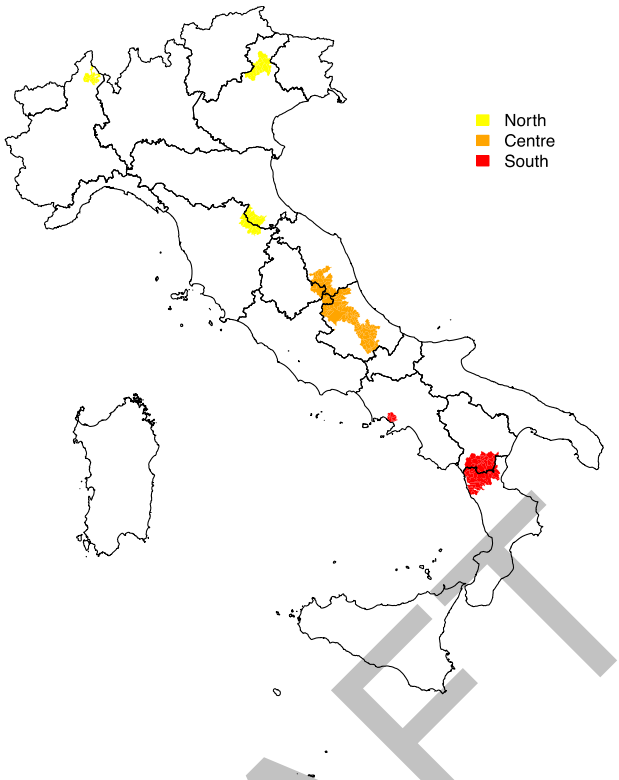



Figure 1: National Parks considered.

The 8 NPs (out of 25 Italian NPs [16]) were all established in the 1990s and include 205 municipalities, for a total surface of 11,148.22 km$^2$ (3.69% of the country area) and 947,703 inhabitants (1.67% of the Italian population). We use data on the geographical characteristics of the municipalities, the 1991, 2001, and 2011 *Census of Population and Housing*, and *Census of Industry, Services and Non-profit Institutions*, the 1990, 2000, and 2010 *Census of Agriculture* [17], the CORINE Land Cover source [6], while other data have been elaborated by the authors based on Istat data.

Two sets of variables are used: 1) the matching variables (covariates), for pairing the "treated" and the "untreated" municipalities; 2) the outcome variables on which the NPs impact is estimated.

7,423 municipalities not included in any of the NPs represent the potential counterfactual units, selected by matching the 205 municipalities in the NPs of interest by means of the altitude above sea level (in meters), the distance from the coast (in km), the area of the municipality (in km$^2$), the number of local establishment (firms or section of firms that operate within the municipality area) at year 1981, the number of workers employed in local establishments at year 1981, the population density (inhabitants/km$^2$) at year 1981, the percentage of urbanized land at year 1990.

The impact of the NPs is assessed from the baseline year 1991 to the follow-up years 2001 and 2011, on the following outcome variables: population density, number of local establishments, number of workers employed in these local establishments, number of local establishments in the tourism sector, number of workers employed by local establishments in the tourism sector, incoming, outgoing, and within-municipalities work-commuters, percentage of forested land.

## 4. Results

For the sake of brevity, we do not depict the balancing results of the best matching approach that is selected (among the 13 approaches tested) for the ATT estimation, i.e., the one producing the highest level of balance between the observable characteristics of the treated and the untreated municipalities: the Bayesian additive regression tree propensity score matching. It performs well in terms of the Standardized Mean Difference (SMD), the variance ratio, the mean and maximum of the empirical cumulative density function (eCDF). Following [2,8,27], the optimality of balance is assessed by considering that

DRAFT

SMD values < 0.1 indicate "perfect balance", values between 0.1 and 0.2 indicate "good balance", while SMD > 0.2 hints at concerns about covariates imbalance. For the northern NPs, 6 out of 8 matching variables are perfectly balanced, while the others present a good level of balance. About the central NPs, 5 matching variables are perfectly balanced, while 3 are well balanced. Finally, considering the southern NPs, 6 matching variables present perfect balance, 1 present good balance, while the distance from the coast has a SMD value slightly higher than 0.2. Considering the neighboring municipalities and the analysis of the spillover effects, the matching strategy performs well for the northern group, perfectly balancing 6 out of 8 matching variables, while the others present a good level of balance. About the central NPs, 6 matching variables are perfectly balanced, 1 is well balanced, while 1 (altitude above sea level) is unbalanced. Regarding the southern NPs, 4 matching variables present perfect balance, while 4 are well balanced.

## 4.1 Municipalities in NPs

Table 1 depicts the ATT estimation on the municipalities in NPs.

Table 1: Estimation of the Average Treatment Effect on the Treated (ATT)

| Outcome variable | 1991-2001 | | | 1991-2011 | | |
|---|---|---|---|---|---|---|
| | Northern NPs | Central NPs | Southern NPs | Northern NPs | Central NPs | Southern NPs |
| Population density | 3.299<br>(44.795)<br>[-86.064, 92.662] | -0.719<br>(0.523)<br>[-1.776, 0.338] | -19.840<br>(15.952)<br>[-52.879, 13.198] | 0.007<br>(1.853)<br>[-3.703, 3.720] | -1.110<br>(1.019)<br>[-3.300, 1.080] | -45.374<br>(26.047)<br>[-105.579, 14.831] |
| Nr. of local establishments | -1.377<br>(6.280)<br>[-14.066, 11.313] | -4.668<br>(10.358)<br>[-29.545, 20.209] | 7.899<br>(23.972)<br>[-42.623, 58.422] | 21.760<br>(21.458)<br>[-24.644, 68.164] | 10.170<br>(25.165)<br>[-59.843, 80.183] | -48.038<br>(54.728)<br>[-160.122, 64.047] |
| Nr. of workers employed in local establishments | 99.987<br>(64.462)<br>[-34.729, 234.702] | -7.772<br>(23.205)<br>[-57.259, 41.715] | -1.054<br>(4.731)<br>[-11.595, 9.487] | 139.215<br>(94.036)<br>[-54.386, 332.816] | 8.519<br>(7.556)<br>[-8.964, 26.002] | -108.841<br>(121.148)<br>[-351.914, 134.232] |
| Nr. of tourism sector establishments | -4.726<br>(2.432)<br>[-9.896, 0.445] | 0.035<br>(0.778)<br>[-1.529, 1.598] | -2.112<br>(1.600)<br>[-5.274, 1.049] | -4.692<br>(4.262)<br>[-11.167, 1.782] | 0.144<br>(1.753)<br>[-3.912, 4.199] | -2.815<br>(3.634)<br>[-10.366, 4.736] |
| Nr. of workers employed in tourism sector establishments | 0.177<br>(5.540)<br>[-11.749, 12.102] | -0.899<br>(1.984)<br>[-5.031, 3.232] | -1.044<br>(5.055)<br>[-12.041, 9.953] | -0.643<br>(13.596)<br>[-30.575, 29.288] | **14.459***<br>**(5.663)**<br>**[4.628, 33.546]** | -16.795<br>(13.105)<br>[-44.308, 10.718] |
| Nr. of incoming commuters | **212.336*****<br>**(71.759)**<br>**[58.033, 366.640]** | **54.859*****<br>**(21.363)**<br>**[5.927, 103.792]** | -47.792<br>(71.256)<br>[-214.729, 119.145] | **283.850*****<br>**(106.158)**<br>**[70.902, 496.797]** | 33.833<br>(43.965)<br>[-65.918, 133.584] | -174.459<br>(107.686)<br>[-397.663, 48.745] |
| Nr. of outgoing commuters | 37.276<br>(34.533)<br>[-30.258, 104.810] | -12.688<br>(14.971)<br>[-44.068, 18.704] | -74.491<br>(79.441)<br>[-254.787, 105.805] | 54.688<br>(59.539)<br>[-65.659,175.036] | -53.562<br>(32.923)<br>[-131.582, 24.457] | -179.901<br>(95.080)<br>[-393.535, 33.733] |
| Nr. of within-municipalities commuters | -44.203<br>(31.772)<br>[-115.058, 26.652] | -5.975<br>(14.233)<br>[-36.103, 24.153] | 31.224<br>(60.312)<br>[-104.120, 166.567] | -17.185<br>(34.039)<br>[-90.830, 56.460] | -39.807<br>(30.332)<br>[-111.498, 31.885] | 8.059<br>(66.948)<br>[-129.150, 145.267] |
| Percentage of forested land | **0.001***<br>**(0.000)**<br>**[0.000, 0.004]** | 0.001<br>(0.001)<br>[-0.002, 0.011] | **0.002*****<br>**(0.000)**<br>**[0.000, 0.004]** | **0.009*****<br>**(0.003)**<br>**[0.004, 0.015]** | **0.024*****<br>**(0.006)**<br>**[0.010, 0.038]** | **0.020*****<br>**(0.006)**<br>**[0.008, 0.032]** |

Standard errors in parenthesis. 95% confidence intervals in brackets. Significance levels: '*': 0.1; '**': 0.05; '***': 0.01.

## 4.2 Spillover effects

Table 2 depicts the ATT estimation on the neighboring municipalities of those in NPs, considered as those municipalities having at least one common border with one of the municipalities included within one of the 8 NPs under consideration.

Table 2: Estimation of the spillover effects (ATT)

| Outcome variable | 1991-2001 | | | 1991-2011 | | |
|---|---|---|---|---|---|---|
| | Neighbor municipalities, northern NPs | Neighbor municipalities, central NPs | Neighbor municipalities, southern NPs | Neighbor municipalities, northern NPs | Neighbor municipalities, central NPs | Neighbor municipalities, southern NPs |
| Population density | 1.053<br>(2.257)<br>[-4.174, 6.280] | 1.464<br>(1.694)<br>[-2.122, 5.049] | -59.673<br>(82.058)<br>[-238.725, 119.379] | -4.166<br>(3.292)<br>[-11.241, 2.908] | 1.352<br>(4.347)<br>[-8.638, 11.343] | -108.651<br>(120.107)<br>[-364.228, 146.926] |
| Nr. of local establishments | -20.951<br>(12.426)<br>[-50.467, 8.566] | 1.601<br>(13.701)<br>[-27.474, 30.687] | 208.439<br>(153.094)<br>[-170.016, 586.893] | -25.879<br>(26.848)<br>[-91.436, 39.679] | 30.392<br>(41.818)<br>[-62.336, 123.120] | 608.433<br>(493.180)<br>[-569.689, |

DRAFT

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | 1,786.554] |
| Nr. of workers employed in local establishments | **69.048*** **(32.088)** **[17.832, 155.927]** | 19.637 (34.131) [-52.216, 91.491] | -265.077 (208.674) [-711.551, 181.398] | **73.227*** **(53.075)** **[9.842, 170.530]** | 74.156 (66.224) [-63.065, 211.376] | 557.434 (570.218) [-827.962, 1,942.829] |
| Nr. of tourism sector establishments | -1.314 (0.957) [-3.264, 0.636] | 1.257 (0.842) [-0.508, 3.022] | 4.689 (6.729) [-11.993, 21.372] | -2.913 (2.080) [-7.430, 1.604] | 2.068 (2.359) [-2.791, 6.927] | 28.077 (26.692) [-36.983, 93.137] |
| Nr. of workers employed in tourism sector establishments | -3.279 (5.575) [-15.606, 9.048] | 0.673 (2.531) [-4.771, 6.117] | 29.538 (35.495) [-53.505, 112.581] | 8.646 (13.864) [-23.791, 41.082] | 13.241 (11.415) [-11.926, 38.409] | 98.417 (109.444) [-157.278, 354.112] |
| Nr. of incoming commuters | **77.718*** **(33.071)** **[5.878, 161.314]** | 17.491 (33.570) [-54.876, 89.858] | -879.618 (692.543) [-2,590.826, 831.590] | 43.699 (62.559) [-86.053, 173.451] | 40.351 (54.340) [-70.416, 151.119] | 386.138 (831.347) [-1,921.076, 2,693.352] |
| Nr. of outgoing commuters | 94.471 (59.399) [-43.706, 232.648] | 16.741 (32.674) [-58.229, 91.711] | -607.578 (483.794) [-1,733.222, 518.065] | 73.799 (62.928) [-67.241, 214.838] | 12.219 (61.734) [-119.649, 144.087] | 175.094 (577.064) [-1,449.971, 1,800.158] |
| Nr. of within-municipalities commuters | 8.805 (27.213) [-46.581, 64.192] | -7.289 (34.319) [-83.943, 69.366] | -487.373 (465.844) [-1,602.785, 628.039] | 1.757 (33.009) [-69.714, 73.227] | -15.431 (46.457) [-118.667, 87.805] | 144.551 (516.278) [-1,462.293, 1,751.395] |
| Percentage of forested land | 0.000 (0.001) [-0.002, 0.002] | 0.002 (0.001) [-0.004, 0.000] | **0.002** **(0.000)** **[0.002, 0.002]** | 0.002 (0.004) [-0.011, 0.007] | **0.012*** **(0.006)** **[0.002, 0.025]** | **0.019***** **(0.007)** **[0.005, 0.033]** |

Standard errors in parenthesis. 95% confidence intervals in brackets. Significance levels: '*': 0.1; '**': 0.05; '***': 0.01.

## 5. Discussion

The evidence from the data at hand suggests that there is a positive impact on the percentage of forested land in the municipalities in a NP, existing specific time dynamics in the effects. In the medium run, the impact is larger than in the short run. The results are more nuanced considering the socioeconomic impacts. It is observed a positive effect on the number of workers employed in the tourism sector establishments in central NPs in the medium run, while there is an increase in the number of incoming work-commuters in the northern NPs both in the short and medium run, and in central NPs limited to the short run. The scrutiny of the role of geographical spillovers highlights socioeconomic leakages to the surrounding municipalities, e.g., in terms of the positive impact on the number of workers employed in local establishments in neighbouring municipalities of the northern NPs (both in the short and medium run), as well as in the number of incoming work-commuters (in the short run) in the neighbouring municipalities of the northern NPs. The spillover effects are supposed to be stronger in environmental terms, with central and southern NPs affected by an increase in the percentage of forested land. Further analyses could be devoted to the investigation of such spillovers by adopting a robust spatial model. The main policy implication is that the data at hand support the idea that NPs benefit not only the environment, but they can support the socioeconomic development of local communities as well. This does not apply limitedly to Global South countries but to high-income countries as well.

DRAFT

## References


[1] Abadie, A., Drukker, D., Herr, J.L., Imbens, G.W.: Implementing Matching Estimators for Average Treatment Effects in Stata. Stata J. (2004) doi: 10.1177/1536867X0400400307

[2] Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. (2009): doi: 10.1002/sim.3697

[3] Braber, B., den Evans, K.L., Oldekop, J.A.: Impact of protected areas on poverty, extreme poverty, and inequality in Nepal. Conserv. Lett. (2018) doi: 10.1111/conl.12576

[4] Chen, H., Zhang, T., Costanza, R., Kubiszewski, I.: Review of the approaches for assessing protected areas' effectiveness. Environ. Impact Asses. (2023) doi: 10.1016/j.eiar.2022.106929

[5] Chen, Y., Lewis, D.J., Weber, B.: Conservation Land Amenities and Regional Economies: A Postmatching Difference-in-Differences Analysis of the Northwest Forest Plan. J. Regional Sci. (2016) doi: 10.1111/jors.12253

[6] Copernicus: CORINE Land Cover, URL: land.copernicus.eu/pan-european/corine-land-cover (2022)

[7] D'Alberto, R., Pagliacci, F., Zavalloni, M.: A socioeconomic impact assessment of three Italian national parks. J. Regional Sci. (2023) doi: 10.1111/jors.12618

[8] Diamond, A., Sekhon, J.S.: Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Rev. Econ. Statistics (2013) doi: 10.1162/REST_a_00318
[9] Ferraro, P.J., Hanauer, M.M.: Protecting Ecosystems and Alleviating Poverty with Parks and Reserves: 'Win-Win' or Tradeoffs? Environ. Resource Econ. (2011) doi: 10.1007/s10640-010-9408-z
[10] Ferraro, P.J., Hanauer, M.M.: Quantifying causal mechanisms to determine how protected areas affect poverty through changes in ecosystem services and infrastructure. P. Natl. Acad. Sci. USA (2014) doi: 10.1073/pnas.1307712111
[11] Fox, H.K., Swearingen, T.C.: Using a difference-in-differences and synthetic control approach to investigate the socioeconomic impacts of Oregon's marine reserves. Ocean Coast. Manage. (2021) doi: 10.1016/j.ocecoaman.2021.105965
[12] Heckman, J.J., Robb, R.Jr.: Alternative methods for evaluating the impact of interventions: An overview. J. Econometrics (1985)
[13] Heckman, J.J., Ichimura, H., Todd, P.E.: Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. Rev. Econ. Stud. (1997) doi: 10.2307/2971733
[14] Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Polit. Anal. (2007) doi: 10.1093/pan/mpl013
[15] Ho, D.E., Imai, K., King, G., Stuart, E.A.: MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Softw. (2011) doi: 10.18637/jss.v042.i0
[16] Italian National Ministry of Ecological Transition: Elenco dei Parchi Nazionali Italiani - Ministero Italiano per la Transizione Ecologica, URL: www.mite.gov.it/pagina/elenco-dei-parchi (2022)
[17] Istat: Italian National Institute of Statistics – Datawarehouse, URL: http://dati.istat.it/ (2022)
[18] Mattioli, W., Ferrara, C., Colonico, M., Gentile, C., Lombardo, E., Presutti Saba, E., Portoghesi, L.: Assessing forest accessibility for the multifunctional management of protected areas in Central Italy. J. Environ. Plann. Man. (2022) doi: 10.1080/09640568.2022.2106554
[19] McNally, C.G., Uchida, E., Gold, A.J.: The effect of a protected area on the tradeoffs between short-run and long-run benefits from mangrove ecosystems. P. Natl. Acad. Sci. USA (2011) doi: 10.1073/pnas.1101825108
[20] Pelegrina-López, A., Ocana-Peinado, F., Henares-Civantos, I., Rosúa-Campos, J.L., Serrano-Bernardo, F.A.: Analyzing social perception as a key factor in the management of protected areas: the case of Sierra Nevada Protected Area (S Spain). J. Environ. Plann. Man. (2017) doi: 10.1080/09640568.2017.1291413
[21] Romano, B., Zullo, F., Fiorini, L., Marucci, A.: "The park effect"? An assessment test of the territorial impacts of Italian National Parks, thirty years after the framework legislation. Land Use Policy (2021) doi: 10.1016/j.landusepol.2020.104920
[22] Rosenbaum, P.R., Rubin, D.B.: Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician (1985) doi: 10.2307/2683903
[23] Rubin, D.B.: Direct and Indirect Causal Effects via Potential Outcomes. Scand. J. Stat. (2004)
[24] Sant'Anna, P.H.C., Zhao, J.: Doubly robust difference-in-differences estimators. J. Econometrics (2020) doi: 10.1016/j.jeconom.2020.06.003
[25] Sims, K.R.E.: Conservation and development: Evidence from Thai protected areas. J. Environ. Econ. Manag. (2010).
[26] Sims, K.R.E., Thompson, J.R., Meyer, S.R., Nolte, C., Plisinski, J.S.: Assessing the local economic impacts of land protection. Conserv. Biol. (2019) doi: 10.1111/cobi.13318
[27] Stuart, E.A., Huskamp, H.A., Duckworth, K., Simmons, J., Song, Z., Chernew, M.E., Barry, C.L.: Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology (2014) doi: 10.1007/s10742-014-0123-z
[28] United Nations: Update of the zero draft of the post-2020 global biodiversity framework, Convention on Biological Diversity (2020)
[29] Visconti, P., Butchart, S.H.M., Brooks, T.M., Langhammer, P.F., Marnewick, D., Vergara, S., Yanosky, A., Watson, J.E.M.: Protected area targets post-2020. Science (2019) doi: 10.1126/science.aav6886
[30] Weiler, S., Seidl, A.: What's in a Name? Extracting Econometric Drivers to Assess The Impact of National Park Designation. J. Regional Sci. (2004) doi: 10.1111/j.0022-4146.2004.00336.x

DRAFT

# On the gap between emitted and absorbed carbon dioxide. Are trees enough to save us?

Lorenzo Mori [a] and Maria Rosaria Ferrante[a]

[a]Department of statistical science P. Fortunati , Via Belle Arti, 41 - Bologna - Italy ;
lorenzo.mori7@unibo.it, maria.ferrante@unibo.it

**Abstract**

The Carbon Footprint (CFP) is one of the most common indicators to quantify environmental pollution and it is based on household consumption. In this paper, we estimate the per capita CFP of the Italian regions and we compare CFP with the $CO_2$ absorbed by forest per inhabitant. The idea that is enough to plant more and more trees to obtain carbon neutrality, i.e. emitted $CO_2$ equal to the intake one, is debunked.

***Keywords:*** Carbon footprint, climate change, $CO2_{eq}$, forest inventory, HBS

DRAFT

## 1. Introduction

Nowadays the problem of climate change is quite evident. Since 1972 (10) researchers have tried to find possible solutions to the challenge of carbon emission reduction. Two of the most recent international meeting on this topic are the 2015 conference about Sustainable Development Goals (14) and the 2021 United Nations Climate Change Conference (12). In particular, the impact of human activities on greenhouse gas emissions is usually studied and measured by the Carbon Footprint (now on CFP). For a complete review of the literature of the CFP measure see (6). An important aim in the field of CFP measure is to obtain carbon neutrality within 2100, but how can it be done? This work aims to contributes to this debate debunking the very widespread idea that to plant trees is enough to reduce the $CO_2$. To achieve this result we estimate the CFP from the household expenditure and we compare it with the absorbed $CO_2$ by the forest.

In order to obtain the CFP estimation matrices relationship based on Input-Output Tables (IoT) and consumption data are usually adopted. Two of the most known approaches are presented in (13) and (7). (13) starting from Swiss data, define the CFP as a sum of three components and use national IoTs. (7) substitute the IoTs with the EXIOBASE dataset that has the same structure of the IoTs, but it is appositely created to estimate the CFP by (15) and it is suitable for more than 41 countries. Both approaches, in our view, could be improved. Indeed, the EXIOBASE dataset has been developed in 2014 and not updated data could bring to unreliable estimates. On the other hand, the method developed by (13) need to have all the involved datasets expressed with the same classification. This request cannot be satisfied by data produced under the guidelines of EUROSTAT that use different classifications for firm production and household consumption. In order to solve this problem, we define a strategy that uses a matrix bridging the two classifications. Using the estimated CFP and the intake carbon dioxide by forests, we verify the idea that to plant several trees could have a central role in the achievement of carbon neutrality. The difference between the two dimensions, that is the gap between per capita emitted and absorbed carbon dioxide, is huge and, mostly, heterogeneous among regions. This suggests that forests will not be able to completely absorb carbon dioxide and that regional estimates (7) can help to design equitable carbon place-based policies leading to engage local actors (9).

## 2. Data

The estimation of the household CFP based on household consumption data requires information coming from the following datasets:

- Household Budget Survey (HBS) data, reporting household expenditure and published by the Italian Statistical Institute (ISTAT). In this survey data are classified following the Classification of Individual Consumption by Purpose (COICOP).
- National Accounting Matrix, including Environmental Accounts (NAMEA) and reporting the emission intensities. NAMEA is an air emissions report calculated starting from the national emission inventory annually produced by Istituto Superiore per la Ricerca e la Protezione Ambientale (Ispra). In this dataset, air pollutants are reported in $CO_2$ equivalent. The aggregation level used is NACE Rev. 2 for the economic activities and the Classification of Products by Activity (CPA) for products.
- Input-Output table (IoT), published by ISTAT, reporting matrix of supply and use for each economic sector. The classification used for economic activities and for products is the same as of NAMEA.

## 3. The carbon footprint estimation

Here we present the strategy to obtain the measure of CFP. Let $i$ ($i = 1, \ldots, n$) indicate the sampled units in the region $j$ ($j = 1, \ldots, J$), $m$ the number of COICOP in the HBS and $\nu$ the number of CPA. Following (13), the $CFP_{ij}$ is defined as:

$$CFP_{ij} = IND_{ij} + DIR_{ij} + EEI_{ij} \tag{1}$$

where IND, DIR and EEI indicates respectively indirect emissions, direct emissions and embodied emissions in import. Indirect emission are the emissions induced by final consumption of domestic, imported goods and services. The measure of $IND_{ij}$ is based on a matrix relationship that involves HBS and IoT. Unfortunately these two informative sets are not directly comparable due to the different classification. This problem is present in each country (also Italy) that follows the European Classification. In order to achieve comparability, we use the conversion matrix (**cf**) obtained by (3) and (4) for each European country, and available in the supplementary material of (4). This matrix, obtained through the count-seed RAS approach, allows us to bridge data on consumption categories (COICOP in HBS) and on final use categories (CPA in IoTs) and then to re-define formulas proposed in (13). At the best of our knowledge, this strategy has never been used to derive $IND_{ij}$ in the estimation of CFP. The new formulation is:

DRAFT

$$IND_{ij} = \underbrace{\begin{bmatrix} e_1 \\ \vdots \\ e_\nu \end{bmatrix}^T}_{\text{e}} \times \underbrace{\left[ \underbrace{\begin{bmatrix} 1 \dots 0 \\ \vdots \ddots \vdots \\ 0 \dots 1 \end{bmatrix}}_{I_\nu} - \underbrace{\begin{bmatrix} z_{1,1} \dots z_{1,\nu} \\ \vdots \ddots \vdots \\ z_{\nu,1} \dots z_{\nu,\nu} \end{bmatrix} \times \begin{bmatrix} \frac{1}{v_1} \dots 0 \\ \vdots \ddots \vdots \\ 0 \dots \frac{1}{v_\nu} \end{bmatrix}}_{\text{coefficient matrix (z×v)}} \right]^{-1}}_{(I-A)^{-1}}$$

$$\times \underbrace{\left[ \underbrace{\left[ \underbrace{\begin{bmatrix} cf_{1,1} \dots cf_{1,m} \\ \vdots \ddots \vdots \\ cf_{\nu,1} \dots cf_{\nu,m} \end{bmatrix}}_{\text{conversion factor (cf)}} \times \underbrace{\begin{bmatrix} c_{1i} \\ \vdots \\ c_{mi} \end{bmatrix}}_{\text{consumption (c)}} \right]^T}_{\substack{\text{converted} \\ \text{household} \\ \text{consumption}}} \times \underbrace{\begin{bmatrix} y_{1,1} \dots y_{1,\nu} \\ \vdots \ddots \vdots \\ y_{\nu,1} \dots y_{\nu,\nu} \end{bmatrix}}_{\text{final use (y)}} \right]^T}_{\text{Y}} \tag{2}$$

where **e** is the vector of emission intensities derived from the NAMEA air emission inventory, (**I**-**A**)$^{-1}$ denotes the Leontief's inverse matrix, and **Y** is the final demand matrix based on a vector of household expenditure across final use categories. (**I**-**A**)$^{-1}$ involves the product of matrix **z**, which represents the transaction between activities, and matrix **v**, which has on the principal diagonal the inverse of the total output. The final demand matrix **Y** involves the conversion matrix **cf**, the household expenditure and the final use matrix **y** . The other two components, DIR and EEI, of eq. 1 don't need to be re-defined and we use the formulation of (13). Direct emissions, which are associated with household consumption of fuels for transportation and housing, are measured as follows:

$$DIR_{ij} = \frac{cc_{ij}}{\sum_{i=1}^{n} cc_{i|j}} \times DHE \frac{n_j}{N_j} \tag{3}$$

where $cc_{ij}$ is the expenditure on combustible of the family $i$ in region $j$, $\sum_{i=1}^{n} cc_{ij}$ is the total expenditure on combustibles for all the families living in the same region and DHE is the total volume of direct household emissions from the NAMEA emissions inventory, scaled accordingly to the number of households surveyed in the HBS.

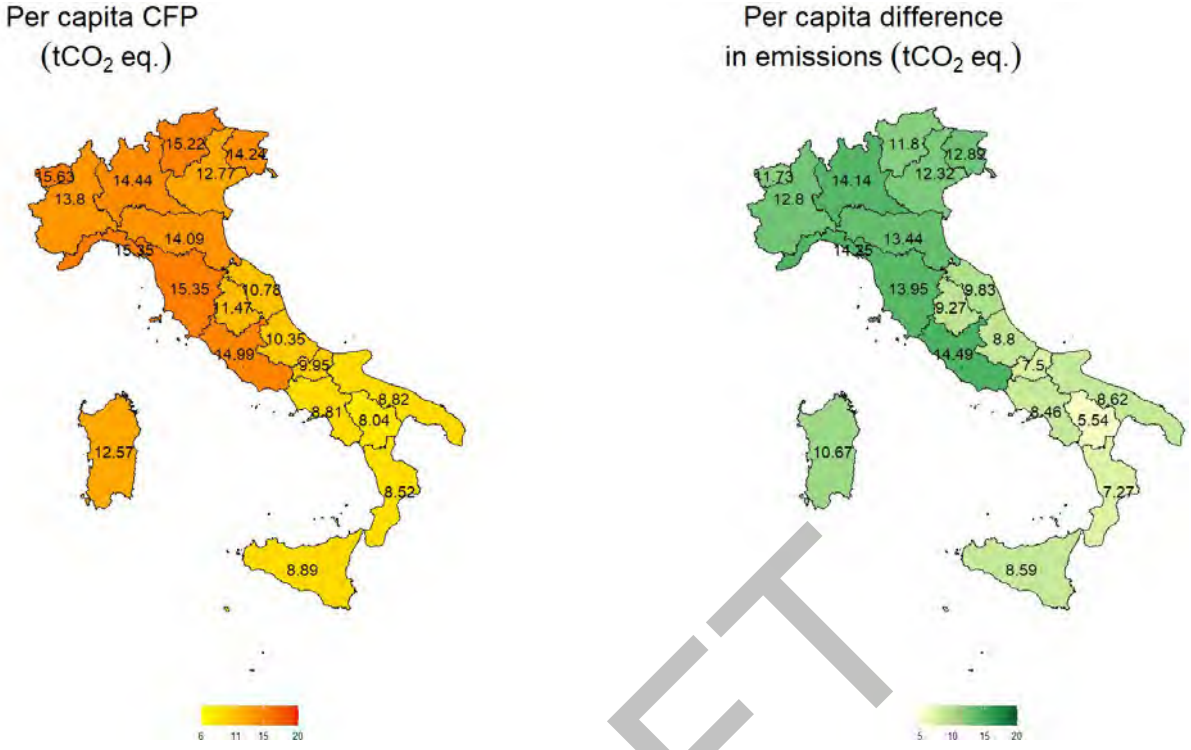
The embodied emissions in import of the household *i* in the region *j* are defined as:

$$EEI_{ij} = \underbrace{\begin{bmatrix} e_1^{EU} \\ \vdots \\ e_\nu^{EU} \end{bmatrix}^T}_{e^{UE}} \times \underbrace{\left[ \begin{bmatrix} Im_1 \dots 0 \\ \vdots \ddots \vdots \\ 0 \dots Im_\nu \end{bmatrix} \times \begin{bmatrix} \frac{1}{x_1} \dots 0 \\ \vdots \ddots \vdots \\ 0 \dots \frac{1}{x_\nu} \end{bmatrix} \right]}_{\text{import coefficient (Im×x)}} \times \underbrace{[I - A_s]^{-1}}_{\substack{\text{Leontief's inverse} \\ \text{of supply}}} \times Y \tag{4}$$

where $\mathbf{e}^{UE}$ is the vector of emission intensities derived from the NAMEA air emission inventory of EU countries, (**I**-$\mathbf{A}_s$)$^{-1}$ denotes the Leontief's inverse matrix this time respect the supply IoT matrices instead of total output, and **Y** is the final demand matrix based on a vector of household expenditure across final use as before. The import coefficient matrix is obtained by multiplying imports (**Im**) and matrix (**x**). This last matrix has on the principal diagonal the inverse of the total supply. To conclude, all tables are re-scaled in order to obtain results in tonnes of equivalent $CO_2$ ($tCO_2$eq.).

DRAFT

Figure 1: Choropleth maps of average CFP and difference between emitted and absorbed $CO_2$




DRAFT

## 4. Results

In this section, we will briefly summarize the main results. First of all, as proof of the goodness of the proposed strategy to measure $IND_{ij}$, it is important to note that the results obtained and those of (7) for 2016, with a completely different approach, present a clearly comparable path. Results on this comparison are available on requests. As expected, per-capita CFP (Fig. 1) is higher in Northern and Central Italy than it is in the South. The region with the highest per-capita CFP is Aosta Valley (15.63 $tCO_2$ eq.), while Basilicata shows the lowest level (8.04 $tCO_2$ eq.). With data published from the last Italian national forest inventory is possible to estimate how much $CO_2$ is intake regionally every year. To do this we use the forest area per inhabitant (per capita $km^2$) estimated by the inventory and the results published by (5) which asserts that the amount of $CO_2$ taken up by forests is around 500 tonnes per square kilometre per year. Subtracting the amount of caught-up $CO_2$ to the produced one (CFP) we obtain the gap between emitted and absorbed carbon dioxide. Basilicata remain the region with the lowest values while Lazio becomes the one with the highest. As expected, Aosta Valley reduces to only 11.73 $tCO_2$ eq. the per capita difference in emission. Other two notable regions are Umbria and Apulia. The first one, known to be the green lung of Italy, thanks to trees markedly reduce the per capita emission going from 11.47 to 9.27 $tCO_2$ eq.. Apulia, on the other hand, is the region with the least decrease in $CO_2$ emissions (-0.2 $tCO_2$ eq.). Is evident that forests are unable to absorb the carbon dioxide produced by consumption. In reality, they will never be, as to be able to do so they should have an extension equal to about double that of Italy. These results, although extreme, are in line with those obtained by (2) for UK.

## 5. Concluding remarks

We present a new method to compute the per capita CFP from consumption expenditure data and we compare them with the quantity of $CO_2$ absorbed by the Italian forests. Estimates, obtained at regional level, clearly show that trees are not sufficient to reduce the quantity of $CO_2$ and mostly that they cannot be, in the future, the only solution for the ambitious and primary goal of 0 emissions in

2100. Obviously, the contribution of trees in reducing $CO_2$ emissions is beyond question, but, as noted by (8), forest carbon sequestration should only be viewed as a component of a mitigation strategy and not as the solution. Then, the question we all need to answer is: which can be the way to achieve the objectives set with the Sustainable Development Goals? Answer to this question is not easy. Regions are different in terms of territorial conformation and of presence of both pollutants and forests. We obtain results showing the heterogeneity among regions of the difference between carbon dioxide absorbed and produced. This highlight the necessity of place-based policies that encourage regions to adopt peculiar tools suitable at the local level. The definition of a regional policy would not mean creating differences but rather aiming at the different strengths that each region has in order to remove their critical points in the path of CFP reduction.

## References


[1] Birdsey, R. A. Carbon storage and accumulation in united states forest ecosystems (Vol. 59). US Department of Agriculture, Forest Service. (1992)

[2] Boysen, L. R., Lucht, W., Gerten, D., Heck, V., Lenton, T. M., Schellnhuber, H. J. The limits to global-warming mitigation by terrestrial carbon removal. Earth's Future, 5(5), 463–474. (2017)

[3] Cai M., Manuel J. Bridging macroeconomic data between statistical classifications: the count-seed RAS approach. Economic Systems Research, 31, 382-403, (2019)

[4] Cai M., Vandyck T. Bridging between economy-wide activity and household-level consumption data: Matrices for European countries. Data in Brief, 30, 105395, (2020)

[5] CREA, Le foreste Italiane. Sintesi dei risultati del terzo Inventario Forestale Nazionale INFC2015. Arma dei Carabinieri Comando Unità Forestali, Ambientali e Agroalimentari. (2021)

[6] Heinonen, Jukka, Ottelin, J., Ala-Mantila, S., Wiedmann, T., Clarke, J., Junnila, S. Spatial consumption-based carbon footprint assessments-A review of recent developments in the field. Journal of Cleaner Production, 256: 120335. (2020)

[7] Ivanova, D., Vita, G., Steen-Olsen, K., Stadler, K., Melo, P. C., Wood, R., Hertwich, E. G. Mapping the carbon footprint of EU regions. Environmental Research Letters, 12(5), 054013. (2017)

[8] Malhi, Y. and Meir, P. and Brown, S. Forests, carbon and global climate. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. The Royal Society. pp 1567-1591 (2002)

[9] McCann, P., Soete, L. Place-based innovation for sustainability. Publications Office of the European Union, Luxembourg. (2020)

[10] Meadows, D. H., Meadows, D. L., Randers, J., Behrens, W. W. The limits to growth. Green planet blues. Routledge. pp. 25–29. (2018)

[11] Miller, R. E., Blair, P. D. Input-output analysis: Foundations and extensions. Cambridge university press. (2009)

[12] Moosmann, L., Siemons, A., Fallasch, F., Schneider, L., Urrutia, C., Wissner, N., Oppelt, D. The cop26 climate change conference (tech. rep.). (2021)

[13] Pang, M., Meirelles, J., Moreau, V., Binder, C. Urban carbon footprints: A consumption-based approach for Swiss households. Environmental Research Communications, 2(1), 011003. (2019)

[14] United-Nations. Transforming our world: The 2030 agenda for sustainable development, 21 october 2015 (tech. rep.). A/RES/70/1. (2015)

[15] Wood, R., Stadler, K., Bulavskaya, T., Lutter, S., Giljum, S., De Koning, A., Kuenen, J., Schütz, H., Acosta-Fernández, J., Usubiaga, A., et al. Global sustainability accounting—developing exiobase for multi-regional footprint analysis. Sustainability, 7(1), 138–163. (2014).

[16] Wright, L., Kemp, S., Williams, I. Carbon footprinting: Towards a universally accepted definition. Carbon management, 2(1), 61–72. (2011)


DRAFT

# Small scale analysis of energy vulnerability in the municipality of Palermo

Giuliana La Mantia[a]

[a]First Address of Institute; giuliana.lamantia@unipa.it

**Abstract**

The current increase in energy prices has raised concerns about the repercussions on households, particularly the most vulnerable ones. This paper aims to provide a preliminary analysis of energy poverty in the municipality of Palermo. Based on the evidence from the most recent literature, we select a set of vulnerability indicators, also taking into account data availability. We then perform a principal component analysis to reduce the dataset into a few components and map them to the urban context under study. The analysis identifies two distinct vulnerability profiles concentrated in the historic centre and the Filippo-Neri, Roccella and Brancaccio-Ciaculli neighbourhoods.

***Keywords:*** energy vulnerability, energy justice, principal component analysis

DRAFT

## 1. Introduction

Domestic energy deprivation or Energy Poverty (EP) concerns the accessibility and affordability of energy goods and services. A household is energy poor if is unable to access adequate levels of energy services to ensure the satisfaction of its essential needs (3). Over the past decade, the literature on EP has grown significantly, with different methodologies for measuring and monitoring it emerging. A strand of literature has focused on the spatial distribution of EP and its interaction with structural deprivation and inequality at various geographic scales (4). While national assessments have helped frame the scale of the issue and enable cross-country comparisons, small-scale studies have revealed significant disparities between groups and locations that country-level analysis cannot capture (8). This report aims to contribute to the literature on local EP analysis by conducting a preliminary study of the municipality of Palermo.

## 2. The new theoretical framework of Energy Poverty

The notion of Energy Poverty was first studied in the UK where it was recognized as an issue of energy affordability, mainly in terms of heating costs. The leading causes of energy poverty were identified as a combination of the triad: low income, high energy prices, and inefficient housing (1; 6). The literature has expanded over the past decade and has embedded energy poverty in new theoretical frameworks that consider the particularities of specific groups and contexts and their interaction with other structural inequalities (2; 9; 12).

The Energy Vulnerability framework attempts to provide a more comprehensive and nuanced analysis of energy poverty, highlighting different dimensions of the concept and identifying spatial patterns. Vulnerable groups include the elderly, disabled, and families with young children which spend more time

at home, and their particular needs require more energy consumption to achieve comfortable temperatures (7). Households in the private rental sector also face significant challenges, including high housing and energy costs, and difficulty negotiating efficiency improvements with landlords. On the other hand, students, precarious workers, residents of informal settlements and ethnic minorities are more likely to be unable to afford adequate energy services (2; 9). Generally speaking, researchers identified six risk factors that increase the risk of fuel poverty among households: access to energy services, affordability, level of energy efficiency, special needs of households and lack of public recognition of the issue (3). These factors interact with each other and interplay with other structural inequalities, such as job insecurity and housing market conditions. Energy poverty can be also viewed as the result of broader dynamics of injustice, as it involves unequal access to energy services, which are necessary for the satisfaction of basic needs (13).

The evidence shows that the driving forces of energy poverty are locally contingent and that there are clear geographical patterns of energy poverty at different scales (4; 12). Bouzarovski and Thomson (5) investigated the links between sociodemographic and housing vulnerabilities to energy poverty, and wider patterns of urban social inequality. The authors analyze how forms of energy injustice can arise from inequalities in the distribution of urban amenities combined with socioeconomic disparities to create distinctive geographies of segregation. Building on this literature, the report examines the factors that increase the risk of energy vulnerability of households in the city of Palermo, in order to contribute to the understanding of spatial patterns of energy poverty in the urban context.

## 2.1 Data and aims

Studies with higher spatial resolution can provide more specific information on energy poverty, leading to better identification of energy vulnerability for different contexts and groups.

Nevertheless, small-scale analyses must consider the limited availability of data at such a granular level. At the Italian level, population and housing census data is the most spatially detailed source of information. The National Institute of Statistics (ISTAT) publishes the spatial basis system associated with the data collected with the 2011 Census, which provides aggregated information on the socio-demographic characteristics of the population and on the status of dwellings.

This analysis focuses on the municipality of Palermo, specifically our observation units are 43 census areas, identified by ISTAT. The areas are census sections aggregated with respect to demographic and social characteristics. Two sets of census variables were selected based on vulnerability factors identified in the literature and data availability. The first group includes demographic and socioeconomic characteristics of the population, such as the percentage of elderly people and children, the number of foreigners per thousand Italian residents, and the percentage of families experiencing economic deprivation (where no member receives income from work or pension). It also includes information on the type of tenure (owned, rented, or other) of the families' homes. The second group includes information about the residential buildings themselves, such as their state of preservation, construction period, and size (number of floors).

We point out two significant constraints in the analysis. Firstly, the data was gathered approximately ten years ago during the last census. Secondly, expenditure and energy consumption data are unavailable, and thus *proxies* such as employment and specialization levels and building conditions are utilized.

In the next stage of the analysis, we will examine the municipal area to see if a pattern of vulnerability to energy poverty can be identified. The aim of the analysis is twofold: first, we want to reduce the multivariate dimension of our dataset into a few components. Then we map the result of the analysis in order to investigate if exists a spatial pattern of vulnerability to energy poverty and how interplays with structural urban inequality. In the next section, we provide some descriptive analysis and then move on to the principal component analysis.

DRAFT

Table 1: Descriptive statistics.*Source*: Census data, 2011.

| *Sociodemographic variables* | Italy | Sicily | Palermo |
|---|---|---|---|
| People above 75 years | 10.4 | 9.4 | 8.9 |
| Children under six years | 5.6 | 5.7 | 5.9 |
| Unemploymente rate | **11.4** | **21.8** | **24.4** |
| Low-skilled employment | 16.2 | 20.7 | 18.9 |
| Young people outside the labour market and training | **12.3** | **19.4** | **21.2** |
| Foreigners | 67.8 | 25 | 21.6 |
| Large household | 1.6 | 1.8 | 2.2 |
| Economic deprivation | **2.7** | **5.9** | **7.2** |
| Homeowner | 70.5 | 69.8 | 60.9 |
| Renters | 21.2 | 16.6 | 29 |
| Other occupation title | 10 | 14.9 | 11 |
| *State of preservation* | | | |
| Buildings in good state of preservation | 83.3 | 72.9 | 72.3 |
| Buildings in a mediocre state of preservation | **15** | **23.8** | **24.5** |
| Buildings in a poor state of preservation | **1.7** | **3.3** | **3.2** [a] |
| *Construction period* | | | |
| Buildings built before 1946 | **23.5** | **19.5** | **29.6** |
| Buildings built between 1946 and 1960 | 17.8 | 14.5 | 20 |
| Buildings built between 1961 and 1980 | 35.1 | 38.9 | 35.5 |
| Buildings after 1981 | 23.6 | 27.2 | 14.9 |
| *Floors number* | | | |
| Buildings with one floor | 14.2 | 26 | 13.8 |
| Buildings with two floors | 37.9 | 34.5 | 28.1 |
| Buildings with three floors | 21.4 | 19.6 | 18.2 |
| Buildings with four or more floors | 26.5 | 19.8 | 39.9 |
| *Census size* | | | |
| Total residents | 59.433.744 | 5.002.904 | 657.561 |
| Total households | 24.611.766 | 1.963.577 | 246.227 |
| Total number of dwelling* | 24.135.177 | 1.940.472 | 244.053 |
| Total number of residential buildings | 12.187.698 | 1.431.419 | 46.293 |

* Refers to dwellings occupied by residents

DRAFT

## 3. Descriptive analysis

Data from the 2011 Census (Table 1) indicates that the municipality of Palermo has higher rates of social and material deprivation compared to both national and regional averages. Specifically, the unemployment rate in Palermo (24.4%) is more than twice the national average of 11.4%. Additionally, 21.2% of the city's young population is neither employed nor pursuing further formation, a figure that exceeds the national average by 9 percentage point. The percentage of residential buildings in Palermo built before 1946 (30%) is 42% higher than the national average of 21.4% and almost double that of Sicily (16%). Moreover, the share of buildings in a poor state of preservation is 3.2%, a value that is in line with the regional average of 3.3% but double the national average of 1.7%.
To sum up, the data show that the municipality of Palermo is characterised by a higher proportion of old residential buildings in a poor state of preservation than the national average, and the vulnerability indicators have a higher value than both the national and regional level.

## 4. Principal Component analysis

Principal component analysis (PCA) allows us to reduce a multivariate data set into a few principal components. Starting from a set of correlated variables, the analysis generates new variables, linear combinations of the original ones. The new components are uncorrelated, so each component reflects distinct statistical dimensions in the data. The contribution of each variable to the formation of the components, known as the *loadings*, is quantified by the correlation between the principal component and the variable. The loading values can be either positive or negative, signifying the direction and strength of the relationship between the variables and the component. We conduct a PCA after standardising our set of variables to have zero means and unit variances, which ensures that the analysis is not affected by variable scales and that all variables are equally important. The first two components explain the 60% of the total variance. Fig. 1 shows the loading values between each variable and the first two components. The positive and negative values of the variables allow us to identify different profiles and the maps (Fig. 2 and Fig. 3) show us the new coordinates for the observations.

The first principal component explains 39% of the total variance and is positively correlated with various vulnerability indicators, including low education, unemployment, and economic deprivation. It is also associated with sociodemographic characteristics related to energy poverty risks, such as large families, young children, and renters, as well as modestly sized and poorly maintained residential buildings. Conversely, the component is negatively associated with owner-occupied homes, elderly people, and well-preserved, medium to large-sized buildings.
Fig. 2 shows the observation score, the higher the score, the greater the variance explained by the first component in those areas. The first component's positive contributions effectively account for the variance in Filippo-Neri, Centro Storico, Roccella, and Brancaccio-Ciaculli neighbourhoods, whereas its negative contributions accurately reflect Resuttana, Villa Sperlinga, and Malaspina-Palagonia.
The second component accounts for 22% of the total variance. It is negatively associated with foreigners, low-skilled workers, elderly persons and households living in rented or other property titles. This component is also negatively related to medium-sized and old residential buildings in poor condition and is mainly found in the historic centre and surrounding areas. On the other hand, the second component is positively correlated with small and newly built residential buildings in good condition (Fig. 3).

The analysis indicates the presence of two profiles of vulnerability, based on indicators positively associated with the first component and those negatively correlated with the second component. The first component is strongly linked to social and economic hardships, indicating low income and limited ability to invest in energy-saving measures. The presence of large families and children is also associated with this component, as families with children spend more time at home and have a higher energy consumption.
The first component well represents the variation of the Filippo Neri neighbourhood, a peripheral area of the city that has been the subject of several studies for its urban development that have made it one of the most segregated areas of the city (10). The historic centre is characterized by the second vulnerability

DRAFT

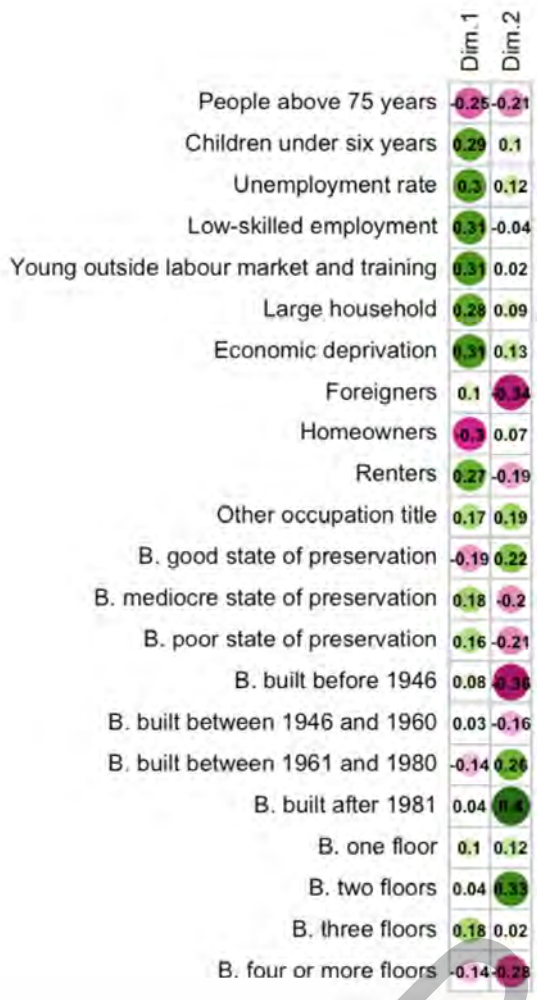



Figure 1: Loading values of the first two components.

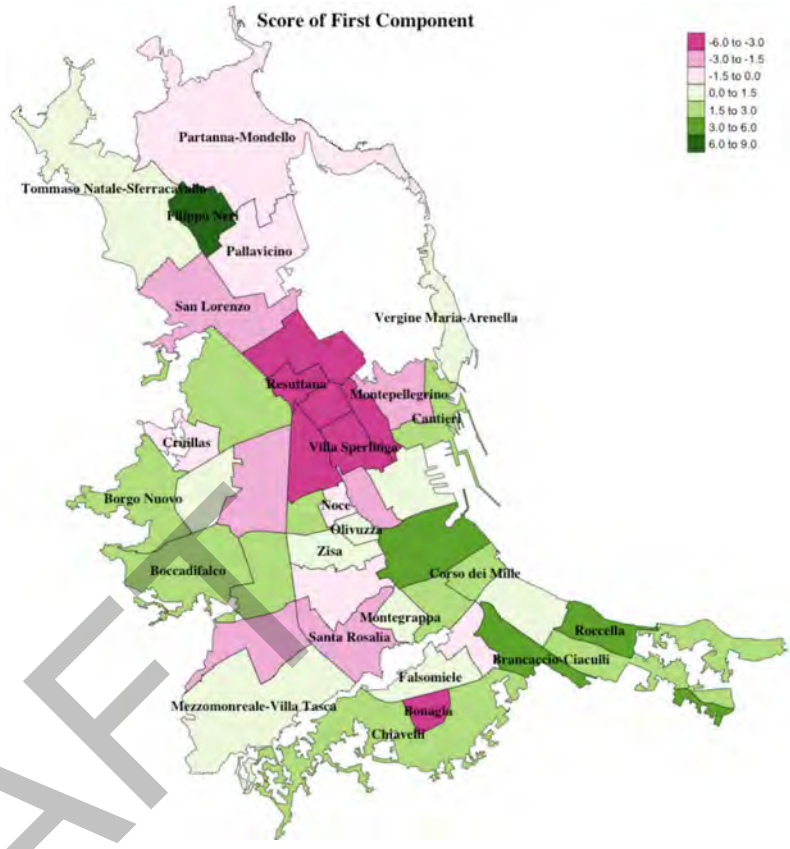



Figure 2: Score of the First Component.

DRAFT

profile identified, in which other types of vulnerability coexist such as the presence of foreigners and the elderly and old and badly preserved real estate. This area in recent years has been interested in several urban redevelopment projects that have changed the urban and social configuration, necessitating a more in-depth and up-to-date analysis (11).

## 5. Conclusion

In conclusion, this report contributes to the local PE analysis literature by conducting a preliminary study of the municipality of Palermo. The analysis revealed two profiles of vulnerability and shows their spatial distribution in the urban context. There are important limitations to this analysis. The available data comes from the 2011 Census, so we are aware that they need an update. The analysis uses only data aggregated by geographical area, therefore it is of the area-target type. Essential information such as spending and energy consumption data is not available, but proxies such as education level, building state and construction period are used. The plan for further analysis is to combine different data sources to get more accurate information.

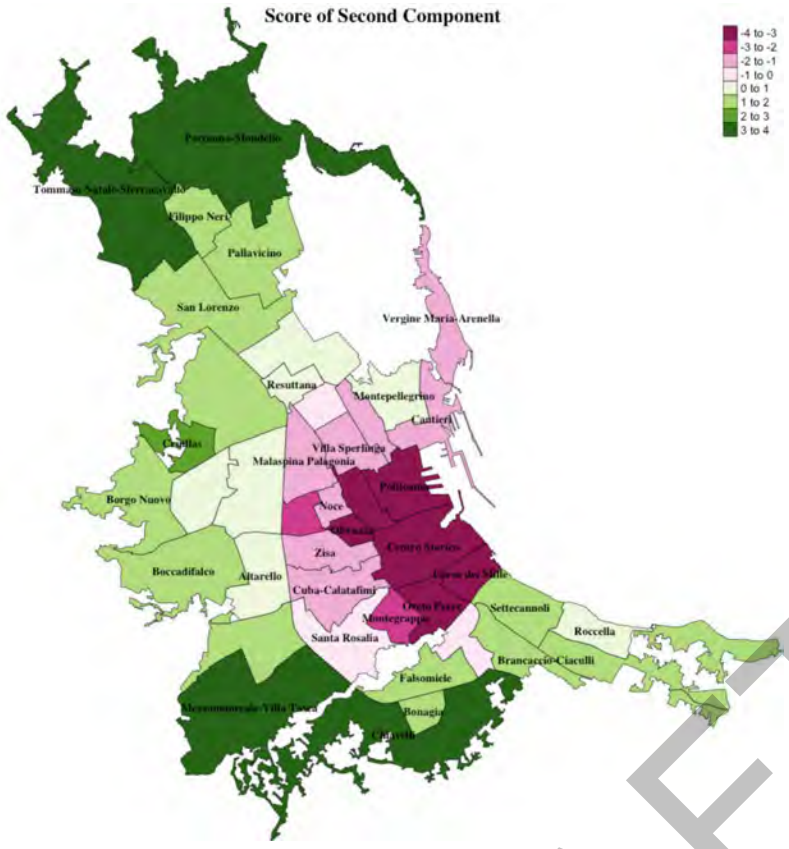



Figure 3: Score of the Second Component.

DRAFT

## References


[1] Boardman, B. (1991). *Fuel poverty: from cold homes to affordable warmth*. Pinter Pub Limited.

[2] Bouzarovski, S. (2014). Energy poverty in the european union: landscapes of vulnerability. *WIREs Energy and Environment*, 3(3):276–289.

[3] Bouzarovski, S. and Petrova, S. (2015). A global perspective on domestic energy deprivation: Overcoming the energy poverty–fuel poverty binary. *Energy Research & Social Science*, 10:31–40.

[4] Bouzarovski, S. and Simcock, N. (2017). Spatializing energy justice. *Energy Policy*, 107:640–648.

[5] Bouzarovski, S. and Thomson, H. (2018). Energy vulnerability in the grain of the city: Toward neighborhood typologies of material deprivation. *Annals of the American Association of Geographers*, 108(3):695–717.

[6] Hills, J. (2012). Getting the measure of fuel poverty: Final report of the fuel poverty review.

[7] Liddell, C. and Morris, C. (2010). Fuel poverty and human health: a review of recent evidence. *Energy policy*, 38(6):2987–2997.

[8] Pedro, P. and Pedro, G. J. (2022). Bringing energy poverty research into local practice: Exploring subnational scale analyses.

[9] Petrova, S. (2018). Encountering energy precarity: Geographies of fuel poverty among young adults in the uk. *Transactions of the Institute of British Geographers*, 43(1):17–30.

[10] Picone, M. (2016). Una segregazione paradossale e multi-scalare: il caso del quartiere zen di palermo. *MÃ©diterranÃ©e [Online]*, 127.

[11] Picone, M. (2021). Shifting imageries: Gentrification and the new touristic images of the inner city of palermo. In Banini, T., I. O., editor, *Representing Place and Territorial Identities in Europe*, pages 37–50. Springer.

[12] Robinson, C., Lindley, S., and Bouzarovski, S. (2019). The spatially varying components of vulnerability to energy poverty. *Annals of the American Association of Geographers*, 109(4):1188–1207.

[13] Walker, G. and Day, R. (2012). Fuel poverty as injustice: Integrating distribution, recognition and procedure in the struggle for affordable warmth. *Energy policy*, 49:69–75.

# A test for non-differential misclassification error in database epidemiological studies

Giorgio Limoncella[a], Leonardo Grilli[a], Emanuela Dreassi[a], Carla Rampichini[a], Robert Platt[b], and Rosa Gini[c]

[a]University of Florence, Florence, Italy;
giorgio.limoncella@unifi.it,leonardo.grilli@unifi.it,
emanuela.dreassi@unifi.it, carla.rampichini@unifi.it
[b]McGill University, Montreal, Canada; robert.platt@mcgill.ca
[b]ARS Toscana, Florence, Italy; rosa.gini@ars.toscana.it

DRAFT

**Abstract**

In epidemiological database studies, the true value of the outcome of interest is unknown, and it is measured with error through indicators implemented by appropriate algorithms. Differential misclassification of sensitivity of the study outcome across exposure groups may severely bias measures of association, in both directions. In this work, we introduce a bootstrap test for differential sensitivity and we conduct a simulation study to explore the properties of the method under multiple scenarios.

***Keywords:*** bootstrap test, measurement error, validity indices

## 1. Background

In epidemiology, when the outcome is measured with error, it is common practice to choose a very specific indicator (no false positives) and to rely on the assumption that sensitivity (SE) is non-differential across exposure strata. Unfortunately, estimating the sensitivity across exposure strata is usually impossible, and the assumption of non-differentiality is not tested (4). In safety studies, when an outcome is a suspected adverse effect of exposure, it is realistic to expect that exposed subjects may have their diagnoses recorded with diagnostic codes having higher precision. Differential sensitivity of the study outcome across exposure groups may severely bias measures of association in both directions.

## 2. Non-differentiality test

Several different indicators can potentially point to one single event (3). We exploited the interrelationships between validation indices of multiple indicators (1) of the same event to build a test for non-differentiality. The formulas are based on the observed prevalence of all indicators and their intersections, and on the values of their $PPV$. More details are included in Appendix. Brenner and Gefeller (2) showed that it is possible to adjust the risk ratio observed by an indicator $A$ as long as its $PPV$ is available across exposure strata and the sensitivity is non-differential, namely equal across exposure strata. Therefore, to implement the adjustment, it is essential to test the hypothesis $H_0 : SE_A^e = SE_A^{\bar{e}}$, which can be written as

$$H_0 : \frac{SE_A^e}{SE_A^{\bar{e}}} - 1 = 0 \tag{1}$$

To carry out the test, we exploit an auxiliary indicator $B$. Under the assumption of non-differentiality of the sensitvity of the union $SE_{A\cup B}$, the hypothesis (1) is equivalent to:

$$H_0 : \frac{P_A^e PPV_A^e (P_A^{\bar{e}} PPV_A^{\bar{e}} + P_B^{\bar{e}} PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} PPV_{A\cap B}^{\bar{e}})}{P_A^{\bar{e}} PPV_A^{\bar{e}} (P_A^e PPV_A^e + P_B^e PPV_B^e - P_{A\cap B}^e PPV_{A\cap B}^e)} - 1 = 0 \tag{2}$$

The test statistic is calculated by replacing the values of $PPV$ in equation (2) with their estimates obtained from validation studies. The validation sample can be extracted by a stratified sampling across exposure groups and within each exposure group stratified by indicator (in our simulation study: 50% of the sample from $A = 1$ and 50% from $B = 1$, for the exposed group, and the same for the non-exposed group). The statistical test is carried out with the bootstrap method.

## 3. Simulation study

To evaluate the performance of the proposed non-differentiality test, we developed a simulation study. We generated one binary vector representing the true outcome of interest $Y$ and one representing the exposure group $E$. Different scenarios were explored, letting $\pi^{\bar{e}}$ ($Y^{\bar{e}}/N^{\bar{e}}$) vary between 1%, 5% and 10% and varying the relative risk, $RR$, between 1.2 and 2. The proportion of exposed was fixed at 5% in every simulation. Then, two different indicators of Y, A and B, were generated, misclassified with specific error rates for each exposure group. $SE_{A\cup B}$ was fixed to 0.8 in both exposure groups, in order to comply with the assumption of non-differentiality of $SE_{A\cup B}$. In each scenario, we defined the value of $SE_A$ and of the specificities, $SP$, while the other validity indices were derived. The sensitivity of $A$ in the unexposed group was set at 0.50, while the sensitivity in the exposed group varied between 0.3 and 0.7. Thus, the sensitivity ratio ranged from 0.6 to 1.4. Only a non-intersection scenario was studied, in which no subject tested positive for both indicators. $SE_{A\cap B}$ was set to 0 and $SP_{A\cap B}$ was set to 1. So, the formula (2) become:

$$H_0 : \frac{P_A^e PPV_A^e (P_A^{\bar{e}} PPV_A^{\bar{e}} + P_B^{\bar{e}} PPV_B^{\bar{e}})}{P_A^{\bar{e}} PPV_A^{\bar{e}} (P_A^e PPV_A^e + P_B^e PPV_B^e)} - 1 = 0 \tag{3}$$

Similar scenarios are plausible, as shown in a case study concerning angioedema disease (5). The specificity was set at $1 - \frac{\pi^{\bar{e}}}{10}$ for $A$ and $1 - \pi^{\bar{e}}$ for $B$ in both exposure groups, and we assumed that $P(A = 1|Y = 0, E = e)$ is independent from $P(B = 1|Y = 0, E = e)$, so $1 - SP_{A\cap B} = 1 - SP_A \times 1 - SP_B$. Once the data had been generated, the validation and the test were performed 1000 times for each scenario using three different sample sizes (200, 400 and 600, of which 50% from $A$ and 50% from $B$) and the proportion of times the test leads to rejecting the null hypothesis, $P_{rej}$, was calculated.

DRAFT

## 3.1 Results

The PPVs of indicators ranged between 78% to 95%, for $A$, and between 11% to 33% for $B$, depending on the scenario. When the sample size was equal to 200 or 400, the power of the test was greater than 80% only in the extreme values of the sensitivity ratio (0.6 and 1.4), irrespective of risk ratio and true prevalence. For a sample size of 600, the power of the test was greater than 80% and for the sensitivity ratio equal to 0.8 and 1.2, but only if the risk ratio was equal to 2 or the prevalence was greater than 0.01. Figure 1 shows how the proportion of times in which the null hypothesis is rejected ($P_{rej}$) varies according to the true sensitivity ratio, for the scenario in which $RR = 2$, and $\pi^{\bar{e}} = 0.05$.

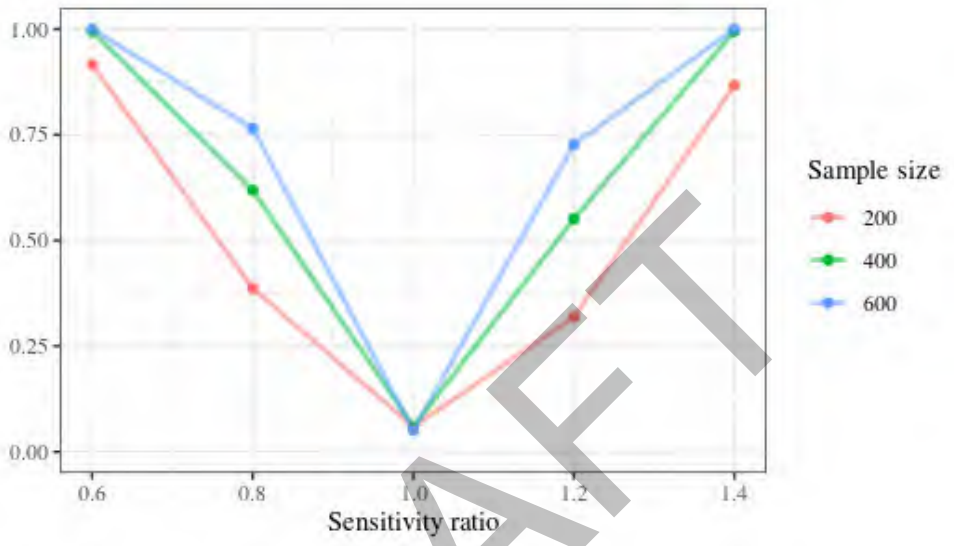



Figure 1: Proportion of times in which the null hypothesis is rejected, $P_{rej}$. In scenarios where the sensitivity ratio is different from 1, $P_{rej}$ corresponds to the power of the test.

DRAFT

**Acknowledgments** This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking ConcePTION grant n. 821520.

## Appendix

Prevalence can be defined as $\pi = \frac{TP+FN}{N}$, where N represents the numerosity of the study population. It can therefore be rewritten as a function of observed prevalence ($P = \frac{TP+FP}{N}$), positive predictive value (PPV), and sensitivity (SE):

$$\begin{aligned} \pi &= \frac{TP+FN}{N} \\ &= \frac{TP+FN}{N} \times \frac{TP+FP}{TP} \times \frac{TP}{TP+FP} \\ &= \frac{TP+FP}{N} \times \frac{TP+FN}{TP} \times \frac{TP}{TP+FP} \\ &= P \times \frac{1}{SE} \times PPV \end{aligned} \tag{4}$$

therefore:

$$SE = \frac{P \times PPV}{\pi} \tag{5}$$

$$\pi = \frac{P \times PPV}{SE} \tag{6}$$

In order to perform the non-differentiality test, the null hypothesis $H_0 : SE_A^E = SE_A^{\bar{E}}$ can be rewritten as:

$$\frac{SE_A^E}{SE_A^{\bar{E}}} - 1 = 0 \tag{7}$$

and using equation (5):

$$SE_A = \frac{P_A \times PPV_A}{\pi} \tag{8}$$

then, using equation (6):

$$SE_A = \frac{P_A \times PPV_A}{\frac{P_A \times PPV_A}{SE_{AUB}} + \frac{P_B \times PPV_B}{SE_{AUB}} - \frac{P_{A\cap B} \times PPV_{A\cap B}}{SE_{AUB}}} \tag{9}$$

thus, the test statistic can be rewritten as:

$$t = \frac{\frac{P_A^E \times PPV_A^E}{P_A^E \times PPV_A^E + P_B^E \times PPV_B^E - P_{A\cap B}^E \times PPV_{A\cap B}^E}}{\frac{P_A^{\bar{E}} \times PPV_A^{\bar{E}}}{P_A^{\bar{E}} \times PPV_A^{\bar{E}} + P_B^{\bar{E}} \times PPV_B^{\bar{E}} - P_{A\cap B}^{\bar{E}} \times PPV_{A\cap B}^{\bar{E}}}} \times \frac{SE_{AUB}^E}{SE_{AUB}^{\bar{E}}} - 1 \tag{10}$$

finally, if $SE_{AUB}^E = SE_{AUB}^{\bar{E}}$:

$$t = \frac{P_A^E PPV_A^E (P_A^{\bar{E}} PPV_A^{\bar{E}} + P_B^{\bar{E}} PPV_B^{\bar{E}} - P_{A\cap B}^{\bar{E}} PPV_{A\cap B}^{\bar{E}})}{P_A^{\bar{E}} PPV_A^{\bar{E}} (P_A^E PPV_A^E + P_B^E PPV_B^E - P_{A\cap B}^E PPV_{A\cap B}^E)} - 1 \tag{11}$$

## References


[1] Bollaerts, K., Rekkas, A., De Smedt, T., Dodd, C., Andrews, N., Gini, R.: Disease misclassification in electronic healthcare database studies: Deriving validity indices-a contribution from the advance project (2020). PLoS ONE. doi 10.1371/journal.pone.0231333

[2] Brenner, H., Gefeller, O.: Use of positive predictive value to correct for disease misclassification in epidemiologic studies. American journal of epidemiology. **138**, 1007–1015 (1994)

[3] Gini, R., Dodd, C., Bollaerts, K., Bartolini, C., Roberto, G., Huerta, C., Martín-Merino, E., Duarte-Salles, T., Picelli, G., Tramontan, L., Danieli, G., Correa, A., McGee, C., Becker, B., Switzer, C., Gandhi-Banga, S., Bauwens, J., Maas, N., Spiteri, G., Sturkenboom, M.: Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. Vaccine. (2019) doi 10.1016/j.vaccine.2019.07.045

[4] Lanes S, Beachler DC. Validation to Correct for Outcome Misclassification Bias. Pharmacoepidemiology and Drug Safety (2023). Available from: doi doi/abs/10.1002/pds.5601

[5] Limoncella, G., Girardi, A., Hyeraci, G., Bartolini, C., Roberto, G., Heintjes, E., De Jong, H., Kuiper, J., Beckmeyer-Borowko, A., Behr, S., Sturkenboom, M., Gini, R. : Validation to estimate sensitivity along with positive predictive value: A case study. Abstract 956, page 440 in: ABSTRACTS of ICPE 2022, the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management (ICPE), Copenhagen, Denmark, 26-28 August, 2022. Pharmacoepidemiology and Drug Safety. 2022;31(S2):3-628. https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.5518


DRAFT

# Is the COVID-19 'color code' of Italian regions subjected to political manipulation?

Giovanni Busetta [a] and Fabio Fiorillo [b]

[a] Department of Economics, University of Messina, Messina, Italy; gbusetta@unime.it
[b] Department of Economic and Social Sciences (DiSES), Marche Polytechnic University; f.fiorillo@staff.univpm.it

**Abstract**

While the first wave of the COVID-19 pandemic saw the Italian government adopt a strategy of strict containment measures (lockdown) to stop the spread of the virus, the second one saw it apply a regional division into colored zones with differentiated restrictions, depending on the specific local pandemic risk. The aim of this paper is to analyze whether in addition to health reasons, some political motivations, such as the will of increasing mobility inflows and outflows, could affect pandemic restrictions based on regions' colors. The idea underlying the paper is that, if no political reasons affect the decision of applying restrictions, the probability for a certain region to obtain a particular color shall depend only on factors concerning public health and included by law as factors influencing the color of the region. By applying a probit model to a regional panel, some clues of political manipulation emerge.

**Keywords:** Regional classification of pandemic risk, airport passengers' flows, political manipulation.

DRAFT

## 1. Introduction

Looking at the Coronavirus Source Data by Our World in Data, Italy was among the countries with the highest mortality rate, with 1226.6 deaths per million of inhabitants (DpM). Notwithstanding the widespread average diffusion of COVID-19 in the country, Italian regions experienced very different levels of contagion, especially during the first wave of pandemic, and thus different mortality rates (De Leo and Araújo 2021). Indeed, the pandemic has affected Italian regions differently because the capacity of the various regional hospital systems is also very different.

Citizens and regional governments did not want to further worsen the economic impact of pandemic. Hence, as of the end of November 2020, central government relaxed lockdown measures, defining a regional classification of pandemic risk. Such a classification was based on parameters related to the diffusion of the virus and the capacity of the hospital system to accept new patients. Based on the results achieved on these parameters, each Italian region was identified by a color which determined the level of restriction to mobility (De Leo and Araújo 2021).

The aim of this paper is to analyze whether the wish for increased people mobility could affect the changes in the color-based risk classification of the regions, in addition to health reasons defined by law. The idea underlying the paper is that, if no political reasons affect the decision of applying

restrictions, the likelihood of a certain region to obtain a particular color should depend only upon the factors influencing the color of the region by law. Restrictions depend on the levels reached by three main health indicators in the region. These indicators are the ratio of hospitalized over positive people (HOSP), COVID-19 occupancy of beds (INTENS) in intensive care units (hereafter ICUs) and the reproduction number (hereafter RT).

Our idea is in line with the results of other studies (Tassinari et al. 2021), in which the authors found a significant link between the cooperative pressure exerted by business organizations against the containment measures and the reduction of the measures themselves. This result undoubtedly indicates a significant role of the 'business world' in shaping the responses to the pandemic by regional authorities. The reason underlying the inclusion of these three parameters as identifiers of the regional COVID-19 danger is awareness of the fact that mortality is not the only problem to face.

On the one side, the pandemic crisis challenged the sustainability of health care system. The risk of the ICU in Italy collapsing was the main concern in facing the pandemic. The presence and the number of such wards is particularly relevant because COVID-19 often induces infected people not to breathe on their own. On the other side, the economic consequences of COVID-19 have been huge. In response to the pandemic crisis, governments have called for policies with unprecedented restrictions on mobility and activities (OECD, 2020a, 2020b). These policies, which are characterized by enormous, short, and long-term economic and social costs, have generated the worst economic crisis since the Second World War. The considerations of Hebert and Curry (2022) may explain why central government changed its strategy after the first wave of pandemic passed and vaccines started to be distributed among population. Indeed, to maintain the levels of contagions sufficiently low for the pandemic not to explode, central government abandoned the strict lockdown approach and defined a 'color based' risk classification to be applied to regions.

Even though the pandemic shock simultaneously affects all countries (Caragliu, 2022), thereby decreasing the trust of citizens, it is worth to note that the mortality rate (De Leo and Araújo 2021), the economic impact and the long run effect of such shock has been uneven among regions, as investigated by Caragliu and Capello (2021). Indeed, the incidence of the virus is affected by differences in socio-economic variables (González-Val and Sanz-Gracia, 2022). As the mortality rate was uneven distributed and the pressure to reduce the economic effect was strong, central government accepted to apply differential rules to regions depending on three indicators particularly relevant in influencing spread and diffusion of the pandemic. The Italian government adopted three sets of restrictions (yellow for low levels of restriction, orange for intermediate levels and red for high levels of restrictions) to be imposed on a regional basis.

It is likely that, in a multi-governmental system, the incentive of central and regional government's incentives could not be aligned. Regional government may have stronger incentives to reduce restrictions to make the economy restart. Following Hebert and Curry (2022) weak lockdown policies are often correlated to minimal levels of tourism, stable population, and high vaccination rates. As Koyama (2021) states, policymakers face incentive and information problems, therefore we may formulate the following hypotheses.

**Hypothesis 1.** Regional governments have some incentive of exploiting their information advantages so as to manipulate regional health indicators and display a better color risk code and have lower restrictions.

The main contribution of our paper is to test such hypothesis by verifying the assumption of independence of regional restrictions (colors) during COVID-19 pandemic from political reasons. Our null hypothesis is the independence of worsening regional color risk code due to the political need of promoting the economy through avoiding restrictions. We used the number of passengers that have passed through regional airports one month before as a proxy of such a need. Indeed, the greatest the number of passengers passing through regional airports, the highest the costs of mobility restrictions for the entire regional economy. If the null hypothesis is true, the only parameters explaining the

DRAFT

probability of worsening the color risk code will be the three parameters used by government, the percentage of people positive to COVID-19 in the region, the color of the region a week before and regional dummies as control variables. The information of such parameters should provide all the information required to determine whether the risk of contagion would increase, and thus the future restrictions to be applied in the region. Eventually, the number of passengers could have an effect in increasing the spread of virus, worsening the risk code. On the contrary, the null hypothesis is rejected in case the impact of the number of passengers a month before significantly reduce the probability of the color code worsening. We interpret such results as a clue that political motivation matters in exploiting all the possible informative advantages when regions provide data for risk codification.

## 2. Data and methods

We used a balanced panel which contains daily data covering the period from 23 November 2020 to 18 January 2022. Such period was divided in two sub periods: the first one covers the time before the introduction of the green pass (from 23 November 2020 to 5 August 2021), while the second one does cover the period after it was introduced (from 6 August 2021 to 18 January 2022).

Since we used daily, weekly, and monthly data, first we presented a specific description for each variable (Table 1). In Table 2, descriptive statistics for all the time spans and the two sub periods are presented.

Table 1: variables description

| Variable | Description | Periodicity | Source |
|---|---|---|---|
| worsening | Dependent variable. Dummy equal to 1 if the color risk code worsens with respect to the code of a week before | Daily | Authors' elaboration on Italian Ministry of Health's decrees |
| orange | Dummy equal to 1 if the color risk code is red, intermediate restrictions | Daily | Authors' elaboration on Italian Ministry of Health's decrees |
| yellow | Dummy equal to 1 if the color risk code is yellow, low restrictions | Daily | Authors' elaboration on Italian Ministry of Health's decrees |
| white | Dummy equal to 1 if the color risk code is yellow, no restrictions | Daily | Authors' elaboration on Italian Ministry of Health's decrees |
| RT | Effective reproduction number | Weekly | Health Ministry, ISS |
| INTENS | % of occupied beds in ICUs | Weekly | Health Ministry, ISS |
| HOSP | Ratio of hospitalized over positive people | Weekly | Health Ministry, ISS |
| positive_PC | Share of positive to the COVID-19 on regional population | Daily | Health Ministry and Gimbe foundation |
| K_passengers | Thousands of passengers | Monthly | Assaeroporti |
| GP | Dummy equal to 1 after the introduction of the green pass, 6 August 2021. | Daily | Authors'elaboration on Italian Ministry of Health's decrees |

DRAFT

If the three indicators (RT, HOSP, INTENS) of a week before, exhibit bad values, the color risk code is worsened. This is the reason why we used weekly data for such indicators. On the contrary, the worsening of the risk code is measured daily, since for specific periods (Christmas and Easter days) the regional color risk code was not applied, and the color was set to orange and red for all the regions regardless of the indicators. The source (Assaeroporti: Italian Airport Association) provides the monthly number of passengers of all Italian airport, which we aggregated on regional basis. Having monthly data permits us to consider the effect of previous inflows on color code. Such data may depend on both regional and airport location, morever it should be influenced by the pandemic restrictions. To avoid endogeneity problems, we considered data on passengers lagged by one month and all other variables lagged by a week. Table 2 presents descriptive statistics.

Table 2: descriptive statistics, from 23 November 2020 to 18 January 2022

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| worsening | 8.855 | 0.0902315 | 0.2865293 | 0 | 1 |
| red | 8.855 | 0.079616 | 0.2707131 | 0 | 1 |
| orange | 8.855 | 0.1586675 | 0.3653863 | 0 | 1 |
| yellow | 8.855 | 0.2866177 | 0.4522069 | 0 | 1 |
| white | 8.855 | 0.4750988 | 0.4994077 | 0 | 1 |
| RT | 8.855 | 1.00349 | 0.3027737 | 0 | 2.44 |
| INTENS | 8.841 | 0.1614184 | 0.1492893 | 0 | 0.67 |
| HOSP | 8.841 | 0.1939815 | 0.1658585 | 0 | 0.8 |
| positive_PC | 8.855 | 0.0058458 | 0.0070748 | 7.77E-05 | 0.0728643 |
| K_passengers | 8.855 | 286.1524 | 530.7765 | 0 | 2942368 |
| GP | 8.855 | 0.3928854 | 0.4884193 | 0 | 1 |

We estimated the probability of worsening the color risk code using a linear probability model, a probit and a logit estimation. The use of three different models provides a robustness check. For comparisons, we present only the marginal effect[1].

## 2. Results

Table 3 presents our results. The sign of the parameters relative to the three indicators and to the share of individuals positive to COVID-19 shows a positive coefficient as expected. Moreover, the null hypothesis of independence of the worsening of the color risk code from the number of passengers a month before is rejected with a negative sign. Indeed, the probability of the restrictive color of the region worsening decreases with the number of passengers a month before in all the performed analyses. This result must necessarily imply either that an increase in the passengers landed in the region reduces the probability of being infected, which is epidemiological very unlikely, or that political reasons influenced the decision of not applying stronger restrictions to a particular region.

Table 3: Linear probability model, logit and probit predictions of increasing pandemic risk.

| | (1) | (2) | (3) |
|---|---|---|---|
| VARIABLES | lin_prob | probit | logit |
| l7.orange | 0.197*** | 0.587 | 0.888 |
| | (0.0116) | (14.40) | (26.01) |
| l7.yellow | 0.341*** | 0.666 | 0.973 |
| | (0.0123) | (14.40) | (26.01) |

1 The estimations of probit and logit coefficients are in appendix (Table A1 and A2).

DRAFT

| | | | |
|---|---|---|---|
| l7.white | 0.440*** | 0.741 | 1.049 |
| | (0.0165) | (14.40) | (26.01) |
| l7.RT | 0.0498*** | 0.0593*** | 0.0620*** |
| | (0.00911) | (0.00888) | (0.00882) |
| l7.INTENS | 0.357*** | 0.362*** | 0.386*** |
| | (0.0525) | (0.0468) | (0.0477) |
| l7.HOSP | 0.656*** | 0.364*** | 0.345*** |
| | (0.0486) | (0.0392) | (0.0397) |
| l7.positive_PC | 3.027*** | 2.026*** | 2.140*** |
| | (0.682) | (0.509) | (0.497) |
| l32.K_passengers | -2.14e-05** | -2.31e-05** | -2.32e-05** |
| | (8.99e-06) | (9.89e-06) | (1.04e-05) |
| GP | -0.0647*** | -0.0483*** | -0.0528*** |
| | (0.00797) | (0.00944) | (0.00998) |
| Constant | -0.474*** | | |
| | (0.0170) | | |
| Observations | 8,176 | 8,176 | 8,176 |
| R-squared | 0.178 | | |
| Regions FE | YES | YES | YES |

Standard errors in parentheses *** p<0.01, **p<0.05, * p<0.1

## 2. Concluding remarks

Results of the analysis show that the increase in the restrictions applied to the regions during COVID-19 pandemic is explained not only by health indicators, but also by economic variables such as people mobility. The relation between this mobility in the previous month and the probability of worsening of the color risk code is not positive as can be expected from epidemiological considerations. On the contrary, it is negative and significant. This result led us to think of some kind of political manipulation, as restrictions applied to regions may impact on inflows and outflows of people, and thus on the general economic development of the region in that period.

Hence, our results provide some strong clues of manipulation mechanisms at work. We suspect that regions would delay the transmission of information not to worsen their color risk code. This way, regions gained an additional week with few restrictions on mobility. Such behavior is not punishable since a delay in transmitting information can depend on many legitimate causes . Even if our suspicion seems likely, further investigation should be conducted to describe the manipulation mechanism in detail.

DRAFT

## References


[1] Capello, R., Caragliu, A.: (2021). Regional growth and disparities in a post-COVID Europe: A new normality scenario. Journal of Regional Science, 61(4), 710– 727. https://doi.org/10.1111/jors.12542

[2] Caragliu, A. (2022). Better together: Untapped potentials in Central Europe. Papers in Regional Science, 101(5), 1051–1085. https://doi.org/10.1111/pirs.12690

[3] De Leo, S., & Araújo, M. P. (2021). A modelling study across the Italian regions: lockdown, testing strategy, colored zones, and skew-normal distributions. How a numerical index of pandemic criticality could be useful in tackling the CoViD-19. arXiv preprint arXiv:2102.03373.

[4] González-Val, R., & Sanz-Gracia, F. (2022). Urbanization and COVID-19 incidence: A cross-country investigation. Papers in Regional Science, 101( 2), 399– 415. https://doi.org/10.1111/pirs.12647

[5] Hebert, D.J., Curry, M.D. Optimal lockdowns. Public Choice (2022). https://doi.org/10.1007/s11127-022-00992-4

[6] Koyama, M. Epidemic disease and the state: Is there a tradeoff between public health and liberty?. Public Choice (2021). https://doi.org/10.1007/s11127-021-00944-4

[7] Milligan GN, Barrett AD (2015). Vaccinology: an essential guide. Chichester, West Sussex: Wiley Blackwell. p. 310. ISBN 978-1-118-63652-7. OCLC 881386962.

[8] OECD (2020a), COVID-19 and fiscal relations across levels of government, Paris, OECD Publishing.

[9] OECD (2020b), The territorial impact of COVID-19: Managing the crisis across levels of government, Paris, OECD Publishing.


DRAFT

# Modelling multilevel ordinal response under endogeneity: application to DTC patients' outcome.

Silvia D'Elia[a]

[a] Sapienza Università di Roma, silvia.delia@uniroma1.it

**Abstract**

Ordinal data are widely used in medicine as they help to summarise health states or responses to treatment. Cumulative link models are useful because they explicitly consider the order in response's categories, and they allow to model the probability of observing a value that does not exceed a specific category. In this contribution, two situations that often arise in real data analysis will be investigated: i) data with hierarchical structure; ii) presence of endogenous covariates. We describe this framework by discussing the application to the response to treatment at 3 years in patients with differentiated thyroid cancer (DTC) diagnosis. The application aims to evaluate potential predictors for the response at 3 years considering both repeated measurements from the same patient, and data from several clinical centres. The analysis of the response over time is of interest as it evaluates whether the improvement can be attributed to the time that allows to obtain clearer eco images. Consequently, it could lead to a reduction in the use of radioactive iodine therapy, which is essential both to limit the side effects it produces on patients and to reduce the production of waste that are difficult to dispose of.

**Keywords:** Ordinal data, Cumulative link models, Multilevel structure, Endogenous covariates.

DRAFT

## 1. Introduction

Ordinal data are widely used in health and social science, where ordinal scales are used to measure response to a treatment or summarise the severity of a disease. Ordinal data are characterized by the fact that they can assume values in a discrete set of ordered categories with no assumption on distance between adjacent categories. Cumulative link models are a powerful models' class since they allow to treat the observations as categorical and explicitly consider the order in response's categories.

The structure of this paper follows. In section 2 we review basic regression models for ordinal data with a focus on random effects and endogenous covariates. In section 3 we introduce the relevant application; the results are briefly described in section 4.

## 2. Regression models for ordinal data

Let us start considering a response variable $Y_i$ which is measured on an ordinal scale taking values in the ordered set $\{1, \dots, J\}$. Cumulative link models are widely used to analyse ordinal response variable as they allow to model the cumulative probability $\gamma_{ij} = P(Y_i \leq j)$[1] as

$$\gamma_{ij} = h(\eta_{ij}), \qquad \eta_{ij} = \theta_j - \boldsymbol{x}_i'\boldsymbol{\beta} \quad i = 1, \dots, n \ , j = 1, \dots, J-1 \tag{1}$$

When we adopt a so-called underlying latent variable framework, the response function $h$ in equation (1), is usually a cumulative distribution function of an unobserved, continuous, latent variable. This choice allows to explicitly consider the order in response categories, by using category-specific thresholds $\theta_j$ and a common parameter vector $\boldsymbol{\beta}$. In the following we will assume that the underlying latent

variable is Gaussian and, therefore, adopt the Gaussian cdf as the response function, leading to the ordinal probit model. To be more specific, the observed categorical variable $Y_i$ is obtained by discretizing a latent continuous variable $Y_i^*$ according to the following scheme:

$$Y_i = j \leftrightarrow \theta_{j-1} < Y_i^* \leq \theta_j \tag{2}$$

where $-\infty = \theta_1 \leq \theta_2 \leq \cdots \leq \theta_J = +\infty$ are non-decreasing category-specific thresholds on the continuous scale, when latent variable falls in the *j*-th interval the observed variable is equal to *j* [2]. The latent continuous variable $Y_i^*$ is described via a standard regression model where error terms are independent and identically distributed Gaussian random variables:

$$Y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1) \tag{3}$$

Given the previous hypotheses, the probability of $Y_i$ not exceeding the *j*-th category can be written as in equation (4). Therefore, fixed a category specific threshold $\theta_j$, a positive value of $\beta$ coefficient means that the probability of not exceeding the *j*-th category decreasing.

$$P(Y_i \leq j) = P(Y_i^* \leq \theta_j) = P(\boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \leq \theta_j) = P(\varepsilon_i \leq \theta_j - \boldsymbol{x}_i'\boldsymbol{\beta}) = \Phi(\theta_j - \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{4}$$

The probability associated with a specific category *j* can in fact be written as the difference between two consecutive values of the cumulative distribution function of a standard normal random variable.

$$P(Y_i = j) = \Phi(\theta_j - \boldsymbol{x}_i'\beta) - \Phi(\theta_{j-1} - \boldsymbol{x}_i'\beta), \qquad j = 1, \dots, J \tag{5}$$

For the last category probability, the following equality holds:

$$P(Y_i = J) = P(Y_i^* \leq \theta_J) - P(Y_i^* \leq \theta_{J-1}) = 1 - \Phi(\theta_{J-1} - \boldsymbol{x}_i'\beta) \tag{6}$$

The likelihood function, given the sample, is defined, according to Wooldridge (2010) [3], as

$$L(\theta, \beta) = \prod_{i=1}^{n}\prod_{j=1}^{J} P(Y_i = j)^{Y_{ij}} \tag{7}$$

The parameters can be estimated by maximizing the corresponding log – likelihood function that can be written as follow:

$$\begin{aligned} l(\theta, \beta) = \sum_{i=1}^{n} \Bigg\{ & 1[Y_i = 1] \log[\Phi(\theta_1 - \boldsymbol{x}_i'\beta)] + \sum_{j=2}^{J-1} 1[Y_i = j] \log\left[\Phi(\theta_j - \boldsymbol{x}_i'\beta) - \Phi(\theta_{j-1} - \boldsymbol{x}_i'\beta)\right] \\ & + 1[Y_i = J] \log\left[1 - \Phi(\theta_{J-1} - \boldsymbol{x}_i'\beta)\right] \Bigg\} \end{aligned} \tag{8}$$

### 2.1 Random effects in ordinal regression models

When data has a hierarchical structure, it is possible to consider the information within the model specification. A random effects ordinal model may be proposed for the analysis of clustered ordinal data to account for potential dependence of responses recorded from the same higher-level unit.

In this case, the model in equation (3) could be modified to account for unobserved, time-constant, features of the higher-level units by considering the model below.

$$\begin{aligned} & \gamma_{ijk} = h(\eta_{ijk}), \quad \eta_{ijk} = \theta_j - \boldsymbol{x}_i'\boldsymbol{\beta} - \alpha_k \\ & \alpha_k \sim N(0, \sigma_\alpha^2),\ \varepsilon_i \sim N(0,1), \qquad \alpha_k \perp \varepsilon_i \\ & i = 1, \dots, n;\ j = 1, \dots, J;\ k = 1, \dots, K. \end{aligned} \tag{9}$$

DRAFT

We must consider that random effects account for unobserved, time-constant, features of the higher-level units. As they may be correlated with observed covariates, according to Nehuaus and Kalbfleish (1998) [4] and Neuhaus and McCulloch (2006) [5], we need consider also higher-level specific averages, obtaining the following model structure:

$$\begin{gathered}\gamma_{ijk} = h(\eta_{ijk}), \qquad \eta_{ijk} = \theta_j - \boldsymbol{x}_i'\boldsymbol{\beta} - \bar{x}_k'\boldsymbol{\gamma} - \alpha_k, \\ \alpha_k \sim N(0, \sigma_\alpha^2),\ \varepsilon_i \sim N(0,1), \qquad \alpha_k \perp \varepsilon_i \\ i = 1, \dots, n, j = 1, \dots, J, k = 1, \dots, K\end{gathered} \tag{10}$$

For sake of simplicity, we did not consider a random effects approach, but we used the linear predictor above and employed a robust estimator for the standard errors of model parameters.

### 2.2 Endogenous variables in ordinal regression models

A complication in the model is given by potential endogenous covariates. The model in equation (3) restricts all the covariates to be independent of unobserved heterogeneity and error terms. Some observed covariates may be endogenous in the sense that they are correlated with the error terms. It is possible to formulate the model for the response variable $Y_i^*$ considering exogenous covariates and an ordinal covariate $w_i$.

$$Y_i^* = \boldsymbol{x}_i'\beta_1 + w_i\beta_{e1} + \varepsilon_i \tag{11}$$

The probability in equation (4) can be written as:

$$\begin{gathered}P(Y_i \le j) = P(Y_i^* \le \theta_j) = P(\boldsymbol{x}_i'\beta + w_i\beta_{e1} + \varepsilon_i \le \theta_j) = \\ = P(\varepsilon_i \le \theta_j - \boldsymbol{x}_i'\beta - w_i\beta_{e1}) = \Phi(\theta_j - \boldsymbol{x}_i'\beta - w_i\beta_{e1})\end{gathered} \tag{12}$$

The endogenous covariate $w_i$ in the model (11) is assumed to be correlated with the error term $\varepsilon_i$ as, for example, both response and the observed covariates are produced by a (partially) common data generation process. In order to account for such data generation process, a model for the endogenous covariate taking values in the ordered set $\{1, \dots, G\}$. is fitted as follow:

$$w_i^* = \boldsymbol{z}_i'\alpha + \eta_i \tag{13}$$

The errors $\eta_i$ have a standard normal distribution.

Given the assumption in formula (2), the cumulative probability for the observed endogenous variable is:

$$P(w_i \le g) = P(w_i^* \le \theta_g) = P(\boldsymbol{z}_i'\alpha + \eta_i \le \theta_g) = P(\eta_i \le \theta_g - \boldsymbol{z}_i'\alpha) = \Phi(\theta_g - \boldsymbol{z}_i'\alpha) \tag{14}$$

To account for the dependence between the errors in equation (11) e (13) we assume that they are bivariate Gaussian random variables:

$$\begin{pmatrix}\varepsilon_i \\ \eta_i\end{pmatrix} \sim MVN(0, \Sigma), \qquad \Sigma = \begin{pmatrix}1 & \rho_{\varepsilon,\eta} \\ \rho_{\varepsilon,\eta} & 1\end{pmatrix} \tag{15}$$

The likelihood function, according to Winship and Mare (1984) [6], can be written as:

$$L(\theta, \beta, \Sigma) = \prod_{i=1}^{n}\prod_{j=1}^{J}\prod_{g=1}^{G} P(Y_i = j, w_i = g)^{d_{ijg}} \tag{16}$$

where $d_{ijg}$ is an indicator for $Y_i = j$ and $w_i = g$. The corresponding conditional probabilities can be written as follow:

DRAFT

$$P(Y_i = j|\boldsymbol{x}_i, \boldsymbol{z}_i, w_i) = \frac{\Phi(\theta_j - \boldsymbol{x}_i'\beta - w_i\beta_{e1}) - \Phi(\theta_{j-1} - \boldsymbol{x}_i'\beta - w_i\beta_{e1})}{\Phi(\theta_g - \boldsymbol{z}_i'\alpha) - \Phi(\theta_{g-1} - \boldsymbol{z}_i'\alpha)} \quad (17)$$

## 3. Application: Italian Thyroid Cancer Observatory

We used the approach described so far to analyse data from the Italian Thyroid Cancer Observatory (ITCO). ITCO, founded in 2013, is a database that collects data about patients with diagnosis of thyroid cancer, treated in different clinical centres in Italy. The thyroid cancer is associated with a high survival rate with mortality less than 1x100.000. The concern for this pathology is due to the considerable prevalence of disease relapse. For this reason, long-term management through follow-up protocols is necessary. ITCO has the aim to evaluate the response to treatment over time, enrolling patients from 54 clinical centres and following each of them periodically after the initial surgical treatment.

The analysis of the response to treatment over time is important to evaluate if there are conditions that define a change in the patient's clinical status among those recorded at each follow-up, such as the residual tumoral tissue, or at baseline such as neck dissection, initial treatment, or assessment of the risk level. The analysis includes only patients that did not undergo further treatments as this allows to check whether changes in the response is associated to information available at baseline or at each follow up. Alternatively, changes may be just due to time passing by that allow to obtain clearer and more definite eco imaging.

The response to treatment is measured on an ordinal scale based on information obtained during the follow up. When compared to other staging systems, that include only primary characteristic of the tumour, the response to treatment we analysed is a dynamic one; it is based on clinical advice, thyroglobulin levels and eco imaging. Given such information, the status of the disease is classified into 4 classes in non-decreasing order of desirability as excellent (ER, no evidence of disease), indeterminate (IND), biochemical incomplete (BIR) and structural incomplete (SIR, evidence of disease).

## 4. Real world analysis of the response to treatment at 3 years.

Table 1: Distribution of the response at 3 years by response at 12 months

| **Response** | | *3 years* | | | | |
|---|---|---|---|---|---|---|
| | | SIR | BIR | IND | ER | TOTAL |
| *12 months* | SIR | 16 (34%) | 6 (12.8%) | 14 (29.8%) | 11 (23.4%) | 47 |
| | BIR | 10 (8.4%) | 37 (31.1%) | 51 (42.9%) | 21 (17.6%) | 119 |
| | IND | 18 (1.9%) | 33 (3.6%) | 434 (46.7%) | 444 (47.8%) | 929 |
| | ER | 2 (0.2%) | 17 (1.5%) | 202 (18%) | 904 (80.4%) | 1125 |

Table 1 shows transitions between the response at two, pre-specified, different timepoints 12 months and 3 years since initial treatment. It is possible to observe some changes with direction from the lower categories (SIR, BIR, IND) at 12 months towards the higher ER category at 3 years. On the main diagonal there are percentage of patients that do not register changes in their disease status after (approximately) 24 months. The percentage are calculated over row totals.

The response to treatment can assume 4 values $j = \{SIR, BIR, IND, ER\}$. According to the literature [7], the response to treatment at 12 months ($rtt_{12m}$) is analysed conditionally to three covariates, that represent different patient's baseline information: level of risk as defined by the American Thyroid Association ($z_1$), the type of initial treatment ($z_2$), and the possible neck dissection ($z_3$). We defined the following ordered probit model:

$$P(rtt_{12m_i} \leq g) = P(rtt^*_{12m_i} \leq \theta_g) = P(z_{1i}\alpha_1 + z_{2i}\alpha_2 + z_{3i}\alpha_3 + \eta_i \leq \theta_g) = \\ = P(\eta_i \leq \theta_{gj} - z_{1i}\alpha_1 - z_{2i}\alpha_2 - z_{3i}\alpha_3) = \Phi(\theta_g - z_{1i}\alpha_1 - z_{2i}\alpha_2 - z_{3i}\alpha_3) \quad (18)$$

Then, we model the response to treatment at 3 years ($rtt_{3y}$) conditional on the response at 12 months and residual tumoral tissue. Obviously, responses to treatment observed at different times cannot be considered independent since they come from the same data generation process.

DRAFT

$$P\left(rtt_{3y_i} \le j\right) = P\left(rtt^*_{3Y_i} \le \theta_j\right) = P\left(x_i\beta + w_{1i}\beta_{e1} + w_{2i}\beta_{e2} + w_{3i}\beta_{e3} + \varepsilon_i \le \theta_j\right) = \\ = P\left(\varepsilon_i \le \theta_j - x_i\beta - w_{1i}\beta_{e1} - w_{2i}\beta_{e2} - w_{3i}\beta_{e3}\right) = \\ = \Phi(\theta_j - x_i\beta - w_{1i}\beta_{e1} - w_{2i}\beta_{e2} - w_{3i}\beta_{e3}) \quad (19)$$

Here, the response to treatment at 3 years $rtt_{3y}$ is analysed conditional on a potential exogenous predictor, the residual suspicious tissue in thyroid bed ($x$), and the endogenous covariate $rtt_{12m}$ . The latter is ordinal and, when we come to the regression model it is represented by a series of binary variables, one for each level minus one ($\boldsymbol{w_i}$).

Considering the response at 12 months as endogenous in the model for the response at 3 years, it is important to evaluate the relation between these two variables and estimate the correlation $\rho_{\varepsilon,\eta}$ between the errors in models (17) and (18).

To properly consider the effects of covariates on response, given that patients are nested in clinical centres, we inserted in both equations the centre-specific mean value of each covariate, as in equation (10). To account for similarities within the same centre, we used a robust estimator for the standard errors of model coefficients.


## Acknowledgments

The drafting of the paper was possible thanks to the contribution of ITCO, Italian Thyroid Cancer Observatory, the research group of the Prof. Cosimo Durante and all the clinical centre that collaborate on the project.


## References


[1] A. Agresti, An Introduction to Categorical Data Analysis, 2nd ed., 2007.

[2] W. H. Greene, D. A. Hensher, Modeling Ordered Choices: A Primer and Recent Developments, 2008.

[3] Wooldridge J. M., Econometric Analysis of Cross Section and Panel Data, 2nd ed., Cambridge MA, MIT Press, 2010.

[4] J.M. Neuhaus, J.D. Kalbeisch, Between- and within-cluster covariate effects in the analysis of clustered data Biometrics, 54, 638-645, 1998.

[5] J.M. Neuhaus and C.E. McCulloch, Separating between- and within-cluster covariate effects by using conditional and partitioning methods Journal of the Royal Statistical Society, Series B, 68, 859-872, 2006.

[6] Winship C., Mare R. D., Regression models with ordinal variables, American Sociological Review, Vol. 49, 1984.

[7] G. Grani et al., Real-World Performance of the American Thyroid Association Risk Estimates in Predicting 1-Year Differentiated Thyroid Cancer Outcomes: A Prospective Multicenter Study of 2000 Patients, Thyroid, 2020.


DRAFT

# Monitoring drugs-based diagnostic therapeutic paths in heart failure patients using state-sequence analysis techniques

Nicole Fontana[a], Laura Savaré[a,b,c], and Francesca Ieva[a,b,c]

[a]MOX, Department of Mathematics, Politecnico di Milano, Milan 20133, Italy;
nicole.fontana@polimi.it, laura.savare@polimi.it
francesca.ieva@polimi.it
[b]HDS, Health Data Science Center, Human Technopole, Milan 20157, Italy;
[c]National Centre for Healthcare Research & Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy;

DRAFT

**Abstract**

The prevalence of heart failure (HF) is increasing globally and its treatment is mainly based on drug therapy. However, non-adherence to therapies is widespread in patients with heart failure and is often associated with worse health conditions and an increase in hospital admissions. This study focuses on developing an innovative method, the state-sequence analysis (SSA), for profiling HF patients based on different drug-utilisation patterns and then investigating if and how the information is extracted by combining clustering algorithms to this technique affects patients' overall survival. This information can be used by people in charge of the healthcare government to properly asses different patterns of care and then plan tailored patient care supporting the proper resource allocation within the National Health Service.

***Keywords:*** Heart failure, State-sequence analysis, Administrative database, Clustering

## 1. Introduction

Heart failure (HF) is currently the most common cardiovascular reason for hospitalisation among individuals over the age of 60. The prevalence of heart failure is increasing globally due to the increasing incidence of the ageing population (8), and the primary goals of its treatments are the reduction in mortality, the prevention of recurrent hospitalisation and improvement in the clinical status and functional capacity (5). The data suggest that non-adherence to drugs is associated with worse patient health conditions and an increase in hospital admissions. According to numerous studies, non-adherence is very common in patients with HF and the corresponding estimated rates range from 10% to 93%, making therapy non-adherence a severe medical problem on a global scale (12). For this reason, we focus on the therapeutic pathways administered to HF patients using an innovative method, at the expense of more traditional ones, to have a deeper description capable of evaluating the temporal order of drug prescription to extract drug-utilisation patterns and their association with health outcomes. The most widely administered therapies to HF patients are Angiotensin-Converting Enzyme Inhibitors (ACE-I). ACE-Is are the first class of drugs shown to reduce mortality and morbidity in HF patients and are recommended in all patients unless contraindicated or not tolerated. In this latter case, Angiotensin-Receptor Blockers (ARBs) are administered. ACE-I and ARB drugs will be grouped into a single drug class, namely

Renin-Angiotensin System (RAS) (6). In addition, Beta-Blocking agents (BB) combined with ACE-I have been shown to reduce mortality and morbidity in patients. Anti-Aldosterone agents (AA) are recommended, in addition to an ACE-I and a BB, to reduce mortality and the risk of HF hospitalisation (10). Moreover, diuretics (DIU) are recommended to reduce the signs and symptoms of congestion, which is one of the main predictors of HF. Finally, antithrombotic agents (AAG) are advised for patients with atrial fibrillation, which increases the risk of thromboembolic events (2). The aim of this work is to develop a suitable methodological framework for the representation of therapies prescription history in HF patients, with the final purpose of obtaining workable mathematical objects to be used in prognostic models. We apply such methods to an administrative database of hospitalised HF patients in the Lombardy region (RL) from 2006 to 2012. Administrative databases allow access to a large amount of heterogeneous data collected systematically. These data are rich in quantity and precision of information but poor in clinical content, as financial reasons drive their collection (11). The availability of many variables defined on a large number of patients makes the population similar to the "real-world" one, allowing the study of low-incidence phenomena and patients for an extended follow-up period. It is possible to extract information about drugs and comorbidities through algorithms based on national systems: the Anatomical Therapeutic Chemical (ATC) drug classification system and the International Classification of Diseases (ICD-9-CM). The information provided by this database allows us to estimate a proxy of the real drug-utilisation starting from drug purchases.

## 2. Cohort selection

The data used in this study is taken from the Healthcare Utilisation Databases of Lombardy, a Region of Italy with approximately 16% of its population (more than 10 million residents) (7). This database describes patients hospitalised for HF using information about hospitalisations and pharmaceutical purchases. For each patient, the follow-up begins with the discharge from the first HF hospitalisation and ends with his death or exit from the study. Then, we extracted the pharmacological class of each drug purchase using the ATC system. Next, using the ICD-9-CM system, each hospitalisation diagnosis encoded by the Clinical Classification Software was converted into comorbidity. This procedure provides the information necessary to compute the Multisource Comorbidity Score (MCS), which describes the patient's overall clinical condition (1). We decide to use this score in our analysis since it is the first that combines hospital diagnosis and drug prescriptions to provide a tool capable of measuring the severity of the clinical condition and it has demonstrated good performance in predicting mortality.

The last step of this pre-processing is the choice of the patients' cohort. As the goal is to study mortality one year after the reference date through drug-based patterns, all patients with censoring data that fall within this period or without pharmaceutical purchases were excluded. In addition, the analysis is focused on the three commonly administered drugs, in this context, RAS, BB and AA. Consequently, only patients who provided at least one purchase of these three drugs were taken into consideration in the first year of follow-up. The final sample is made up of 35,842 patients.

## 3. Methodologies

The analysis is developed as follows. First, SSA was applied to study patients' behaviour in terms of proper drug assumption and capture informative patterns that may affect patients' prognosis through sequence clustering. Then, sequence analysis is integrated into the predictive models through the clustering result to evaluate the association between drug patterns and health-related outcomes.

### 3.1 State-sequence analysis

State-sequence analysis was born to understand how events in ordered sequences can be related to the outcome of the observation. It is a well-defined technique in sociology that assesses how the chronological order of events in subgroups can lead to different social behaviours (9).

DRAFT

The main objective of sequence methods is to extract simplified, workable information from longitudinal data. Combined with cluster analysis, this method allows us to identify sequence descriptors that can be used in predictive models. In general, it is possible to divide the SSA into three main steps:

1. Identification sequences' states
2. Measuring sequence dissimilarity
3. Clustering sequences

State-sequences are characterised by two properties: an *alphabet* which is the list of all possible states or events, and a *time axis*, which assigns to each state a time stamp (3). Sequences are mathematical objects that contain elaborated information coming from the original data. An essential part of this study is the analysis of these objects, which may be conducted using visual tools and extracting statistics that characterise the sequences, like entropy or turbulence.

After analysing the sequences, it is necessary to define a metric to measure their dissimilarity. The sequences can be represented by different aspects that are not independent of each other, but the similarity of two sequences can be measured over one or many of these dimensions. So, comparing sequences is a significant issue of SSA, which may result in different partitions according to the metric adopted. We use a commonly used method called Optimal Matching distance. This edit-based metric generates edit distances that are the minimal cost, in terms of insertions, deletions and substitutions, for transforming one sequence into another. The insertion/deletion of an element, which generates a one-position shift of all the elements on its right, has the cost set to a default value of 1. At the same time, the substitution cost of one element by another derives from the observed transition rates between states. The parameterisation of the operations costs makes this measure flexible and allows for finding a trade-off between sequentiality and contemporaneity of the states (4). Once the dissimilarity matrix is computed, a cluster analysis can be performed to construct partitions of the sequences into distinct groups. Two clustering algorithms are applied to our work: hierarchical clustering and partitioning around medoids (PAM). In order to evaluate the partition obtained, we use three different metrics to assess their capability:

- Point Biserial Correlation (PBC) which measures the partition's capacity to reproduce the distance matrix
- Hubert's C (HC) index which measures the gap between the partition obtained with the best partition that could have been obtained with this number of groups and this distance matrix
- Average Silhouette Width (ASW) which measures the coherence of observation assignments to a given group

We obtain the final clustering by applying hierarchical clustering to the whole set of sequences and choosing the best number of clusters maximising the PBC and ASW and minimising the HC. Then we initialise the PAM algorithm with the best result of the hierarchical clustering. Finally, we choose the partition which gives the best quality metrics and a significant interpretation of the clusters.

DRAFT

# 4. Results

## 4.1 State-sequence construction

Using SSA, we study RAS, BB, AA, DIU and AAG drug classes prescribed during the first year after the first hospital discharge of each patient. This period allows us to observe patients' behaviour long enough to characterise them but at the same time not excessively long to limit the effect of the immortal bias. To construct the sequences, we keep all the purchases made during the first year of observation with the condition that if the drug coverage expires after the year, we truncate the coverage days to the final day of observation. Then we eliminate all the purchase overlaps, which occur when the purchase date of a drug happens during the coverage period of a previous purchase of the same type of drug. Patient purchases are grouped by pharmaceutical class and studied separately; in this way, each patient is associated with a sequence for each of the five drug classes taken at least once in the first year of observation. Sequences are composed of 52 elements, each of which represents a week of observation. Each state is assigned

zero if the drug prescription does not cover the patient for at least four out of seven days of the specific week and one otherwise.

For the RAS, BB and AA therapies, we introduce an additional phase that enables us to analyse the simultaneous intake of these drugs. For each patient starting from the three single sequences of RAS, BB and AA, we build a sequence that combines their states, forming an alphabet of eight elements. The states that make up the alphabet represent which of the three drugs are taken by the patient in any week. At this point, we have a combined-sequence of RAS & BB & AA for each patient and two separate sequences indicating the intake of DIU e AAG only for those patients who assume them.

## 4.2 State-sequence analysis

The analysis of the combined-sequences RAS & BB & AA performed through visual tools and statistical measures show that over the first two months, the proportion of patients who do not take any therapy decreased in favour of a rise in the RAS assumption. A small fraction of patients takes AA and BB separately, showing roughly the same proportion across the observation period. BB and AA are rarely taken by people who do not already take RAS. Looking at the transition rates, the probability of remaining in states which represent no treatments or monotherapy is elevated. However, BB and AA states are more unstable since the probability of stopping these therapies when accompanied by RAS is relatively high. Generally, it is more likely for a patient to leave a therapy than to start or add a new one.

The diuretics and antithrombotic sequences show similar behaviour. In the first weeks, the proportion of patients taking them increases by about 20%, then stabilises. This proportion remains almost constant in the weeks of observation, counting about 60% of the patients who take the therapy.

## 4.3 State-sequence clustering

Following the clustering procedure, we obtain for the combined-sequence RAS & BB & AA an eight-cluster partition from the hierarchical clustering. This partition captures the main behavioural patterns of the patients assuming the three classes of drugs, as shown in 1. Clusters 1, 2 and 3 denote patients with constant therapy patterns for most of the observation period. In particular, cluster 1 identifies patients who are non-adopters of therapies, cluster 2 the combined intake of RAS and BB therapies and cluster 3 RAS as monotherapy. Clusters 4, 5 and 6 identify those patients who, on average, take monotherapy for half of the observation period and no therapy for the remaining weeks. Finally, clusters 7 and 8 identify heterogeneous behaviours about the three pharmacological classes.

A two-cluster partition from the PAM method is obtained for the DIU and AAG sequence. This grouping recognises patients as *low-adopters* and *adopters* of these two drugs. The choice of such a small partition is supported by the quality metrics and the interpretation we can make. Compared to the combined-sequences, these sequences represent more simplistic patterns since they only have two states. We underline that this innovative procedure provides insights into how diuretics or antithrombotic agents are prescribed over time, not only recording the flag of having or not the prescription.

## 4.4 Including SSA-based representation into predictive models

To each patient, is associate the information about the three sequences clustering to examine, through the use of the Cox regression model, whether and to what extent this information adds value in predicting the overall survival of patients. We add to the model some explanatory variables, such as age, gender, the multisource comorbidity score, and the total number of procedures the patient has undergone.

Looking at the p-values of the stratified log-rank tests, we can state that the difference in survival between all the groups is significant. In particular, being diuretics adopters decreases the chances of survival by 40% (CI, [35% - 46%]); this could be explained by the fact that those who take diuretics are in a severe clinical condition. On the contrary, being antithrombotic adopters increases the survival probability of 11% (CI, [8% - 14%]), which identifies this drug as a protective factor.

Looking at the cluster of the combined-sequence of RAS & BB & AA, we can state that all the therapies increase the probability of survival since the Hazard Ratios are all significantly less than one.

DRAFT

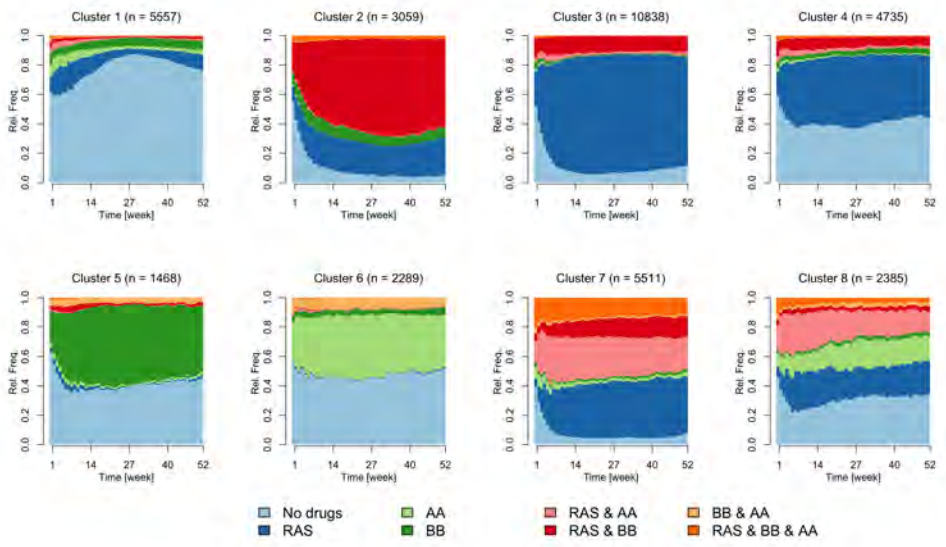



Figure 1: Sequence distribution plot by cluster of the combined-sequence of RAS & BB & AA.

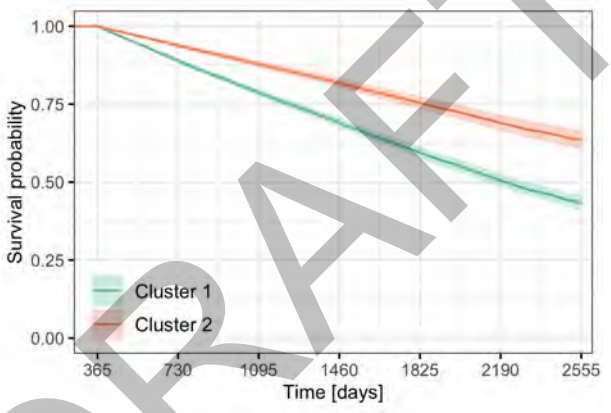



Figure 2: Survival probability plot for male patients aged 75 with 15 days of hospital stay, low category of MCS, without undergoing procedures, which are low-adopters of DIU and AAG and that belongs to clusters 1 and 2 of the combined-sequence of RAS & BB & AA.

DRAFT

In particular, patients belonging to cluster 2 have the highest survival probability, which is 46% (CI, [40% - 51%]) higher than the reference cluster, as shown in the KM survival curves in 2. So, we can infer that despite the diagnosis severity being unknown, taking drugs for an extended period after the first hospitalisation seems essential to improve the prognosis of HF.

## 5. Discussion and conclusions

The state-sequence analysis allows the description of drug-utilisation patterns for each patient through a mathematical object, which provides tools for assessing and associating the primary endpoint of interest. Compared to commonly used baseline measures, which omit some time-dependent information, this technique yields more realistic and valuable results since it considers the entire evolution of each patient's clinical path. In particular, it allows observing which weeks and how long the patient is covered by the therapies after the first hospitalisation. In addition, it allows the analysis of the polytherapy taken by the patient using a single patient descriptor (the sequence). The SSA, therefore, allows a change of perspective in the analysis of the prescriptions, moving from a transversal and syntactical approach to

a holistic one that exploits the information available through the application of statistical tools, slightly more complex than traditional methods. One of the limitations, deriving from the use of administrative databases, is the lack of detailed and contextual clinical information that would have supported the predictive models.

This work represents an important step in evaluating HF patients' drug-based path and paves the way for future developments. Latent Markov Models can be applied to study hidden patterns in this data to investigate the evolution of an individual characteristic which is not directly observable. Future expansions regard the application of these techniques to the so-called multichannel sequence analysis, i.e., the case where multiple sources of information are described through SSA, allowing us to describe individual trajectories on several dimensions simultaneously.

In conclusion, SSA has provided impressive results in studying drug-utilisation patterns of heart failure patients. However, more specific and updated data can lead to even more effective results that can support healthcare specialists in evaluating the pathways provided to patients and the National Health Service in allocating the most appropriate resources.

## References


[1] Corrao, G., Rea, F., Di Martino, M., De Palma, R. & Others: Developing and validating a novel multisource comorbidity score from administrative data: a large population-based cohort study from Italy. BMJ Open. **7** (2017)

[2] Ferreira, J. P., Girerd, N., Alshalash, S.: Antithrombotic therapy in heart failure patients with and without atrial fibrillation: update and future challenges. Eur. Heart J. **37**, 2455-2464 (2016)

[3] Gabadinho, A., Ritschard, G., Müller, N. S., Studer, M.: Analyzing and visualizing state sequences in R with TraMineR. J. Stat. Softw. **40** (2011)

[4] Lesnard, L.: Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. Sociological Methods & Research. **38**, 389-419 (2010)

[5] Pazos-López, P., Peteiro-Vázquez, J. & Others: 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur. Heart J. **42**, 3599-3726 (2021)

[6] Probstfield, J., O'Brien, K.: Progression of Cardiovascular Damage: The Role of Renin–Angiotensin System Blockade. The American Journal Of Cardiology. **105**, 10A-20A (2010)

[7] Rea, F., Savaré, L., Franchi, M., Corrao, G., Mancia, G.: Adherence to Treatment by Initial Antihypertensive Mono and Combination Therapies. American Journal Of Hypertension. **34**, 1083-1091 (2021)

[8] Rossignol, P., Hernandez, A. F., Solomon, S. D., Zannad, F.: Heart failure drug treatment. Lancet. **393**, 1034-1044 (2019)

[9] Studer, M., Ritschard, G.: A comparative review of sequence dissimilarity measures. (2014)

[10] Swedberg, K. & Others: Guidelines for the diagnosis and treatment of chronic heart failure: executive summary (update 2005): The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology. European Heart Journal. **26**, 1115-1140 (2005)

[11] Timofte, D., Pantea Stoian, A., & Others: A Review on the Advantages and Disadvantages of Using Administrative Data in Surgery Outcome Studies. The Journal Of Surgery. **14** (2018)

[12] Wu, J., Moser, D. K. and Lennie, T. A., Burkhart, P. V.: Medication adherence in patients who have heart failure: a review of the literature. Nursing Clinics Of North America. **43**, 133-53; vii–viii (2008)


DRAFT

# Optimal two-stage design based on error rates under a Bayesian perspective

Susanna Gentile[a] and Valeria Sambucini[a]

[a]Dipartimento di Scienze Statistiche, Sapienza Università di Roma;
susanna.gentile@uniroma1.it, valeria.sambucini@uniroma1.it

**Abstract**

In phase II clinical trials, two-stage designs allowing early stopping for lack of efficacy are frequently used. We present a Bayesian two-stage design that ensures high posterior probabilities that the response rate of the experimental drug exceeds a desirable level, when the decision is to proceed with treatment evaluation. Moreover, the design exploits the distinction between *analysis* and *design* prior distributions, to control the predictive probability of Type I and II errors, while minimizing the expected sample size under the null hypothesis.

***Keywords:*** analysis and design priors, Bayesian approach, error rates, two-stage design

DRAFT

## 1. Introduction

Single-arm two-stage designs are frequently used in phase II clinical trials with binary endpoints. The aim is to stop the study early for futility if the response rate of the experimental drug is not sufficiently high. According to the original scheme suggested by Simon (1989), at the first stage $n_1$ patients are enrolled and if $s_1 \leq r_1$, the trial terminates for lack of efficacy, where $s_1$ denotes the observed number of responders. Otherwise, the trial continues to the second stage and $n_2$ additional patients are treated. Then, if the observed number of responders $s$ out of the total sample $n = n_1 + n_2$ is not grater than $r$, the trial stops and the treatment is declared not effective. Otherwise, it is recommended for a more rigorous evaluation in a phase III trial.

The most popular two-stage designs are the optimal and the minimax designs developed by Simon (1989) under a frequentist framework, with many extensions and modifications presented in the literature. Several Bayesian two-stage designs have been also proposed (see, among others, Tan and Machin, 2002; Sambucini, 2008; Dong et al., 2012; Matano and Sambucini, 2016). In a recent work, Shi and Yin (2018) proposed a Bayesian enhancement two-stage (BET) design, that $(i)$ guarantees a high posterior probability of the response rate exceeding the target of interest, when the observed number of responders reaches the minimum required level to continue with the experimentation and $(ii)$ controls the length of the HPD interval for the response rate. In line with Shi and Yin (2018), we propose a Bayesian two-stage design that satisfies condition $(i)$ and also yields the minimum expected sample size under the null hypothesis, while controlling the probability of Type I and Type II errors at a certain prespecified levels. To compute the error probabilities, we use a predictive approach by exploiting different kinds of prior distributions, which play different roles in the design of the trial. More specifically, we introduce an *analysis prior distribution* to express pre-experimental information and to construct posterior distributions. On the other hand, we elicit two different *design prior distributions* to represent suitable design scenarios to evaluate the Type I and Type II error rates.

The outline of the article is as follows. In Section 2, we formalize the Bayesian problem. In Section 3, the procedure to obtain the predictive probabilities of errors is described and the Bayesian strategy to find the optimal design is illustrated. Some numerical results are given in Sections 4 and, finally, Section 5 contains a brief discussion.

## 2. Preliminaries

Let $\theta$ be the unknown response rate of the experimental treatment and assume that the interest is focused on testing $H_0 : \theta \leq \theta^*$ vs $H_1 : \theta > \theta^*$. Thus, the treatment is considered sufficiently promising if $\theta$ exceeds the target value $\theta^*$. Moreover, let us denote by $S_1$ and $S$ the total number of responders at the end of the first and the second stage, respectively.

We introduce a beta prior density, $\pi^A(\theta) = \text{Beta}(\theta; \alpha^A, \beta^A)$, and use it to obtain the posterior distributions of $\theta$ at the end of both the stages. $\pi^A(\theta)$ is called the *analysis prior distribution*, because it is used to represent pre-experimental knowledge available at the analysis stage, that can be for instance gathered by previous studies or derived from expert opinions. We have that $S_1|\theta \sim \text{Bin}(n_1, \theta)$ and, from standard conjugate analysis, the first stage posterior distribution of $\theta$ is $\text{Beta}(\theta; \alpha^A + s_1, \beta^A + n_1 - s_1)$. Since $S$ includes the number of successes of the first stage, the posterior distribution for $\theta$ at the end of the second stage should be conditional on the event $S_1 > r_1$. However, it is possible to show that this condition does not affect the second stage posterior distribution (see Sambucini, 2008), that results to be $\text{Beta}(\theta; \alpha^A + s, \beta^A + n - s)$.

In each stage, we assume that the trial terminates if the posterior probability assigned to the alternative hypothesis is not sufficiently high. More specifically, at the end of the first stage the trial proceeds to the second one if

$$Pr(\theta > \theta^* | S_1 = s_1, n_1) > \lambda_1, \tag{1}$$

and, similarly, at the end of the second stage we claim the experimental treatment promising if

$$Pr(\theta > \theta^* | S = s, n) > \lambda_2, \tag{2}$$

where $\lambda_1$ and $\lambda_2$ are two desired probability thresholds, typically fixed so that $\lambda_2 > \lambda_1$.

DRAFT

## 3. The proposed two-stage design

For fixed values of $n_1$ and $n$, the posterior probabilities in (1) and (2) are increasing functions of $s_1$ and $s$, respectively. We use an algorithm that searches the optimal design by varying $n$ from $n^{min} = 10$ to $n^{max} = 100$. For each value of $n$, we consider $n_1$ in the range from $n_1^{min} = \max\{5, \frac{n}{3}\}$ to $n_1^{max} = n-1$. This choice for $n_1^{min}$ aims at avoiding that the sample size of the first stage may be relatively small compared to the total sample size. For each couple of $n$ and $n_1$, the two-stage boundaries are selected as

$$r_1 = \min\left\{s_1 \in \{0, ..., n_1\} : \; Pr(\theta > \theta^* | S_1 = s_1, n_1) > \lambda_1\right\} - 1 \tag{3}$$

$$r = \min\left\{s \in \{r_1 + 1, ..., n\} : \; Pr(\theta > \theta^* | S = s, n) > \lambda_2\right\} - 1 \tag{4}$$

In this way, we obtain a set of designs $(n_1, r_1, n, r)$ such that in each stage the posterior probability assigned to the alternative hypothesis is larger than the threshold of interest, when the observed number of responders exceeds the corresponding boundary, $r_1$ or $r$. Among these designs, we select the one that satisfies error probability constraints at the end of the second stage and minimizes the expected sample size under $H_0$.

### 3.1 Error probability computation

We consider the standard two types of errors that can occur in hypothesis testing: rejecting the null hypothesis when it is actually true (Type I) and failing to reject the null hypothesis, when the alternative is

true (Type II). When studying the operating characteristic of two-stage designs under a frequentist framework, the probabilities of these errors are evaluated by specifying a single value for $\theta$, suitably selected under $H_0$ or $H_1$ according to the type of error we are interested in. By adopting a Bayesian approach, we instead introduce two different prior distributions that model uncertainty on the single values of the parameter specified in the classical framework and add flexibility to the procedure. In the statistical literature, these priors are typically called *design prior distributions*, because they are used at the design stage of the study to describe a scenario of interest and to derive the prior predictive distributions of the data (see, Wang and Gelfand, 2002; Sahu and Smith, 2006; Brutti et al., 2008; Sambucini, 2008). This allows to obtain the probability model that generates the data, under the assumption that $\theta$ is highly likely to be in a certain subset of the parameter space.

More specifically, to compute the Type I error rate, we need to realize the conjecture that $H_0$ is true. Thus, we introduce a beta design prior distribution for $\theta$, $\pi^D_{H_0}(\theta) = \text{Beta}(\theta; \alpha^D_{H_0}, \beta^D_{H_0})$, which has mode $\theta^D_0$ smaller than $\theta^*$ and assigns negligible probability to values of the parameter under the alternative hypothesis. By marginalizing the sampling distribution of the data over $\pi^D_{H_0}(\theta)$, we obtain the prior predictive distribution of $S_1$ and $S - S_1$, that are

$$m^D_{H_0}(s_1) = \text{Bin-Beta}(s_1;\, n, \alpha^D_{H_0}, \beta^D_{H_0}), \quad \forall\, s_1 = 0, ..., n_1,$$

$$m^D_{H_0}(s - s_1) = \text{Bin-Beta}(s - s_1;\, n - n_1, \alpha^D_{H_0}, \beta^D_{H_0}), \quad \forall\, s - s_1 = 0, ..., n - n_1.$$

Here, Bin-Beta$(\cdot\,; m, a, b)$ denotes the probability mass function of a Binomial-Beta distribution with parameters $m$, $a$ and $b$. Therefore, given the four values $(n_1, r_1, n, r)$, the predictive probability of a Type I error is provided by

$$Pr(\text{Type I error}) = \sum_{i=r_1+1}^{n_1} \sum_{j=r+1}^{n} \text{Bin-Beta}(i;\, n_1, \alpha^D_{H_0}, \beta^D_{H_0}) \text{Bin-Beta}(j - i;\, n - n_1, \alpha^D_{H_0}, \beta^D_{H_0}).$$

When the focus is on the Type II error, we elicit a beta design prior distribution for $\theta$, $\pi^D_{H_1}(\theta) = \text{Beta}(\theta; \alpha^D_{H_1}, \beta^D_{H_1})$, used to realize the assumption that the alternative hypothesis is true. Its prior mode is $\theta^D_1 > \theta^*$ and the prior probability assigned to values of the $\theta$ under the null hypothesis is negligible. Analogously to the previuos case, $\pi^D_{H_1}(\theta)$ is exploited to obtain the prior predictive distribution of $S_1$ and $S - S_1$, that are still Binomial-Beta. Therefore, the predictive probability of a Type II error is given by

$$Pr(\text{Type II error}) = 1 - \sum_{i=r_1+1}^{n_1} \sum_{j=r+1}^{n} \text{Bin-Beta}(i;\, n_1, \alpha^D_{H_1}, \beta^D_{H_1}) \text{Bin-Beta}(j - i;\, n - n_1, \alpha^D_{H_1}, \beta^D_{H_1}).$$

## 3.2 Optimal design strategy

As described before, the first step of the proposed strategy to find the optimal design consists in considering all the possible couples of $n$ and $n_1$ and identifying the corresponding boundaries $r_1$ and $r$ by exploiting the conditions (3) and (4) about the posterior probabilities that the alternative hypothesis is true. Among the set of $(n_1, r_1, n, r)$ identified using this procedure, we select the one that

1. satisfies the error constraints $Pr(\text{Type I error}) < \alpha$ and $Pr(\text{Type II error}) < \beta$, where $\alpha$ and $\beta$ are desired levels for the error rates;
2. minimizes the expected sample size under $H_0$, $E(N|H_0) = n_1 + (n - n_1)(1 - PET(H_0))$, where $PET(H_0)$ is the *probability of early termination*, that is computed for $\theta$ equal to the mode of the design prior $\pi^D_{H_0}(\theta)$ and is given by

$$PET(H_0) = \sum_{i=0}^{r_1} \binom{n_1}{i} (\theta^D_0)^i (1 - \theta^D_0)^{n_1 - i}.$$

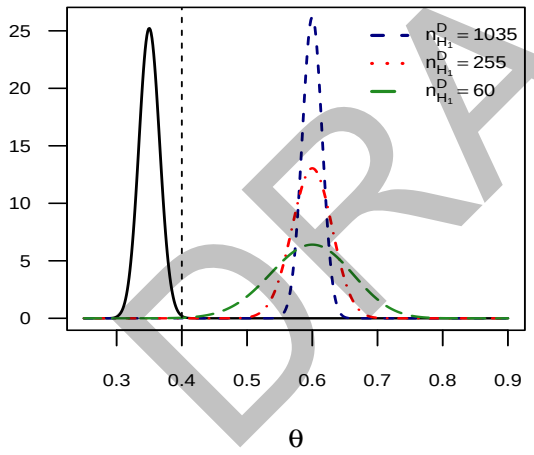
DRAFT

## 4. Numerical results

In this Section, we report some numerical results to illustrate the main features of the proposed design. We set $\theta^* = 0.4$ and consider a non-informative analysis prior distribution, $\pi^A(\theta) = \text{Beta}(\theta; 1, 1)$. To elicit the design prior distributions, we resort to a typical way of proceeding by expressing the hyperparameters in terms of prior mode and prior sample size. For instance, the hyperparameters of the design prior $\pi^D_{H_1}(\theta)$ can be fixed as

$$\alpha^D_{H_1} = n^D_{H_1}\theta^D_1 + 1 \qquad \text{and} \qquad \beta^D_{H_1} = n^D_{H_1}(1 - \theta^D_1) + 1,$$

where the prior sample size $n^D_{H_1}$ regulates the concentration of the prior around its mode. In particular, we set $\theta^D_1 = 0.6$ and select $n^D_{H_1}$ so that it is equal to 0.99 the prior probability assigned to the intervals $[0.55, 0.65]$, $[0.5, 0.7]$ and $[0.4, 0.8]$. Consequently, we obtain three design priors with $n^D_{H_1}$ equal to 1035, 255 and 60, respectively, that assign negligible probability to values of $\theta$ smaller than $\theta^*$. These distributions are represented in Figure 1 along with the design prior $\pi^D_{H_0}(\theta)$. This latter density has mode $\theta^D_0 = 0.35$ and is based on a prior sample size, $n^D_{H_0}$, equal to 909. On the right of the graph, a small Table shows the optimal designs that correspond to the diffrent design priors $\pi^D_{H_1}(\theta)$. As expected, the corresponding optimal sample sizes increase when $n^D_{H_1}$ decreases, as a consequence of the greater dispersion of the design distribution.




| $\theta^*$ | $n^D_{H_1}$ | $r_1/n_1$ | $r/n$ |
|---|---|---|---|
| 0.4 | 1035 | 7/16 | 16/33 |
| | 255 | 7/16 | 19/40 |
| | 60 | 9/20 | 23/49 |

DRAFT

Figure 1: Design prior distributions $\pi^D_{H_0}(\theta)$ and $\pi^D_{H_1}(\theta)$ for different values $n^D_{H_1}$, when $\theta^* = 0.4$. The Table shows the corresponding optimal designs when $\pi^A(\theta) = \text{Beta}(\theta; 1, 1)$ and $(\alpha, \beta) = (0.05, 0.2)$

In Table 1, we report the optimal designs for different values of $\theta^*$ and $(\alpha, \beta)$, when $\theta^D_0 = \theta^* - 0.05$, $\theta^D_1 = \theta^* + 0.2$, $\alpha^A = \beta^A = 1$, $\lambda_1 = 0.8$ and $\lambda_2 = 0.9$. The prior sample sizes of the design priors, $n^D_{H_0}$ and $n^D_{H_1}$, are selected as the minimum values so that the probabilities assigned to $H_0$ and $H_1$, respectively, are equal to 0.999. The corresponding probabilities of early termination and the expected sample sizes are also reported. Regardless of $\theta^*$, the optimal sample sizes obtained when $\alpha = \beta = 0.1$ in both the stages are greater than the ones obtained with $\alpha = 0.05$ and $\beta = 0.1$. Furthermore, in the first case the probabilities of early termination are higher, resulting in expected sample sizes closer to $n_1$. As for the analysis prior distributions, we expect that their impact varies according to the degree of skepticism expressed towards the treatment efficacy.

Table 1: Optimal proposed two-stage design for different values of $\theta^*$, $\alpha$ and $\beta$, when $\theta_0^D = \theta^* - 0.05$, $\theta_1^D = \theta^* + 0.2$, $\alpha^A = \beta^A = 1$, $\lambda_1 = 0.8$ and $\lambda_2 = 0.9$

| $\theta^*$ | $(\alpha, \beta)$ | $r_1/n_1$ | $r/n$ | $EN(H_0)$ | $PET(H_0)$ |
|---|---|---|---|---|---|
| 0.2 | (0.1, 0.1) | 6/27 | 12/48 | 27.070 | 0.901 |
| | (0.05, 0.2) | 3/14 | 7/27 | 15.094 | 0.853 |
| 0.3 | (0.1, 0.1) | 11/33 | 25/70 | 36.652 | 0.901 |
| | (0.05, 0.2) | 6/18 | 14/38 | 20.780 | 0.861 |
| 0.4 | (0.1, 0.1) | 18/41 | 27/58 | 42.509 | 0.911 |
| | (0.05, 0.2) | 9/20 | 27/58 | 24.628 | 0.878 |
| 0.5 | (0.1, 0.1) | 22/40 | 40/71 | 42.377 | 0.923 |
| | (0.05, 0.2) | 11/20 | 27/47 | 23.531 | 0.869 |

## 5. Discussion

The most popular and commonly used two-stage designs have been developed by Simon (1989) under a frequentist framework. From a Bayesian perspective, Shi and Yin (2018) showed that, given promising results according to Simon's designs, the posterior probabilities that the response rate reaches the desirable target level are very low. Thus, this Author proposed a Bayesian two-stage design that ensures high posterior probabilities of the response rate exceeding the target of interest, when the observed results suggest to proceed with treatment investigation.

In this paper, we present a Bayesian two-stage design based on the same condition used by Shi and Yin (2018) and that in addition minimizes the expected sample size under null hypothesis, while controlling the Type I and Type II error rates at fixed desired levels. To offer a more flexible evaluation of the errors probabilities, we adopt a predictive approach by using design prior distributions to specify design expectations and to realize the assumption that $\theta$ belongs to a specific subset of the parameter space. These prior densities are employed to obtain the prior predictive distribution of the data, used to compute the error probabilities. A different prior distribution is used to represent prior information and to compute the posterior distribution of the parameter.

DRAFT

## References


[1] Brutti P., De Santis F., Gubbiotti S.: Robust Bayesian sample size determination in clinical trials. Stat. Med., **27**, 2290–2306 (2008)

[2] Dong G., Shih, W.J., Moore, D., Quand, H., Marcella, S.: A Bayesian-frequentist two-stage single-arm phase II clinical trial design. Stat. Med., **31**, 2055–2067 (2012)

[3] Matano, F., Sambucini, V.: Accounting for uncertainty in the historical response rate of the standard treatment in single-arm two-stage designs based on Bayesian power functions Pharm Stat., **15**, 517–530 (2016)

[4] Sahu, S.K., Smith, T.M.F.: A Bayesian method of sample size determination with practical applications. J. R. Stat. Soc., **169**, 235–253 (2006)

[5] Sambucini V.: A Bayesian predictive two-stage design for phase II clinical trials. Stat. Med., **27**, 1199–1224 (2008)

[6] Shih, H., Yin, G.: Bayesian enhancement two-stage design for single-arm phase II clinical trials with binary and time-to-event endpoints. Biometrics, **74**, 1055–1064 (2018)

[7] Tan, S. B., Machin, D.: Bayesian two stage designs for phase II clinical trials. Stat. Med., **21**, 1991–2012 (2002)

[8] Wang, F., Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for

performance under a given model and for separating models. Stat. Sci., **2**, 193–218 (2002)

DRAFT

# EU-Border crisis on Twitter: sentiments and misinformation analysis

Elena Ambrosetti[a], Cecilia Fortunato[a], and Sara Miccoli[b]

[a] Sapienza University of Rome; elena.ambrosetti@uniroma1.it, cecilia.fortunato@uniroma1.it

[b] Istat; sara.miccoli@istat.it

**Abstract**

The objective of this paper is to investigate the information and to detect the presence of misinformation on Twitter posts circulating in relation to migration events happened in 2020 at the Greek-Turkish border and in 2021 at the Polish-Belarusian border. Data were retrieved through API by using keywords referring to the two border events. The study was carried out by applying text mining and sentiment analysis techniques on tweets and retweets related to these two events, and by conducting a qualitative analysis on specific subsets of tweets. Our results show that in both borders' crises migration is perceived as an emergency issue, migration-related narratives mainly refer to "war", "attacks", "tension", "invasion" and the emotions expressed are mostly negative. In addition, in outbreaking crisis, the identification of misinformation in social media is extremely challenging, because of the rapid circulation of rumours related to facts that are rather difficult to ascertain.

**Key words:** misinformation, twitter, migration, Europe, border

DRAFT

## 1. Migration events at the Greece-Turkey and Poland-Belarus borders

This study focuses on the analysis of information spreading on social media during specific migration-related events in Europe and it is part of the work conducted within the H2020 Perceptions project[1]. The study focuses on a relevant topic related to migration phenomenon, i.e., the spread of misinformation and rumours that can have a great impact on the lives of migrants and influence the ideas circulating among migrants themselves and the inhabitants of receiving countries.

We focus on two borders' crises. The first one is the border between Greece and Turkey. This border has been the scene of several humanitarian crises involving migrants in recent years. One of the last events originated in late February 2020, when the Turkish president declared the opening of the border for migrants directed to Europe. During spring 2020 videos, pictures, and news on migrants at the Greek-Turkish border started to circulate. The two sides involved, Greece and Turkey, began to make accusations against each other regarding the treatment of migrants and the circulation of misinformation. The Greek and Turkish governments themselves have actively participated in this exchange of accusations, also by using social media. In this context, not only hardly verifiable news began to be spread, but real news was possibly being pointed as fake news or made confusing for political purposes.

Beyond specific ad hoc created fake news then, there was an underlying misinformation and confusion about real events. Misinformation also arose regarding the size of the phenomenon, with the Turkish side tending to overestimate the number of migrants who left Turkish territory. In addition to distorted news, real fake news was artfully created.

The second border's crisis that we analyse is the one on Poland-Border border in 2021. It was triggered by the tensions' escalation in Belarus–European Union (EU) relations started in August 2020, following the Belarusian presidential election, the repression of mass-protests in the country, and other events such as the hijacking of the Ryanair Flight 4978 for the arrest of Belarusian journalist Roman Protasevich. As a response, United States Treasury Department, and the EU imposed sanctions to President Aleksander Lukashenko and to exponents of Belarusian administration and economy. In May 2021, President Aleksander Lukashenko warned the EU of the possibility for Belarus to stop patrolling

[1] The project PERCEPTIONS is funded by the European Commission H2020 Research & Innovation Action under Grant Agreement No 833870. It aims to identify and understand narratives, images, and perceptions of the EU outside Europe, and to study ways, channels and actors through which narratives are distributed.

the EU border, threatening EU to provoke a migrant's crises in response to sanctions. From that moment, Frontex reports thousands of border crossing attempts from Belarus into EU, with the growing apprehension of Polish and European authorities and public opinion as well as the rising tension in the whole region, involving Russia, Latvia, Lithuania and Ukraine. The president of Belarus has been publicly accused by European authorities of offering tourists' visas and flights to migrants from Iraq and Syria. While thousands of migrants were stuck on the Belarus side of the border, tensions between Belarusian government and EU increased also in contents posted on Twitter. Both sides were accusing the other to use migrants as a destabilizing weapon, neglecting international protection's procedures and keeping migrants in life danger.

Both the contexts under study involved two EU border countries (Poland and Greece) facing migrant's events perceived as crises, under the guidance on the EU in a controversy with third countries (Belarus and Turkey). Both crises have echoed widely in the mainstream media, prompted discussions and news circulations (accurate or distorted) in social media. The spread of misinformation on social media has undoubtedly exacerbated the tension in crisis phases and could have also misled migrants and potential migrants' idea about the migration issue in Europe.

## 2. Research design, data and method

Focusing on misinformation this concept has numerous, and sometimes contradictory, definitions (Treen et al. 2020, p. 3). According to the Oxford English Dictionary 2018, misinformation is defined as "wrong or misleading information". In many cases, misinformation is used as a synonym of false facts, rumours (e.g., Donato et al, 2022). Slightly different is the concept of disinformation, defined in literature as a "deliberately false information". This differentiation has been embraced by many authors (e.g. Stahl 2006, Alonso et al. 2021, Treen et al. 2020), while others, such as Ruokolainen & Widén (2020), do not rigidly classify the terms misinformation and disinformation. In the framework of this analysis, we decided to adopt the umbrella concept of *misinformation*, as ambivalent, distorted or falsified information (Zhou and Zhang 2007), without distinction as to intention or source of origin of
the information. Rumours and fake news are thus equally included in this concept.

The objective of the work is to investigate the information and misinformation circulating in Twitter during specific geopolitical crisis related to migration flows at the abovementioned border areas, by examining tweets and retweets retrieved using targeted keywords.

In particular, the first research question of the study is: (RQ1) what words are most frequently expressed in Twitter environment during migration-related events at border areas and what sentiments do they communicate about the discussion taking place in social media? Our aim is to examine the nature of discourse developed on Twitter around these issues. We hypothesize that the words used, and the sentiment expressed, may indicate the presence of certain topics related to specific actors of these migration events.

The second research question is: (RQ2) is it possible to identify misinformation in Twitter contents about migration-related events at the borders? We hypothesize that tweets posted during the uprising of these kind of crises are a huge, heterogeneous and fragmented corpus of text and they can be hardly analysed as whole. Therefore, a qualitative approach on subsets of tweets containing selected keywords could be more informative for the analysis of patterns in narratives and misinformation about specific topics.

The data used for the analysis were downloaded using the Academic Research Twitter API, by searching for the hashtags #GreeceTurkeyBorder and #PolandBorder and #BelarusBorder. Our analysis is restricted to the content in English language, since English messages are more likely to cross national borders and spread all over the world.

For these three complete datasets, we applied text mining and sentiment analysis techniques to analyse the words used and to understand what kind of sentiment they express.

Using the *TM* package in R, we have built a term-document matrix that describes the frequency of single terms that occur in the corpus of documents. To perform the analysis, we cleaned the text data by applying a series of filters and we only selected semantically significant terms. We analysed the association between words through a correlation analysis, a statistical technique that can demonstrate whether, and how strongly, pairs of words could be related, helping to map the discourses around these words.

DRAFT

We investigated the most recurrent words related to the migrant situation at the borders to evaluate the prevalence of specific topics in tweets' contents.

We then conducted a sentiment analysis to understand what sentiments and emotions the tweets expressed, by using the *syuzhet* R package. We used the NRC Emotion Lexicon, developed by Saif and Turney (2010), that assigns words to 8 different types of emotions (Naldi 2019; Widyaningrum et al. 2019). The 8 emotions are anger, fear, sadness, disgust (negative ones) and anticipation, trust, surprise, and joy (positive ones). For each tweet the method counts the number of words associated with each category (Naldi 2019, Widyaningrum et al. 2019), providing a cumulative score representing the sentiments for the whole corpus of text.

We then created various subsets of specific tweets (excluding retweets). First, we selected tweets containing the words "misinformation", "disinformation", and "fake". On these we conducted a manual qualitative analysis, trying to understand which news items were categorized as misinformation by the users in social media. Then, we created targeted subsets of tweets, by searching for selected keywords related to specific topics that generated a major flow of information during the two events. We manually analyzed the contents to understand what kind of information was conveyed and in what ways, in order to identify biased or inaccurate information.

## 3. Section Heading

The analysis produced some interesting results, useful in understanding what kind of discourse emerges during critical situations concerning the migration phenomenon, and what role misinformation can play. Although these results derive from contents posted on Twitter only in English and related to specific geographic contexts, they can be representative of general discourses spread via social media related to migration phenomena.

We found common patterns among the contents concerning the situations on the Greek-Turkish and Polish-Belarusian borders. In both cases, the words circulating in social media about these migration-related events refer to wars, attacks, tension, crises, invasion. In addition to terminology related to the theme of conflict, the use of words connected with the most vulnerable groups, such as children and women, is also frequent. In both cases, the perception of migration as an emergency issue emerges. The feelings and emotions expressed by the words contained in these tweets are mostly negative. The messages and information existing in social media regarding these issues are therefore dominated by negative elements, mainly related to anger and fear (RQ1).

The qualitative analysis carried out on more specific subsets of data made it possible to integrate and deepen the results emerged from the text mining analysis and to understand whether misinformation could be identified in tweets concerning specific topics. Regarding the different topics related to critical situations at the borders, confused discourses emerged, with polarized messages and viewpoints, in which it is very difficult to distinguish between misinformation and accurate information (RQ2). This is true both in the context of events on the Greek-Turkish and on the Polish-Belarusian borders.

Opposing views emerged regarding migrants. On the one hand, they are depicted as victims of a cynical policy in a game for power assets, on the other hand, they are often identified as an invasion force and a source of violent crime. As emerged by the analysis of words' associations in the entire datasets, and confirmed by the qualitative analysis of specific subsets, refugees and asylum seekers, as being subject to a more careful regulation and therefore more entitled to enter Europe, presented a slightly different representation, more empathetic and supportive in comparison to generic migrants. Around migrants' arrival to Europe, their crossing the borders, even on social media (as happens in the public debate), a war on numbers has been fought. Also on this issue, social media acts as an echo chamber of the confusion existing in society regarding the real numbers of migratory flows.

In the analysed tweets, Europe as a political entity is often defined as a weak, silent, role-playing spectator in crisis involving migrants. In this context, both the idea of Europe as a political entity completely helpless and disinterested in migratory events for political ends, and the idea of Europe as an entity that is victim of the events and under attacks, emerges. Discourses that can be based on real facts are therefore mixed with preconceived ideas that do not sink into real information.

In some tweets concerning situations on both borders, Europe is also considered weak in countering the spread of misinformation.

DRAFT

Names of politicians frequently appear in circulating social media discourse. Social media seem to function as echo chambers for politicians' declarations, and this implies a fundamental role that they should play in providing correct and non-distorted information. The reconstruction of what happened on the Greek-Turkish and Polish-Belarusian borders showed that some public actors have been found to push misinformation, by accident or not, spreading confusion, fear, anger, or prejudice.

In social media, we found discourses and accusations very similar to those that emerged and were disseminated in the mainstream media during the days of the two crises. On some specific episodes mentioned in tweets, we found different interpretations regarding the protagonists and the parties involved, demonstrating the various narratives that can circulate and the confusion that can arise about what is true and what is fake. How difficult is identifying misinformation from real information was evidenced not only by the presence of various points of view, but also by the lack of reliable reference points. Indeed, the information circulating in social media in relation to the migration-related events at the borders are often "black box information", without mentioning any verifiable source, often found to be supported by misleading or manipulative images and videos, providing inaccurate messages or marked by simplistic or sensationalist slogans.

Our findings underlined the difficulty in identifying misinformation from real news in social media. The coexistence of so many points of view, often unsupported by reliable data and sources, is an indication of a tendency to provide inaccurate information in social media. Indirectly, this leads to the idea that misinformation, now understood in a very broad sense, can influence the lives of migrants and potential migrants and the perceptions of migrants travelling to Europe, but also of Europeans themselves.

Certainly, the results that emerged from these analyses cannot be considered as definitive. The issue of misinformation can be investigated with further methods and by drawing on further data. Still, these findings are an important indication of how information is developing and circulating on social media, shedding further light on the theme of narratives and discourses, their construction and dissemination, and their potential role on people lives.

DRAFT

## References


[1] Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J.: Sentiment analysis for fake news detection. In: Electronics, 10(11), 1348 (2021)

[2] Donato, K. M., Singh, L., Arab, A., Jacobs, E., & Post, D.: Misinformation about COVID-19 and Venezuelan migration: Trends in Twitter conversation during a pandemic. In: Harvard Data Science Review, 4(1) (2022)

[3] Naldi, M.: A review of sentiment computation methods with R packages. In: arXiv preprint arXiv:1901.08319 (2019)

[4] Ruokolainen, H., Widén, G.: Conceptualising misinformation in the context of asylum seekers. In: Information Processing & Management, 57(3), 102127 (2020)

[5] Saif M., Turney P.: Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, LA, California. (2010)

[6] Stahl, B. C.: On the difference or equality of information, misinformation, and disinformation: A critical research perspective. In: Informing Science, 9, 83 (2006)

[7] Treen, K. M. D. I., Williams, H. T., & O'Neill, S. J.: Online misinformation about climate change. In: Wiley Interdisciplinary Reviews: Climate Change, 11(5), e665 (2020)

[8] Zhou, L., & Zhang, D.: An ontology-supported misinformation model: Toward a digital misinformation library. In: IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(5), 804-813 (2007)

# Graduates' interregional migration in times of crisis: the Italian case

Thaís García-Pereiro[a], Ivano Dileo[b] and Anna Paterno[a]

[a] Università degli Studi di Bari Aldo Moro; C.A. t.garcia.pereiro@uniba.it
[b] Università degli Studi di Napoli Parthenope;

**Abstract**

The economic crisis has affected migration of young adults while experiencing increasing entering and remaining in the labour market. In Italy the share of graduates working in a different territory than those of graduation increased during this period. Previous studies mostly focused on regular business-cycles highlighting the relationship between employment and leaving the place of origin, but studies considering the impact of the Great Recession on interregional migration of graduates is still missing. To fill this gap, the main purpose of this paper is to analyze the effect of economic conditions on graduates' interregional migration, and whether and how these effects change over time in Italy. Our dataset merged two micro-waves of survey data with macrolevel variables, from which we estimated multilevel models for binary responses. Our results shown that the positive effect of regional Youth Unemployment Rates on graduates' likelihood of interregional migration was stronger for those obtaining their degree in 2011 compared to those who graduated in 2007.

**Keywords:** interregional mobility, economic crisis, graduates, multilevel models, Italy.

DRAFT

## 1. Introduction

Research has paid attention to migration patterns of high-skilled individuals founding that interregional movements are more likely to be acted by skilled individuals (Greenwood 1975, Faggian and McCann 2006, 2009, Miguélez and Moreno 2013). Also, the highly-educated are highly-mobile, not only internationally but also internally (Faggian et al. 2017).

Interregional migration of graduates strongly depends on individual factors such as gender (Faggian, McCann and Sheppard 2007), age (Busch and Weigert 2010, Marinelli 2013), geographical origin (Iammarino and Marinelli 2015) as well as on context variables (Florida 2002, Boschma and Fritsch 2009, Nifo and Vecchione 2014, Crescenzi et al. 2017). Interregional migration has been changing so fast recently because of the interfering of uncertain job expectations and socioeconomic challenges (Faggian et al. 2017).

The recession has especially affected migration of the younger age groups, who have experienced high unemployment and difficulties entering the labour market (Bonifazi and Heinz, 2017). In fact, during the crisis, graduates working in a different territory than those of graduation increased (Almalaurea, 2005; 2010; 2012; 2017).

Previous studies mostly focused on regular business cycles highlighting the relationship between the likelihood of being employed after graduation and/or the likelihood of moving out the place of origin (Di Pietro 2005, Dalmazio and De Blasio 2011, Marinelli 2013, Dotti et al. 2013, Etzo 2013) but studies considering the impact that the Great Economic Recession has had on interregional migration of graduates is still missing.

The main purpose of this paper is to analyse the effect of economic conditions on graduates' interregional migration at both micro (individual) and macro (context) levels, and whether and how these effects change over time (pre and during the great recession) in Italy.

## 2. Brief state of the art and hypotheses

At the microlevel, research has demonstrated that employment reduces the propensity toward migration since territorial movements often imply losing an already achieved position in the labor market (Card et al. 2004, Busch and Weigert 2010).

Even though the interregional migration process is a phenomenon mostly related to individual choices, literature has demonstrated that the characteristics of regions of origin and destination also play an important role in explaining the mobility of graduates.

Most research showed that not only individual unemployment, but also structural unemployment is positively associated with out-migration, which is consistent with a push-pull model of migration (Sjastaas 1962, Harris and Todaro 1970).

Evidence focused on interregional migration during recession periods is still limited, and the existent evidence is mixed. On one hand, Davies et al. (2001) find that the relationship between unemployment and interstate migration is stronger during years with a high mean unemployment. Nifo and Vecchione (2014) follow this line showing that as unemployment in the province of origin increases, the probability of migration increases. On the other, some scholars stressed that the disadvantage of migration during the Great Recession for graduates might be more severe (Doherty 2009, Frey 2009), this way increasing obstacles and uncertainty from migratory decisions (postponement).

The direction and magnitude of these relationships in the context of the Great Recession are unclear, but Davies et al. (2001) and Nifo and Vecchione (2014), demonstrated that origin-specific economic characteristics, measured by Youth Unemployment Rates (YUR) of the region of origin, might be especially important for interregional migration.

Following the existing literature and according to available data, we formulate three research hypotheses:

RH1: we expect to find higher likelihoods of performing interregional migration among graduates who were non-employed after graduation, if compared to those who were already employed.

RH2: we suppose graduates' unemployment to be even more important in times of economic recession.

RH3: we suppose that Youth Unemployment Rates will positively influence graduates' interregional migration and, given the impact that the Great Recession might has had on mobility across regions, we hypothesize a change on this influence over time, being Youth Unemployment Rates more important for interregional migration of graduates appertaining to the 2011 cohort (interviewed in 2015) respect to those of the 2007 cohort (interviewed in 2011).

## 3. Data and methods

Individual level microdata were drawn from the last two waves of the survey "Indagine sull'Inserimento Professionale dei Laureati" conducted by the Italian National Statistical Institute (ISTAT) in 2011 and 2015, respectively. Contextual economic macrolevel data, on yearly basis, were drawn from ISTAT regional indicators.

As we analyze interregional migration in Italy during the period between obtaining the degree and the time of surveys (2007-2011 and 2011-2015), we have harmonized and clean survey microdata on both cohorts of graduates into one micro-dataset and merged this information to macro-data on macroeconomic indicators[1] of the region in which the interviewed was living before moving into another region logically matched to each one, nesting graduates in regions (20). Our final sample consisted of 53,880 graduates of the 2007 and 50,106 graduates of the 2011 cohorts.

Our dependent variable measures interregional mobility of graduates in Italy considering graduates who performed interregional migration after graduation vs those who do not moved. The main set of independent variables comprises, a measure indicating if the graduate was unemployed after graduation

[1] Using the mean values of the indicator (Billiet et al. 2014) relative to the five years that preceded graduation for each wave of the survey (between 2002-2006 for graduates interviewed in 2011, graduated in 2007; and 2006-2010 for those interviewed in 2015, graduated in 2011).

DRAFT

as proxy to its own economic conditions before deciding to move to another region, graduates' gender and the survey year (corresponding to a specific cohort of graduates). At the macro/contextual level, our variables of interest were the Youth Unemployment Rates of the region where graduates were living before moving to a different region after obtaining their degree. Model estimations also controls for a set of variables that the literature on graduates' mobility has previously identified as important determinants[2].

Empirical analyses were divided into three stages. On the first, after the null model, the first model includes graduate's unemployment after graduation, to test for graduate's own economic situation; and survey year, to test for changes in graduates' interregional migration over time; and all control variables considered. In the second stage, we added our context variable (Youth Unemployment Rates) to explore its influence on graduates' likelihood of performing interregional migration. Finally, we explore time effects of the economic conditions throughout interactions to a) test if the effect of the employment situation on interregional migration depends on survey year (first-level interaction) and b) test changes over time of the effect of the context on graduates' mobility (cross-level interaction). We have clustered robust standard errors by regions.

## 4. Selected results

The estimates from models of our empirical strategy are showed in Table 1. Estimates of Model 1 show the effect (OR) of individual-level variables on graduates' likelihood of performing interregional migration. Strong significant effects are found for these variables. First, after graduation, unemployed graduates have a 14% higher likelihood of leaving the region where living while attending university than those that were employed. Second, the odds of having performed interregional migration were 38% higher for graduates appertaining to the latest cohort (2011) respect to the previous one (2007).

The second stage is dedicated to test the effects of the Youth Unemployment Rates as contextual predictor on interregional migration. As expected, Youth Unemployment Rates positively influence graduates' interregional migration (Model 3 in Table 1 OR = 1.03, $p<0.001$). Interestingly, the variance of the random effect at the region level diminishes significantly from M1 (0.11) to M2 (0.06), indicating that the inclusion of Youth Unemployment Rates reduces differences on interregional migration across regions in Italy.

The first-level interaction (third stage of empirical analyses) shows that the association of being unemployment after graduation with interregional migration is consistent over time and across cohorts. This means that unemployed graduates were more likely to leave the region in which were living while attending university than those who were employed.

Regarding the final stage, the cross-level interaction shows that graduates' interregional migration has changed during a period of economic recession. The positive effect of regional Youth Unemployment Rates on graduates' likelihood of interregional migration was stronger for those obtaining their degree in 2007 compared to those who graduated in 2011. Thus, the trigger effect of Youth Unemployment Rates on interregional mobility increased over time, being a key factor to understand why graduates moved after graduation during the economic recession.

Table 1: Estimates of a series of binary multilevel regression models for graduates' interregional migration as function variables of interest (micro-macro levels, Odds Ratio).

| **Variables of interest** | **M0 (Null)** | | **M1 (Micro)** | | **M2 (Micro-macro)** | | **M3 (first-level int.)** | | **M4 (cross-level int.)** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | OR | Sign | OR | Sign | OR | Sign | OR | Sign | OR | Sign |
| Unemployed after graduation | - | - | 1,14 | *** | 1,12 | *** | 0,96 | | 1,11 | *** |
| S2015 | - | - | 1,40 | *** | 1,38 | *** | 1,27 | *** | 1,90 | *** |
| YUR5ymean | - | - | - | - | 1,03 | *** | 1,04 | *** | 1,01 | * |

[2] Controls: previous mobility, nationality, more than 30 at graduation (age), married, macro-area of residence, father's and mother's tertiary education, father's and mother's employment status, graduation mark, type of degree (3 years duration), subject of study, graduated later, international mobility program -Erasmus-, have achieved or is currently studying for a Master, have achieved or is currently studying for a PhD, have enrolled or is currently enrolled for a stage.

DRAFT

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *First-level interaction* | | | | | | | | | | | | |
| Unemployed*S2015 | - | - | - | - | - | - | 1,28 | *** | - | - | | |
| *Cross-level interaction* | | | | | | | | | | | | |
| YUR5ymean*S2015 | - | - | - | - | - | - | - | - | 0,99 | *** | | |
| Constant | 0,19 | ** | 0,06 | *** | 0,04 | *** | 0,03 | *** | 0,05 | *** | | |
| Var(region) | 0,13 | *** | 0,11 | *** | 0,06 | *** | 0,05 | *** | 0,48 | *** | | |

[a] * p<0.05; ** p<0.01; *** p<0.001; control variables were included on model estimations but not shown here; Var(region) indicates the variance of the random effect at the second level (region).

## 5. Brief discussion of main findings

This article was aimed at studying the influence of individual and structural economic conditions on graduates' interregional migration while exploring whether and how these effects change over time (pre and during the great recession).

At the micro-level, results show that interregional migration of graduates in Italy is strongly influenced by their individual economic conditions after graduation. Confirming our RH1, likelihoods of interregional migration are higher among non-employed graduates. Regarding changes over time, the role of graduates' unemployment becomes even more important among graduates obtaining their degree in 2011 than among those appertaining to the previous cohort (2007). Thus, as expected (RH2), being unemployed increases interregional migration of graduates in times of economic recession.

Macro-level relations confirm our RH3, since coming from a region with high levels of Youth Unemployment Rates increases the likelihood of changing region of residence after graduation. Its positive effect on interregional migration is stronger over time, as observed for micro-level relations, being more important among graduates obtaining their degree in 2011 than among those of the previous cohort (2007). This points out to the Great Recession as a sort of "trigger event" of graduates' decision to move, increasing their interregional migration likelihood in Italy.

## Selected references


1. Billiet, J., Meuleman, B., & De Witte, H. (2014). The relationship between ethnic threat and economic insecurity in times of economic crisis: Analysis of European Social Survey data. Migration Studies, 2(2), 135-161.
2. Bonifazi, C., & Heins, F. (2017). Internal migration patterns in Italy: Continuity and change before and during the Great Recession. RIEDS, 71(2), 2-10.
3. Boschma R. A., Fritsch M. (2009), Creative Class and Regional Growth: Empirical Evidence from Seven European Countries. Economic Geography 85(4): 391-423.
4. Busch O., Weigert B. (2010), Where have all the graduates gone? Internal cross-state migration of graduates in Germany 1984-2004. The Annals of Regional Science 44(3), 559–572.
5. Ciriaci D. (2014), Does University Quality Influence the Interregional Mobility of Students and Graduates? The Case of Italy, Regional Studies, 48(10), 1592-1608.
6. Crescenzi R., Orru' E., Holman N. 2017, Why do they return? Beyond the economic drivers of graduate return migration. The Annals of Regional Science, 59(3), 603-627.
7. Dalmazzo A., de Blasio G. (2007), Skill-biased agglomeration effects and amenities: theory with an application to italian cities, available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.579.1555&rep=rep1&type=p
8. Di Pietro G. (2005), On Migration and Unemployment: Evidence from Italian graduates. Economic Issues, 10(2).
9. Dotti N.F., Fratesi U., Lenzi C., Percoco M. (2013), Local labour markets and the interregional mobility of Italian university students. Spatial Economic Analysis, 8(4), 443-468.
10. Etzo I. (2010), The determinants of the recent interregional migration flows in Italy: A panel data analysis, Munich Personal RePEc ArchiveMPRA, Online at https://mpra.ub.uni-muenchen.de/26245/1/MPRA_paper_26245.pdf.
11. Faggian A., McCann P. (2006), Human capital flows and regional knowledge assets: A simultaneous equation approach. Oxford Economic Papers, 58(3), 475-500.
12. Faggian A., McCann P. (2009), Human capital, graduate migration and innovation in British regions. Cambridge Journal of Economics, 33(2), 317–333.
13. Faggian A., Rajbhandari I., Dotzel K. R. (2017), The interregional migration of human capital and its regional consequences: a review. Regional Studies, 51(1), 128-143.
14. Florida R. (2002), The Economic Geography of Talent. Annals of the Association of American Geographers 94(2), 743-755.
15. Greenwood M. J. (1975), Research on internal migration in the United States: A survey. Journal of Economic Literature, 13(2), 397-433.
16. Iammarino S., Marinelli E. (2015), Education–Job (Mis)Match and Interregional Migration: Italian University Graduates' Transition to Work, Regional Studies, 49(5), 866-882.
17. Marinelli E. (2013), Sub-national graduate mobility and knowledge flows: an exploratory analysis of onward- and return-migrants in Italy, Regional Studies, 47(10), 1618-1633.


DRAFT

18. Miguélez E., Moreno R. (2013), Research networks and inventors' mobility as drivers of innovation: Evidence from Europe. Regional Studies, 47(10), 1668-1685.
19. Nifo A., Vecchione G. (2014), Do Institutions play a role in skilled migration? The case of Italy. Regional Studies, 48(10), 1628-1649.

DRAFT

# Intentions to stay:
# The experiences of return migrants in Albania

Maria Carella[a], Thaís García-Pereiro[a], Roberta Pace[a] and Anna Paterno[a]

[a] Università degli Studi di Bari Aldo Moro; C.A. t.garcia.pereiro@uniba.it

**Abstract**

The economic crisis intensified both the magnitude and complexity of return migration dynamics, increasing migrants' likelihood of returning to their country of origin mainly as a consequence of unemployment and the worsening of their economic conditions. Despite the importance gained recently, little research has been focused on the relationship between economic consitions and return migration due to the lack of data on the subject. This paper is aimed at contributing to the existent literature on the subject by analysing the case of Albanian migrants who returned between 2009-2013. Data drawn from the 2013 Albanian Survey on Return Migration and Reintegration is used to estimate logistic regression models to predict the intentions to stay in the country of origin after returning. These analyses help us to better understand how return trajectories arise and develop according to the characteristics of returnees, those of their initial migratory project and the situations experienced during and after migration.

**Keywords:** return migration, Albania, migratory intentions, migratory projects, economic conditions.

DRAFT

## 1. Introduction

Return migration is one of the most complex and multifaceted phenomena concerning migratory biographies. The decision behind the return to the country of origin may depend on push factors that have motivated leaving the country of destination (economic circumstances, family or health-related issues) and on pull factors, that may attract migrants back.

The economic crisis started in 2007 intensified both the magnitude and complexity of return migration dynamics. As demonstrated in previous studies (Castles and Vezzoli 2009, Papademetriou and Terrazas 2009, García-Pereiro 2019) the crisis increased migrants' likelihood of returning to their country of origin mainly because of unemployment and worsening economic conditions. Unfortunately, due to the lack of data on the subject and its reduced visibility, little research has been focused on shedding some lights on this relationship.

This paper is aimed at contributing to the existent literature on the subject by analysing the case of Albanian returnees between 2009-2013 to disentangle the decision-making process that have led migrants to return in times of crises. In particular, we analyse returnees' intentions to stay in Albania after returning. We study the effect of variables related to the initial migratory project and situations lived before leaving Albania (before migration) and whether and how its influence change after taking into account situations lived by returnees both in the host country and right after their return.

## 2. Theoretical background and state of the art (in brief)

Theoretical approaches regarding international migration have differently interpreted migrant's return to the country of origin. One set of theories have read return as a positive and successful event which has been part or become part of the migratory project. This is the case of the New Economics of Labour Migration (Stark and Bloom 1985, Taylor 1999, Constant and Massey 2002), Transnationalism (Portes et al. 1999) and return preparation (Cassarino 2004). Other theories, instead, have deemed return as a negative event, signing the failure of the migratory project. Behind this second set is the Neoclassical Theory for which return is the result of the failure of the migratory project (Sjaastad 1962, Todaro 1969, Constant and Massey 2002).

With reference to migration intentions, there is little research that has been focused on whether and how migration intentions change and evolve according to the experiences lived during the development of the migratory project in different territories (country of origin, country of destinations, etc.). Although, we consider that some findings might be somehow interesting to our subject of study. The papers of Balaz et al. (2004) and Khoo et al. (2008) highlight migrants' adaptation experiences in destination places as one of the main drivers behind migration intentions. For Alberts and Hazen (2005) and Lu et al. (2009), these intentions might be strongly dependent on changes in social and personal factors (such as family structure) and mediated by family financial situation. It was also highlighted that migrants living abroad can experienced serious difficulties in accomplishing economic-related migration goals as a result of unemployment, which has been directly related to an increased in the likelihood of remigration (Pecoraro 2012). Instead, being employed or self-employed in the country of destination reduces this likelihood (de Haas and Fokkema 2011, Paparusso and Ambrosetti 2017).

In this paper, deepening the determinants of the emigration and return projects, we analyze the likelihood of being intended to stay in Albania permanently after returning to test whether and how the influence of the initial migratory project changes and if experiences lived in the host country and after return affect such relation. Based on the above-mentioned literature, we are able to build the following research hypotheses:

H1: the return to Albania is a result of the success of the migration project because it is linked to the intention to improve family, individual and working conditions, that, at the time of departure were worse than those at the time of return.

H2: the return to Albania is a result of the failure of the migration project due to the worsening of family, individual working conditions; in this case, the migration experience has not accomplished intentions regarding the improvement of these conditions. H1 and H2 are mutually exclusive competing hypotheses.

H3: the intention to remain in Albania after returning is a result of the modification of the migratory project, due to the experiences and changes that have occurred during this experience.

## 3. Data and methods

Data are drawn from the Survey on Return Migration and Reintegration in Albania conducted by INSTAT and IOM in 2013. This survey has a nationally representative sample size of 1,878 individuals aged 18 years and over who returned to Albania between 2009 and 2013. According to the survey, returnees are those individuals who returned to Albania (permanently or temporarily), after living in another country for at least one year. The dataset includes some information about three stages of migration, asking about situations and conditions experienced before leaving Albania (before migrating), those experienced abroad (while living in the host country) and also post-return (after returning).

We use binary logistic regression models to predict the intentions to permanently stay (intended to stay vs other intention) in the country of origin after returning, which might help us understanding the decision-making process taking place after the return considering as important determinants both main return motivations and initial migratory intentions.

We examine the effects of several predictors that are expected to be associated with the reasons behind these intentions, while controlling for other variables. Model estimations are aimed at getting a deeper understanding of the intention to stay in Albania after returning that, taking part at a well ahead stage of the migratory experience, allowed us to add variables regarding not only the initial migratory project but also experiences lived while settled abroad (financial situation) and after the return (financial situation, having returned with family members). These analyses are aimed at testing H1 vs H2. In order to test for H3, further steps include information regarding experiences lived in the host countries that

DRAFT

were not economic in nature (such as: changes in the marital status and children ever had, having pursued studies or having integration difficulties while living abroad).

Model estimations control for returnees' characteristics (gender, age at migration, marital status, number of dependent family members and educational attainment before migration), year of return, NUTS2 where living before departure, last country of emigration, and length of stay.

In future steps, our attention will be focused on developing robustness checks in which we plan to further disaggregate the intention using multinomial regression models with 3 categories (permanently, temporally, do not know) to disentangle the effect of those whose intentions were not certain. We are also working on the construction of a typology of returnees based on the combination of their initial and final migratory intentions.

## 4. Summarizing results

Table 1 displays the results regarding the intention to stay permanently in Albania after return considering situations experienced by migrants before migrating, during their stay abroad, and after returning. As expected, each step of the migratory experience plays a well differentiated role on returnees' intentions to permanently stay in Albania. The initial migratory project (Table 1, M1) seems to lose relevance after considering experiences lived in the host country (Table 1, M2). In this stage of the migratory project, having experienced a not good-very bad financial situation, changes in family structure and having encountered integration difficulties are of particular importance while controlling for the initial characteristics of the migratory project. In fact, all these covariates exert a positive influence on returnees' likelihood of intending to permanently stay in Albania after having returned.

Although, the effects of two covariates referred to the initial stage (before migrating) remain unchanged: having had prior migration experiences (in terms of international migration) and having planned to migrate alone (without friends or family members).

It is interesting to note that the effects of the covariates illustrating situations experienced by returnees while living in the host country diminished after introducing situations experienced after return. The positive likelihood of being intended to stay in Albania decreases the most when considering non-economic situations (integration difficulties or changes in the family structure). This is not the case of returnees who pursued their studies abroad. The influence of economic conditions, measured thru the financial situation, increases when passing from M2 to M3 (Table 1). Returnees who experienced a not good-very bad financial situation were 19% more likely to be intended to permanently stay when considering situations experienced before migrating and while living abroad (M2), and this likelihood increased to 29% after adding situations experienced after return (M3).

For staying intentions, remains even more important the effect of variables related to situations experienced by return migrants after returning to Albania (Table 1, M3). However, there are some differences between economic and non-economic determinants that deserve to be highlighted. About economic conditions experienced after having return, we found that the likelihood of being intended to stay of return migrants whose financial situation remained unchanged or decrease upon returning (if compared to the situation experienced in the host country) was 17% lower than the one declared by those whose financial situation improved. Then emerges the need to consider the impact that non-economic situations have on their own intentions to stay. Results show that returnees that were back to Albania without their family members were 48% less likely to be intended to stay if compared to those who returned together whit their family.

Table 1: Results from binary logistic regression on the likelihood of being intended to stay in Albania permanently after returning in M1: base model, M2: base model + situations lived abroad, and M3: base model + situations lived abroad + situations lived after the return (Odds Ratio).

| **Main variables of interest** [a] | **M1** | | **M2** | | **M3** | |
|---|---|---|---|---|---|---|
| | OR | Sign. | OR | Sign. | OR | Sign. |
| **Before migration** | | | | | | |
| Travel with legal docs | 0.97 | * | 0.95 | * | 1.04 | |
| *Not good-very bad financial situation* | *1.04* | | *1.03* | | *1.13* | |
| Planned to travel alone | 1.34 | *** | 1.39 | *** | 1.81 | *** |
| Prior migration experiences | 0.57 | *** | 0.54 | *** | 0.53 | *** |
| **In the host country** | | | | | | |

DRAFT

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Not good-very bad financial situation* | - | | - | | *1.19* | ** | *1.29* | ** |
| Pursued studies | - | | - | | 1.21 | * | 1.20 | |
| Encountered integration difficulties | - | | - | | 1.33 | *** | 1.14 | ** |
| Marital status change or having children | - | | - | | 1.36 | * | 1.23 | ** |
| **After return** | | | | | | | | |
| Back without with family members | - | | - | | - | - | 0.52 | *** |
| *Unchanged/decreased financial situation* | - | | - | | - | - | *0.83* | * |
| *N* | | *1,878* | | *1,878* | | | *1,878* | |
| *Pseudo R2* | | *0.17* | | *0.20* | | | *0.38* | |
| *Log likelihood* | | *-1187.29* | | *-1171.9* | | | *-1048.8* | |

[a] Models control for returnees' characteristics (age at migration, gender, marital status, number of children and educational attainment), NUTS2 of residence before leaving Albania, host country and year of return.

Source: own elaboration, 2013 Return Survey microdata.

## 5. Concluding remarks

The central purpose of this article was to study the intentions to permanently stay in Albania after returning of return migrants while considering the initial characteristics of their migratory projects and the economic and non-economic situations experienced both in the host country and right after their return.

Following theories on return migration, we hypostasized that returning to Albania may be interpreted or as a result of the success (H1) or as a result of the failure (H2) of the migration project. These are competing hypotheses that were built comparing situations experienced by return migrants before migration, while living abroad and after returning. In the first case, the success is proxied through the improvement of individual or family economic conditions. In the second, the migratory experience has not significantly improved such conditions. Here, findings pointed to the confirmation of H2 giving that the influence economic conditions remained of crucial relevance and positive intentions to stay are strongly dependent on returnees having experienced a poor financial situation while living in the host country.

Going further, our third hypothesis (H3) was aimed at recognizing how relevant non-economic conditions are in shaping future intentions to stay rather than to leave the country again. We found support for it, calling the attention to the intrinsic dynamic nature of the migratory project which is constantly evolving over time and across space. The most important role on post-return migratory intentions is played by experiences and changes that have occurred during this experience, which are non-economic in nature.

DRAFT

## References


1. Alberts, H.C., Hazen, H.D.: "There are always two voices...": International students' intentions to stay in the United States or return to their home countries. International Migration {43(3)}, 131-152 (2005)
2. Balaz, V., Williams, M., Kollar, D.: Temporary versus permanent youth brain drain: Economic implications. International Migration, {42(4)}, 3-32. (2004)
3. Cassarino, J. P.: Theorising return migration: The conceptual approach to return migrants revisited. International Journal on Multicultural Societies (IJMS), {6(2)}, 253-279 (2004)
4. Castles, S., Vezzoli, S.: The global economic crisis and migration: temporary interruption or structural change? Paradigmes: economia productiva i coneixement (2009)
5. Constant, A., Massey, D. S.: Return migration by German guestworkers: Neoclassical versus new economic theories. International migration, {40(4)}, 5-38 (2002)
6. De Haas, H., Fokkema, T.: The effects of integration and transnational ties on international return migration intentions. Demographic research, {25}, 755-782 (2011)
7. García-Pereiro, T.: Clustering reasons for returning: An overview of return migration in Albania. Journal of International Migration and Integration, {20(2)}, 361-374 (2019)
8. Khoo, S. E., Hugo, G., Mcdonald, P.: Which skilled temporary migrants become permanent residents and why? International Migration Review, {42(1)}, 193–226 (2008)
9. Lu, Y., Zong, L., Schissel, B.: To stay or return: Migration intentions of students from People's Republic of China in Saskatchewan, Canada. Journal of International Migration and Integration/Revue, {10(3)}, 283-310 (2009)
10. Papademetriou, D.G., Terrazas, A.: Immigrants and the current economic crisis: Research Evidence. Policy Challenges and Implications. Migration Policy Institute, Washington DC. (2009).
11. Paparusso, A., Ambrosetti, E.: To stay or to return? Return migration intentions of Moroccans in Italy. International Migration, {55(6)}, 137-155 (2017)

12. Pecoraro, M.: Rester ou partir: les déterminants de l'émigration hors de Suisse. In: Wanner, P. (eds.) La démographie des étrangers en Suisse, pp. 141-155. Seismo, Genève (2012)
13. Portes, A., Guarnizo, L. E., Landolt, P.: The study of transnationalism: pitfalls and promise of an emergent research field. Ethnic and Racial Studies, {22(2)}, pp. 217-237 (1999)
14. Sjaastad, L.A.: The Costs and Returns of Human Migration. Journal of Political Economy, {70}, pp. 80-93 (1962)
15. Stark, O., Bloom, D.: The New Economics of Labor Migration. American Economic Review, {75}, pp. 173-178 (1985)
16. Taylor, E. J.: The new economics of labour migration and the role of remittances in the migration process. International migration, {37(1)}, pp. 63-88 (1999)
17. Todaro, M.P.: A Model of Labor Migration and Urban Unemployment in Less-developed Countries, American Economic Review, {59}, pp. 138-148 (1969)

DRAFT

# Return migration to home country: a systematic literature review with text mining and topic modelling

Cecilia Fortunato[a], Andrea Iacobucci[b], and Elena Ambrosetti[a]

[a] Sapienza University of Rome; cecilia.fortunato@uniroma1.it, elena.ambrosetti@uniroma1.it

[b] Regione Emilia-Romagna; iacobucciandrea@gmail.com

DRAFT

**Abstract**

A crucial (and less developed) part of migration studies is the exploration of migrant's further mobility and the intention of return to home country at some point in life. Knowing who, why and when returns matters for both the host and the home country. Very few studies have focused on return in a wider perspective, adopting a comparative approach. The present study aims at providing a systematic review of peer-reviewed literature indexed in Scopus database, to understand how return has been dealt with by researchers. The main objectives are: collecting and synthetizing previous studies; comparing approaches, conceptualizations of return, methods and variable of interest. A bibliometric analysis on metadata and content analysis based on text mining and topic modelling techniques has been conducted on a sample of approximately 3,000 publications. With our contribution, we expect to implement a baseline for theoretical development and empirical research, presenting an overview on the evolution of trend topics, with regional and temporal patterns of research focus and identifying knowledge gaps in literature.

**Key words:** return migration, remigration, home country, systematic literature review, topic modelling

## 1. Introduction

Migration studies have developed rapidly in recent decades, in terms of size, methods used, interdisciplinarity, heterogeneity of theoretical frameworks and internationalization of research groups (Pisarevskaya et al., 2020; Vargas-Silva, 2012; Bonifazi and Strozza, 2006; King and Skeldon, 2010; Massey et al., 1993).

Migration flows and many aspects of the migration cycle have been widely discussed in the literature, relating to a growing diversity of migratory categories (from refugees and forced migrants to economic or climate migrants) and specific case-studies at national and global level (Abel, 2018; Castels and Miller,2014).

Researchers have concentrated their efforts in analyzing why people decide to migrate. But a crucial (and poorly developed) part of migration knowledge is the exploration of migrant's further mobility over the life course. A fundamental theoretical and practical question should be: what are future plans and prospects of migrants? What factors influence their decision for permanent stay in host country, for onward migration to a third destination, or for the return to home country, at some point in life? Awareness on the whole migratory cycle is scarce. Yet, knowing who, why and when returns, estimating future migration scenarios and forecasting returns, matters for both the host and the home country. And this aspect could be even more relevant in times of crisis, as shown by the COVID-19 pandemic, when millions of migrants have been returned to their countries of origin, we don't know if permanently, and large numbers of migrants have found themselves "stranded abroad and in need of assistance" (Le Coz and Newland, 2021).

Modern fluid societies, characterized by the globalization of movements, communications, technologies and information, make the ground extremely rough for analysis, because of the volatility of events, the heterogeneity in biographies and experiences and the increasing transnationalism of migrants. Return decision making is a complex process (Carling and Pettersen, 2014; Waldorf, 1995): it is influenced by a variety of

factors related to conditions both in origin and destination countries and motivated by needs at individual, household and social level (Krasniqi and Williams, 2018; Paparusso and Ambrosetti, 2017). Migrants' characteristics and situational conditions affect return probabilities and reintegration prospects of returnees. Furthermore, return to home country is particularly difficult to measure because of the lack of reliable and consistent register data and large-scale survey data, as it often goes spontaneous and unrecorded.

We know from literature and international reports that 1/4 of total migratory events are estimated to be returns; up to 50% of immigrants leave the host country within 5 years (return home or secondary emigration) and return migrant's selection tends to be the reverse of the initial selection process for migration (Azose and Raftery 2019; Abel 2018; OECD, 2017; Åkesson and Baaz, 2015).

However, very few studies have focused on return in a wider perspective, adopting a comprehensive and comparative approach (see Cassarino, 2015; Mohamed and Abdul-Talib, 2020)

## 2. Main objectives and Research Questions

Knowledge production within the field of migration is accelerating at a tremendous speed, while risking at the same time to remain over-specialized, fragmented and inconsistent (Pisarevskaya, 2020; Snyder, 2019; Denyer and Tranfield, 2009; Kitchenham, 2004; Tranfield, et al., 2003; Webster and Watson, 2002,).

The present study aimed at providing a theoretical basis necessary for the further analysis of factors driving intentions and realizations of migrant's homeland return, and the relationship between initial migration drivers and return intentions. Mapping academic landscape is an important step in addressing new research questions, proposing new conceptual framework, facilitating theory development and identifying knowledge gaps and topics that need further research.

The systematic review of peer-reviewed literature indexed on citation databases is a powerful tool for understanding how return has been dealt with by researchers. What has been studied so far may shed light on what has been relevant in time, assuming that fundamental research questions arise with the development of societies and follow the evolution of phenomena under study (Snyder, 2019).

The main objective of this study was collecting and synthetizing previous studies, comparing approaches, conceptualizations of return, methods and variable of interest, integrating findings and perspectives, testing return theories against international migration theories and comparing the assumptions on which they rest. We aimed at providing a detailed overview of the existing literature through a comprehensive search in publication's repositories, using relevant keywords.

The following questions guided our study:

**RQ1:** How the studies on return migration developed from 1960 to 2020?

**RQ2:** How has return been conceptualized in terms of topical focus?

**RQ3:** What theoretical framework and motivational factors have been taken into account?

**RQ4:** Is the evolution of "return theories" following the evolution of "migration theories"?

DRAFT

## 3. Data and Method

Bibliometric analysis and content analysis based on text mining and topic modelling algorithm have been applied to publications, gathered from Sci-Verse Scopus.

The first step was building a valid search query (combinations of key terms and logical operators) that would retrieve (in title, author's keywords or abstract) as many documents as possible with minimum irrelevant results. When possible, we gathered publication and metadata via API, alternatively, we gathered data with search-by-string method (Belter, 2020; Aria and Cuccurullo, 2017-2018). We applied a massive data cleaning procedure to exclude out of scope articles and to clean text with lemmatization and tokenization techniques. For data analysis, we selected specific packages of the statistical software R: dplyr, ldatuning, topicmodels, rscopus, bibliometrix.

The descriptive analysis of the collection and the bibliometric analysis of meta-data aimed at discovering the temporal and spatial distribution of publications, evaluating the impact of a publication in the scientific community, identifying important theoretical contributions and mapping the evolution in international research cooperation networks.

The content analysis of title-keywords-abstract with text mining and topic modelling techniques identified topics or topic clusters that figure centrally in the specific textual landscape. We aimed at

discovering patterns and regularities within the corpus of texts, elaborating classification and categorization of topics, evaluating correlations among topics and following the evolution of trend topics. For text analysis we use the TF-IDF technique (term frequency-inverse document frequency), assessing how relevant a word is to a document in a collection of documents, on the base of how many times the word occurs in the document; and the inverse document frequency of the word across a set of documents. We considered the absolute frequency of a term in the whole corpus of text (frequency) as well as the numbers of documents containing the specific token (count).

For topic modelling, Latent Dirichlet Allocation model has been selected (Farren, 2019). LDA model is a Bayesian probabilistic model of latent topics from the contents of abstracts, based on the assumptions that each document in a corpus discusses multiple topics in differing proportions. A topic can be considered as a collection of words ordered by their probability of occurrence. The topic structure in the corpus of words is hidden and this method seeks to provide an ideal number of topics (k), a matrix with per-topic word probabilities and a matrix with per-document topic proportions. As a result, it was possible to summarize and label the topic in topic clusters and their probability of occurrence. Two metrics have been defined to summarize topic distribution across documents and classify articles: the topic presence, showing the number of document in which the topic is present, equal to 1 if the gamma score of LDA model is above the per topic average of gamma (and 0 otherwise); and the topic dominance, showing the number of document in which the topic is prevailing over other topics, selecting the topic with maximum gamma score for each document.

## 4. Main results

Our final sample included almost 3.000 documents published in English between 1960 and 2020 by more than 1.400 publishers.

The great majority of the publications focusing on return were articles published in specialized journals (2469) and only the 12.9% of them were published in open access. The first publishers, in terms of number of publications and citation obtained were migration's specialized scientific journals such as International Migration and the Journal of Ethnic and Migration Studies; the only regional journal among the top publishers was the Asian and Pacific Migration Journal. Among the top 15 publishers, two were journals specialized on refugees and forced migration, a fundamental component of the migration phenomenon (Journal of Refugee Studies and the Refugee Survey Quarterly).

Our results from the descriptive and bibliometric analysis of meta-data showed that first return migration studies can be traced back to the beginning of the 1960s, with an exponential growth of scientific production on the topic starting from the late seventies. However, a proper corpus of studies took shape only in the 1980s, with the politization of international migration movement and the growing importance of the discourse on (economic) development in origin countries, both in academia and in the institutional and public debate (Cassarino, 2015). The great acceleration of the topic's relevance started at the beginning of the new millennium, with researchers looking at the effects of the global economic crises on millions of migrants living in recessing high industrialized countries. Indeed, 60% of total studies have been published between 2010 and 2020.

Looking at the spatial distribution of publications, we highlighted the dominant role of receiving countries involved in the "knowledge production" on return. The 50% of total documents have been published in only 5 countries, namely United States, United Kingdom, Canada, Germany and Australia.

We than analysed the textual landscape presented in publications retrieved, mining the text of title, abstract and keywords. In the overall corpus of text, most frequent tokens present a variegated nature, including all the topical aspects of different migration theories. Among the top frequent tokens, a major role can be attributed to economic, push and pull factors and labour migration theories, followed by systemic and macrolevel theories and transnationalism and social network theories.

We then analysed most frequent tokens divided in three subperiods (1960-1980, 1981-2000 and 2001-2020), showing a shift in scientific focus over the last 40 years, with economic and labor market associated keywords leaving room for more subjective aspects of migration.

The first period considered (1960-1980) has been dominated by the economic paradigm. In this perspective, the return has been considered as the "logical outcome of a calculated strategy" (Cassarino, 2015), underlining the importance of wage differentials, accumulations of savings and, more important, remittances as explanatory factors of the return decision. In the second period considered (1981-2000),

DRAFT

words related to the economic perspective are still relevant, but a growing attention has been paid to the macro-level influence and constraints on return migration, especially related to the increasing number of conflicts and humanitarian crises around the world during the '90s. In that period, experts and researchers from different discipline implemented a series of new empirical tools ("model") to study migration trends, and most of them started to take into account meso-level factors (group and community) behind migration and return, such as the new role of family in migration strategies and the increasing number of women among international migrants. In the beginning of the new millennium (2001-2020), the phenomenon of return gained another level of interest as well as complexity of analytical frameworks. The increasing globalization of information, facilitated by communication and transportation, the cyclical economic crisis and the higher mobility of high skilled workers, led the researcher to focus on social capital and transnational practices. Many studies still point out that economic factors play a crucial role in homeland return's decision, but there is a slight cultural shift in the economic paradigm as the focus has turned into a matter of "aspirations" for an upward-mobility. As it was predictable, in the last twenty years the discourse on return has been dominated by an increasing attention on the refugee return to post-conflict areas, and it is worth mentioning the growing importance of gender perspective in research landscape.

Topic models applied to our textual data with the LDA algorithm provided the 17 topics that best describe our data. Each topic is defined by a string of words that are semantically connected, according to LDA model. We analysed 10 selected words among the top 30 words defining the topic and assigned a label to each topic. We than analysed relevance of each topic in the whole corpus.

In the overall period, the dominant topic focuses on cultural, ethnic and identity issues, defined by words related to diasporic communities and the difficult compromise between the need of preserving the own cultural identity and the need of getting integrated in the host society. The second dominant topic is the classical economic paradigm anticipated by the top frequent words, concerning skills, labour market opportunities and wage differentials, highlighting the rational benefit-cost perspective upon the possibility of return based on migrants' selection, skills, education, wage and returnees' prospects once back to their origin country (for temporary or permanent return). The third dominant topic is the topic on refugee and post-conflict return. From our results, lower attention has been paid to the topic of circular mobility in conflict areas, a topic mainly defined by a set of words related to specific territory and repeated border crossing between two "national states". Also, the topic of European migration flows, mainly related to internal temporary movement in Europe has attracted a lower interest compared to other topics.

Dividing the whole period in the above mentioned three subperiods, we presented our results starting with topics that have gained importance in researchers' works in the last 60 years. The most relevant topic in 2000-2020 concerns "Cultural, ethnic and political identity", marginal in the first period (dominant in only 3% of documents) but gaining importance in the last twenty years of the last century and dominant in almost the 16% of total document published in the last period considered. The second topic in terms of relevance in the last period (2001- 2020) is related to "Refugee and post-conflict return", experiencing a constant increase in relevance and dominant in the 13% of publications of recent years. As already reported, the topic of "Skills, labour market opportunities and wage differentials" has been relevant in the whole scientific production on return migration, but at the same pace of the growing importance of the neoclassical economic paradigm proposed by Todaro in 1969, it reached the maximum interest in the eighties, being dominant in the 15% of total publications of the period. It can be considered one of the most important topical focus also in the last 20 years, being dominant in the 13% of scientific publications. As anticipated by the top tokens cited in publications in the last period considered, the topic on "Gender, family and vulnerabilities" has gained growing importance in research already in the central period, following an increasing female migration flow, primarily from former Soviet Union Republics, and becoming crucial with the increasing process of family reunifications in "mature" host countries and the aging process and the subsequent higher need of caregivers in high income countries. The topic defined by words expressing a wider "Systemic approach and multiple focus", is dominant in a growing number of publications in the whole period considered, maintaining however a stable proportion of 10% of publications in which the topic is dominant in the three different periods considered.

As for topic that have experienced a vigorous decline in the interest received from the scientific community, we first mention the topic on "South-North migration, Africa and demographic imbalances", which appears to be the most discussed between 1960 and 1980, dominant in almost the 28% of researches published in that period. In recent decades, instead, the topic has been the focal point only in

DRAFT

the 3% of publications. Such decline in relevance could be partially explained by the persistent lack of consistent data in African countries, discouraging researchers always seeking for large-scale data to conduct robust quantitative analysis. Another reason could be the growing complexity of migration phenomena, the consequent questioning of the "South-North" migration model and the growing attention referred to emergent "South-South" mobility patterns. The topic identified as "Rural-urban mobility" and mainly referring to the context of rural Asian countries, has also experienced a sustained and constant decline, being dominant in 15% of publications in the sixties and only in the 7% in the beginning of the millennium. Same destiny has the topic on "Economic-differential and cost-opportunities assessment", with declining dominance from 10% to 4% of total publications in the whole period. This lower interest in a topic defined by words that are key points of the New Economics of Labour Migration approach (NELM), can be interpreted as an overcome of the classical view of return migration as the logical outcome of a "calculated strategy" in favour of a greater attention toward new paradigms focused on identity and cultural issues, transnational practices and social network effects on return's decision-making process. An expected result is the decline of interest in the topic related to "EU migration flows and temporary/permanent stay", dominant in only 1% of recent publications. The first 10 selected words suggested a focus on mobility occurred in the last half of nineteenth century, presenting very different characteristics from nowadays migration patterns in terms of migrants' selection, education and skills, information and opportunities, migration policies and mobility in Europe, social networks and family settlement. However, understanding what happened to guest workers migrated to European countries in that period could be of great importance to understand present patterns of aspirations to return.

The topic of "Education/qualification opportunities and high skilled mobility", attracted a small interest in the eighties, probably because of the average higher proportion of unskilled migrants, but regained importance in last 20 years. The same trend is shown for the topic of "Eastern to Northern Europe migration", probably related to old migration pattern, not highly representative of current situation. And also, the decline in relevance of the topic on "Optimal strategies, return/ onward mobility" can be explained with an overcome of the neoclassical and the NELM paradigms.

Lastly, marginal topics in our dataset, meaning dominant in less than 3% of the publications from the first to the last period are "Circular mobility in conflict areas", "Southern Asia-Pacific labor migration", "Socialization, linguistic skills and integration", (Higher education mobility and opportunities".

## 5. Conclusions and further development

Topic modelling provides an affordable way to classify scientific papers and discover hidden characteristics of documents in a corpus of text. However, natural language processing still needs expert's supervision and interpretation and, in some cases, the full interpretability of topics can be challenging. Also, the way of indexing publications with meta-data and keywords, the query selected for data retrieving, filters applied in data cleaning process and the authors' classification procedure are, intuitively, of fundamental importance for systematic review results.

We found consistent results with what has been highlighted by the scientific literature in the field and the evolution of international migration theories, keeping in mind that ongoing processes of social transformation will perpetually create the need for theoretical innovation (de Haas 2021).

Our results clearly show the dominance of receiving countries in the knowledge's production of return migration, stressing the need of sending countries perspectives on return prospects of reintegration and effective returnees.

Moreover, the number of topics identified and the evolution of their relevance in time, suggests an over-fragmentation of the field and the need for a more comprehensive, wider social-scientific perspective, supported by an historical-comparative approach, which is completely absent in our dataset of publications.

The aspiration-capability framework and the conceptualization of return capability within the interaction between structural constraint (macro and meso-level factors) and agency are "great absents" in the scientific field.

With the increasing role of both emigration and immigration states in controlling and managing migration and return, as suggested by literature (de Haas et al. 2019, Hollifield et al. 2014, Waldinger 2015, Cassarino 2004) we suggest the relevance of the focus on how changing political contexts influence the possibility of return and reintegration experiences of returnees.

Our results also suggested that lower attention has been paid so far to the development of concept of

DRAFT

return in terms of intrinsic aspirations (subjective: identity, perceptions, networks and ties), instrumental aspirations (objective: costs/opportunities, policies and external factors) and effective capabilities, as mainly stressed by Carling and others (Carling, 2002; Carling and Petersen, 2014; Carling and Schewel, 2018; de Haas, 2014).

We finally reported the need of more attention on emerging migration patterns, such as the possibility of return for climate and environmental refugees, as well as the relevant question on what the age and gender composition of migrants' population can tell us about their probability of return, a topic that can be further investigated with the exploitation of decomposition and microsimulations techniques.

## References

## References

[1] Abel G.J.: Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015. In: International Migration, Volume: 52 issue: 3, pp. 809-852 (2018)
[2] Akesson, L., Baaz, M.E.: Africa's Return Migrants: The New Developers? In: Zed Books Ltd. (2015)
[3] Aria M: Bibliometrix: Data Importing and Converting, https://www.bibliometrix.org/vignettes/Data-Importing-and-Converting.html (2018)
[4] Aria, M., Cuccurullo, C.: Bibliometrix: An R-tool for comprehensive science mapping analysis. In: Journal of Informetrics, 11(4), pp 959-975, Elsevier (2017)
[5] Aria M., Cuccurullo C.: Science Mapping Analysis with bibliometrix R-package, https://bibliometrix.org/documents/bibliometrix_Report.html (2018)
[6] Azose, J.J., Raftery, A. E.: Estimation of emigration, return migration, and transit migration between all pairs of countries. In: Proceedings of the National Academy of Sciences Jan 2019, 116 (1) 116-122 (2019)
[7] Belter C.: Text mining in R https://github.com/christopherBelter/textmining (2020)
[8] Bonifazi C., Strozza S.: Conceptual Framework and Data Collection in International Migration. In: Caselli G., Vallin J., Wunsch G. (eds.), Demography: Analysis and Synthesis. A Treatise in Population, Volume IV, Elsevier Inc., USA, 2006, pp. 537-554 (2006)
[9] Carling, J.: Migration in the age of involuntary immobility: Theoretical reflections and Cape Verdean experiences, Journal of Ethnic and Migration Studies, 28:1, 5-42, DOI: 10.1080/13691830120103912 (2002)
[10] Carling, J., Pettersen, S.V.: Return migration intentions in the integration–transnationalism matrix. In: International Migration, Vol. 52 No. 6, pp. 13-30 (2014)
[11] Carling, J., Schewel, K.: Revisiting aspiration and ability in international migration. In: Journal of Ethnic and Migration Studies, 44:6, 945-963, DOI: 10.1080/1369183X.2017.1384146 (2018)
[12] Cassarino, J.P.: Theorising return migration: the conceptual approach to return migrants revisited. In: International Journal on Multicultural Societies, Vol. 6 No. 2, pp. 253-279 (2015)
[13] Castles, S., Miller, M. J.:The Age of Migration: International Population Movements in the Modern World, 5th edn. Basingstoke: Palgrave Macmillan (2014)
[14] Denyer, D., Tranfield, D.: Producing a systematic review. In: Buchanan, D.A. and Bryman, A. (Eds), The SAGE Handbook of Organizational Research Methods, pp. 671-689. Entrepreneurship: Theory and Practice, Vol. 20 No. 1, pp. 55-72 (2009)
[15] Farren T.: Beginner's Guide to LDA Topic Modelling with R: Identifying topics within unstructured text, https://towardsdatascience.com/beginners-guideto lda-topic-modelling-with-r-e57a5a8e7a25 (2019)
[16] King, R. Skeldon, R.: Mind the Gap: Integrating Approaches to Internal and International Migration. In: Journal of Ethnic and Migration Studies, 36/10: 1619–46 (2010)
[17] Kitchenham, B.: Procedures for performing systematic reviews, Keele, UK, Keele University, Vol. 33 No. 2004, pp. 1-26 (2004)
[18] Krasniqi, B.A., Williams, N.: Migration and intention to return: entrepreneurial intentions of the diaspora in post-conflict economies. In: PostCommunist Economies, Vol. 31 No. 4, pp. 464-483 (2018)
[19] Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E.: Theories of International Migration: A Review and Appraisal. In: Population and Development Review, 19/3: 431–66 (1993)
[20] Mohamed, M. , Abdul-Talib, A.N.: Push–pull factors influencing international return migration intentions: a systematic literature review. In: Journal of Enterprising Communities: People and Places in the Global Economy, Vol. 14 No. 2, pp. 231-246 (2020)
[21] OECD: International Migration Outlook 2017, OECD Publishing, Paris, https://doi.org/10.1787/migr_outlook-2017-en (2017)
[22] Paparusso, A., Ambrosetti, E.: To stay or to return? Return migration intentions of Moroccans in Italy. In: International Migration, Vol. 55 No. 6, pp. 137-155 (2017)

DRAFT

[23] Pisarevskaya, A., Levy, N., Scholten, P., Jansen, J.: Mapping migration studies: An empirical analysis of the coming of age of a research field. In: Migration Studies, Volume 8, Issue 3, pp. 455–481 (2020)
[24] Snyder, H.: Literature review as a research methodology: An overview and guidelines. In: Journal of Business Research, Elsevier, No. 104, pp. 333-339 (2019)
[25] Tranfield, D., Denyer, D., Smart, P.: Towards a methodology for developing evidence in formed management knowledge by means of systematic review. In: British Journal of Management, Vol. 14 No. 3, pp. 207-222 (2003)
[26] Vargas-Silva, C.: Handbook of Research Methods in Migration. Cheltenham: Edward Elgar (2012)
[27] Waldorf, B.: Determinants of international return migration intentions. In: The Professional Geographer, Vol. 47 No. 2, pp. 125-136 (1995)
[28] Webster, J., Watson, R.T.: Analysing the past to prepare for the future: writing a literature review. In: MIS Quarterly, Vol. 26 No. 2, pp. 13-2 (2002)

DRAFT

# The allocation of time within native and foreign couples living in Italy

Giovanni Busetta[a], Maria Gabriella Campolo[a], and Antonino Di Pino Incognito[a]

[a]Department of Economics, University of Messina, Via Dei Verdi 75, 98122 Messina; gbusetta@unime.it, mgcampolo@unime.it,dipino@unime.it

DRAFT

**Abstract**

The empirical analysis of the intra-household allocation of time suggests that female labour market participation in Italy is lower than in other European countries. Conversely, the contribution in terms of domestic work and childcare by Italian men is very low. Studies found that the gender gap in time allocation strongly depends on the traditional social norms, and on the cultural background of the subjects. The aim of this paper is to investigate if people resident in Italy but born in other countries, suffer from the same decision-making process of Italian families, involving a gendered division of roles in the intra-household allocation of time. Using the Italian Time Use survey 2013, a Seemingly Unrelated Regression Estimation model of paid and unpaid work, leisure and personal care, is estimated to compare couples in which both partners have Italian nationality with couples in which at least one partner is an immigrant.

**Key words:** allocation of time, households, natives and immigrants, SURE

## 1. Introduction

The empirical analysis of the intra-household allocation of time has had a remarkable growth in re-cent years with the increase of interdisciplinary studies that affect the economic, demographic, social and statistical area. Studying the use of time and, therefore, different ways of allocation of time, we analyse a plurality of phenomena which are interconnected with each other. National and international literature suggests that Italy is characterized by a low rate of female participation in the labour market compared to most of the European countries (Mangiavacchi and Rapallini, 2012), especially considering the subsample of mothers with young children (Anxo et. al, 2011).

Several studies have confirmed that the birth of a child revolutionizes the allocation of time with-in the family, with an increase of the time dedicated to paid work by fathers and a decrease in that of mothers who, at the same time, increase in domestic and childcare activities (Anxo et al., 2011; Bloemen et al., 2010; Campolo and Di Pino, 2012; Campolo et al., 2016; Del Boca and Vuri, 2007).

Failure or difficult to access to the labour market, as well as the difficult reintegration after the childbirth, makes the women economically disadvantaged and most vulnerable to poverty. Only 30% of them, in fact, continue to work after the childbirth (Del Boca and Rosina, 2009). In addition, the unequal distribution of domestic work and the division of gender roles begins in the family of origin, and it passes on from generation to generation (Del Boca et al., 2012). Intra-household allocation of time is usually affected not only by the bargaining power of the partners, but also by social contexts and norms, cultural and family background. For this reason, one could wonder whether immigrants residents in Italy suffer either from gender roles imposed by Italian families, or from the ones of their country of origin.

The aim of this work is to analyse whether factors characterizing the intra-household division of labour of Italian couples (i.e., couple in which both partners are Italians), are the same explaining the reasons underlying gender differentiations within foreign couples (i.e. couples in which both partners are foreigners), or mixed couples (i.e. only one partner is Italian). Deeply, we want to understand whether standard theoretical models predicting immigrant wives' housework time are comparable with to the one operating in native-born and mixed couples, as immigrants usually tend to show more traditional gendered division of labour (Blau et al. 2020)

Throughout the specification of a Seemingly Unrelated Regression Estimation Model (SURE), we jointly estimated the time spent by men and women in different kinds of activities. To do so, we used data coming from the last survey of ISTAT "Use of Time 2013". To do so, we distinguished data in two subsamples: one formed only by Italian couples and the other formed only by foreign-born or mixed couples. The results obtained from an analysis of time use may be supportive not only of the development and support of policies for family/work conciliation, removing obstacles that makes women and foreign-born economically disadvantaged, but also of policies to encourage other kind of activities as leisure time. As shown by Hamermesh and Trejo (2013), the leisure time, in fact, includes both socializing activities, carried out in company and non-socializing activities. To this regard, a (provisional) result obtained from our analysis shows how the female partners of foreign and mixed couples reduce the frequency of contacts with friends to devote more time to housework.

In particular, the socializing activities allow a social integration of the immigrant into the social mainstream of the country in which he lives (Depalo et al., 2007). To incentive foreigners to spend their time socializing could be a way to make them less alone and less exposed to the risk of poverty in a foreign country, as "the allocation and efficiency of non-working time may now be more important to economic welfare than that of working time" (Becker, 1965).

## 2. Data and methods

The study sample was drawn from the 2013 Use of Time Survey carried out by Istat with the time diary method. The diary method is an important instrument that allows to investigate how people organize their days and the relationship between the different timing of the same family's members with an extremely high level of accuracy.

The sample is formed by 12727 Italian couples and 1183 foreign couples in working age.

The aim of this analysis is to compare how Italians and foreigners manage their daily lives and examine whether there are significant differences. Unfortunately, we cannot focus the analysis on the country of origin, because in the Use of Time survey, the only information that can identify ethnic origin is the nationality (Italian or foreign). From the answer to this question, we can identify couples in which both partners are Italians (IC = Italian Couples), both or only one of them is foreigner (FC = Foreign Couples). We performed our analysis not only at the individual level but also at familiar one, as our final goal is the joint estimation of the time spent by the partners in paid and unpaid work, personal care and leisure, considering the latter both as aggregating and not aggregating activity.

The different activities have been grouped into the five following sectors: 1) Paid work (including main and secondary work, break, job search); 2) unpaid work (including housework; family care: washing, cooking, cleaning, gardening, pet care, shopping, care and support to adults in the family, repairs; and child care: physical care, help with homework, play and accompaniment); 3) Personal care (sleeping, eating, or other activities; 4) Media Leisure (TV, radio, reading books and magazines); 5) Non Media Leisure. About this last activity, according to the distinction suggested by Robinson and Godbey (1997), it includes both informal leisure activities which allow an exchange of feelings and opinions (socializing, conversations, sports, and hobbies), and formal activities that allow, however, a cultural development of the individual (voluntary work, aid to other families, social participation and religious, adult education, cultural events).

To estimate simultaneously the time spent from both partners in these five groups of activities, a SURE model of ten simultaneous equations (five for each partner) was implemented. Our assumption is that the time spent on each activity by the subject, and specified as a dependent variable, depends also on the time spent by his/her partner. Therefore, the specification of the system of equations SURE (Campolo, 2012; Campolo et al., 2016; Zellner, 1962), is based on the assumption that de-pendent variables are jointly related and depend on common factors usually not observable. The joint correlation

DRAFT

between dependent variables can be formalized by modeling the error terms of the five model equations for men and women, considering separately Italian and foreign couples ($\varepsilon_W$, $\varepsilon_D$, $\varepsilon_P$, $\varepsilon_{NM}$, $\varepsilon_M$).

In particular, at the first stage of our model, we estimated the five equations using a Tobit model, which allows us to estimate dependent variables also in case of censoring. At the second stage, we used the residuals terms of the five regressions iteratively to estimate the cross-correlation between errors of the individual equations, considering also that we simultaneously estimate both activities performed by the same subjects, and also those carried out by partner. The estimated cross-correlations between residuals of the ten equations (5 equations for men and 5 for women), will be used as correction terms for the dependent variables. We used the same procedure separately for Italians couples and foreign and mixed couples (where one or both partners are foreigners).

The demographic explanatory variables used to explain different ways of allocating time among men and women in Italian and foreign couples, include: the age of the subject (Age); if the subjects in the couples are married or cohabiting (Married: 1= yes; 0=otherwise); where he/she lives (South or Islands: 1=Yes; 0=otherwise); the level of education expressed in years of schooling; the working status (1=no work as reference; 2=part time; 3=full time); an index concerning the housing space (House Index) calculated as the ratio between the number of rooms and the number of components. Moreover, to evaluate the relationship with neighbors and the territorial context, we considered two dummy variables, indicating the level of integration with the surrounding environment and the opportunity for the individual to weave the networks of relationships outside the family: a dummy variable about the frequency of contacts with friends (Friends = 1 if subject meet his/her friends during free time at least once a week, and 0 otherwise), and a dummy variable about the level of trust in others (Trust: 1=most people are trustworthy;0=otherwise). A dummy variable about the level of stress (Stress=1 if the subject feel always or often stressed, 0 otherwise), the general level of life satisfaction (Satisfaction: score 0-10); a categorical variable about the satisfaction of the economic situation (Economics: 1 = not at all as reference; 2= little; 3= quite; 4= much).

## 3. Results

Our econometric model specification allows us to jointly estimate five different activities concerning the use of time, considering men and women in couple (married or cohabiting). Moreover, in a comparative perspective, we have estimated the same model separately for Italian and mixed couples (where at least one of two partner is foreigner). The unobservable correlations within the ten equations (five for each partner) are then simultaneously estimated for both types of couples.

By comparing coefficients, it is possible to identify a first significant difference in the distribution of time within Italians (Tab. 1) and foreign or mixed couples (Tab. 2).

DRAFT

Table 1: SURE estimation results: Italian couples

| Italian couples: MAN | Paid work | Unpaid work | Personal care | Media leisure | Non-media leisure |
|---|---|---|---|---|---|
| Age (ref: 18-24) | | | | | |
| 25-34 | -0.30 | 1.07 | -0.02 | -0.36 | 0.07 |
| 35-44 | 0.10 | 1.17 | -0.04 | -0.49 | 0.15 |
| 45-54 | 0.09 | 0.70 | -0.04 | -0.28 | 0.47 |
| 55-64 | -0.36 | 0.29 | -0.04 | -0.05 | 0.64 |
| Married (1=Yes) | -0.22* | -0.04 | 0.001 | -0.04 | 0.16* |
| South or Islands(1=Yes) | 0.99*** | 0.11 | -0.02*** | 0.04 | -0.13** |
| Years of Schooling | -0.023** | 0.030** | -0.0012* | 0.0053 | 0.0057 |
| Working status (ref: No work) | | | | | |
| Part Time | 5.57*** | 3.12*** | -0.07*** | -0.86*** | -0.57*** |
| Full Time | 6.29*** | 2.85*** | -0.11*** | -1.08*** | -0.67*** |
| House_Index | -0.46*** | -0.11 | 0.0064 | 0.15* | -0.0024 |
| Friends | 0.01 | -0.14 | -0.0027 | 0.30*** | -0.034 |
| Stress | 0.91*** | 0.11 | -0.013** | -0.14** | -0.13** |
| Trust | 0.11 | 0.20* | -0.0068 | -0.022 | -0.094* |

| | | | | | |
|---|---|---|---|---|---|
| Economic satisfaction (Ref: Not at all) | | | | | |
| little | -0.22* | -0.078 | 0.0034 | 0.036 | 0.11 |
| quite | -0.13 | -0.1 | 0.0053 | 0.032 | 0.12 |
| much | 0.083 | -0.4 | 0.015 | 0.10 | 0.086 |
| Life satisfaction | -0.091*** | -0.025 | -0.0008 | 0.040** | -0.015 |
| Constant | -4.16*** | -3.13*** | 6.66*** | 4.81*** | 4.32*** |
| R-sq | 0.54 | 0.17 | 0.08 | 0.11 | 0.06 |
| Italian couples: WOMAN | Paid work | Unpaid work | Personal care | Media leisure | Non-media leisure |
| Age (ref: 18-24) | | | | | |
| 25-34 | -1.34*** | -1.62*** | -0.024 | 0.0079 | -0.14 |
| 35-44 | -1.56*** | -1.63*** | -0.044* | -0.029 | 0.13 |
| 45-54 | -1.39*** | -1.47*** | -0.051* | 0.093 | 0.44 |
| 55-64 | -1.43*** | -1.56*** | -0.049* | 0.31 | 0.52* |
| Married (1=Yes) | -0.33*** | 0.14 | -0.0073 | -0.091 | -0.01 |
| South or Islands(1=Yes) | 0.44*** | 0.52*** | -0.026*** | -0.0017 | -0.15** |
| Years of Schooling | -0.08*** | -0.07*** | 0.00026 | 0.03*** | 0.005 |
| Working status (ref: No work) | | | | | |
| Part Time | 11.8*** | 11.3*** | -0.037*** | -0.44*** | -0.54*** |
| Full Time | 13.3*** | 11.7*** | -0.051*** | -0.64*** | -0.67*** |
| House_Index | -0.85*** | -0.40*** | -0.012* | -0.11 | -0.22** |
| Friends | -0.16** | -0.15* | 0.0068 | 0.25*** | -0.13** |
| Stress | 0.59*** | 0.38*** | -0.011** | -0.063 | -0.13** |
| Trust | 0.31*** | 0.16 | -0.0093* | 0.11* | 0.013 |
| Economic satisfaction (Ref: Not at all) | | | | | |
| little | -0.18* | -0.34*** | 0.0056 | 0.016 | 0.075 |
| quite | 0.21* | 0.0091 | 0.0096 | -0.066 | 0.12 |
| much | -0.31 | -0.79*** | 0.032** | -0.0029 | 0.20 |
| Life satisfaction | -0.25*** | -0.22*** | -0.0006 | 0.048** | -0.023 |
| Constant | -8.10*** | -8.14*** | 6.60*** | 3.55*** | 3.99*** |
| R-sq | 0.89 | 0.83 | 0.04 | 0.04 | 0.05 |

*Note*: p-value: * 0.05>p>0.01; **0.01>=p>0.001; ***p<=0.001

Results on individual's age show a negative and significant effect only for women both for paid and unpaid work, while it does not show any statistical significance for men. The time devoted to personal care also decreases for women with ages, while the same does not happen for men.

As predictable, having both a part and a full-time job, increases the time spent in paid work and decrease the one spent in personal care and media and non-media leisure both for men and women. Less predictable, it increases also the time spent in unpaid work both for men and women. Being stressed increases the time spent in paid work for men and in paid and unpaid work for women, while it does not affect the time spent in unpaid work for men. Finally, life satisfaction is negatively correlated to paid and unpaid hours of work and positive correlated to media leisure both for men and women. Years of schooling reduces the number of hours spent both in paid and unpaid work for both men and women, reduces the time spent in personal care for men and increases the time spent in media leisure for women (Table 1).

We are not able to exclude the possibility of an inverse causality between stress and use of time and life satisfaction and use of time, because this result risks being affected by endogeneity problems.

Concerning foreigners and mixed couples (Table 2), similar results are obtained for the relationship between, part- and full-time job, and the time spent in paid and unpaid work both for men and women.

An opposite result is shown instead in terms of the relationship between ages of individuals and time spent in paid work, which used to be significant in the previous case only for women and is instead only significant for men in the present case. It means that foreigners men or men living in mixed couples tend to decrease the hours spent in paid work becoming older while the same does not happen for women in the same situation. No relationship appears mixed or foreign couples on respect to stress and life satisfaction. Different from the previous analysis, having friends decreases in this case the

DRAFT

amount of paid and unpaid hours of work for women. Years of schooling increases the number of hours spent in unpaid work for men and decreases the number of hours spent in paid work for women.

Table 2: SURE estimation results: Foreign or mixed couples

| Foreign or mixed couples: MAN | Paid work | Unpaid work | Personal care | Media leisure | Non-media leisure |
|---|---|---|---|---|---|
| Age (ref: 18-24) | | | | | |
| 25-34 | -3.13*** | 1.80* | 0.038 | 1.62* | 0.44 |
| 35-44 | -2.68** | 1.47 | 0.0085 | 1.44* | 0.3 |
| 45-54 | -2.86*** | 1.03 | 0.002 | 1.71** | 0.53 |
| 55-64 | -3.03*** | 1.38 | 0.0018 | 1.55* | 0.63 |
| Married (1=Yes) | 0.52* | 0.088 | -0.013 | -0.26 | 0.024 |
| South or Islands(1=Yes) | 1.18*** | -0.17 | -0.03* | -0.08 | -0.66*** |
| Years of Schooling | -0.04 | 0.07** | 0.001 | 0.037* | -0.004 |
| Working status (ref: No work) | | | | | |
| Part Time | 2.63*** | 0.21 | -0.07** | -1.07*** | -1.16*** |
| Full Time | 2.68*** | -0.49* | -0.12*** | -1.24*** | -0.73*** |
| House_Index | -0.09 | 0.096 | 0.012 | -0.031 | -0.18 |
| Friends | -0.33 | -0.31 | 0.024* | 0.2 | -0.14 |
| Stress | 0.024 | 0.16 | -0.013 | -0.046 | -0.13 |
| Trust | 0.43* | 0.11 | -0.030* | -0.29 | -0.12 |
| Economic satisfaction (Ref: Not at all) | | | | | |
| little | 0.55* | -0.0062 | -0.034* | -0.23 | 0.026 |
| quite | 0.098 | 0.42 | -0.011 | -0.27 | -0.12 |
| much | 1.49 | -0.67 | -0.006 | -0.64 | -0.81 |
| Life satisfaction | 0.045 | -0.021 | 0.003 | -0.025 | 0.017 |
| Constant | 0.8 | -1.00 | 6.61*** | 3.57*** | 4.94*** |
| R-sq | 0.26 | 0.06 | 0.11 | 0.12 | 0.10 |
| Foreign or mixed couples: WOMAN | Paid work | Unpaid work | Personal care | Media leisure | Non-media leisure |
| Age (ref: 18-24) | | | | | |
| 25-34 | -0.77 | -0.84 | -0.026 | 0.22 | 0.19 |
| 35-44 | -0.34 | -0.11 | -0.032 | -0.03 | 0.16 |
| 45-54 | -0.75 | -0.6 | -0.064* | 0.3 | 0.42 |
| 55-64 | 1.22* | 1.53* | -0.064 | 0.22 | 0.39 |
| Married (1=Yes) | 0.44 | 0.69* | -0.022 | 0.098 | -0.19 |
| South or Islands(1=Yes) | 0.83*** | 0.80** | -0.04* | 0.09 | -0.26 |
| Years of Schooling | 0.052* | 0.05 | -0.0006 | 0.037* | -0.032* |
| Working status (ref: No work) | | | | | |
| Part Time | 10.8*** | 9.65*** | -0.073*** | -0.83*** | -0.62*** |
| Full Time | 11.5*** | 9.54*** | -0.057*** | -1.09*** | -0.44** |
| House_Index | 0.53** | 0.29 | 0.001 | -0.37* | -0.38* |
| Friends | -0.88*** | -0.94*** | 0.0026 | 0.13 | 0.015 |
| Stress | -0.022 | -0.021 | -0.017 | 0.19 | -0.15 |
| Trust | -0.89*** | -0.45 | -0.028 | 0.28 | 0.11 |
| Economic satisfaction (Ref: Not at all) | | | | | |
| little | 0.51* | 0.35 | -0.013 | -0.18 | 0.28 |
| quite | 0.65** | 0.35 | 0.0072 | -0.11 | 0.021 |
| much | 1.67** | 0.24 | 0.064 | -0.40 | -0.66 |
| Life satisfaction | -0.20** | -0.19** | 0.002 | -0.009 | 0.088 |
| Constant | -9.72*** | -8.84*** | 6.62*** | 3.82*** | 4.00*** |
| R-sq | 0.87 | 0.76 | 0.08 | 0.09 | 0.06 |

*Note*: p-value: * 0.05>p>0.01; **0.01>=p>0.001; ***p<=0.001

DRAFT

## 4. Conclusions

National statistics available confirm that in most cases foreigners come to Italy for work reasons and tend to settle in areas that offer more job opportunities. Notwithstanding our estimates confirm this trend, foreign women appear to be less active both at home and outside if they live in the Centre and North just like Italian ones. On the contrary, their male partners are less active only in paid work in the same situation, both in cases of Italian couples than foreign or mixed. Analysing Italian couples, however, the contraction of media and non-media leisure time is often determined by stress for men, while for women is determined by stress only the contraction of non-media leisure. Considering foreign and mixed the same does not happen for both media and non-media leisure neither for men nor for women.

Consider now the effect of the education level. While education plays an important role in the distribution of women's time in Italian couple (negative for paid and unpaid work both for men and women, negative for personal care for men and positive for media leisure time for women), for women in foreign couple or mixed, however, the greater the educational level the less time is devoted to non-media leisure, while a positive and significant correlation appears with unpaid work and non-media leisure for men, and with media leisure for women. An inverse relationship, for men and women in mixed couples, is observed in estimating relationship between unpaid work and frequency of contacts with friends. The opposite occurs, in-stead, for natives.

## References


[1] Anxo, D., Mencarini, L., Pailhé A., Solaz A., Tanturri, M.L., Flood, L.: Gender differences in time use over the life course in France, Italy, Sweden, and the US. Fem. Econ. 17(3), 159--195 (2011)

[2] Becker, G.S.: A Theory of the Allocation of Time. Econ. J. 75(299), 493--517 (1965)

[3] Blau, F.D., Kahn, L.M., Comey, M. et al.: Culture and gender allocation of tasks: source country characteristics and the division of non-market work among US immigrants. Rev. Econ. Househ. 18, 907--958 (2020)

[4] Bloemen, H.G., Pasqua S., Stancanelli, E.G.F.: An empirical analysis of the time allocation of Italian couples: are they responsive?. Rev. Econ. Househ. 8, 345--369 (2010)

[5] Campolo, M.G.: L'allocazione del tempo lavorativo nelle famiglie italiane. Problemi di specificazione e di stima. Aracne, Roma (2012)

[6] Campolo, M.G., Di Pino Incognito A.: An Empirical Analysis of Women's Working Time, And an Estimation of Female Labour Supply in Italy. Statistica 72(2), 173--193 (2012)

[7] Campolo, M.G., Di Pino Incognito, A., Rizzi E.L.: How do life course events affect paid and unpaid work of Italian couples?. In Alleva, G., Giommi, A. (eds.) Topics in theoretical and applied statistics, pp. 193-204. Springer, Switzerland (2016)

[8] Depalo, D., Faini R., Venturini A.: The social assimilation of immigrants. Social Protection Discussion Paper, n. 0701, The World Bank (2007)

[9] Del Boca, D., Mencarini, L., Pasqua, S.: Valorizzare le donne conviene. Il Mulino, Bologna (2012)

[10] Del Boca, D., Rosina A.: Famiglie sole, Il Mulino, Bologna (2009)

[11] Del Boca, D., Vuri D.: The mismatch between employment and childcare in Italy: the impact of rationing, J. Popul. Econ. 20(4), 805--832 (2007)

[12] Hamermesh, D.S., Trejo S. J.: How do immigrants spend their time? The process of assimilation. J. Popul. Econ. 26, 507-530 (2013)

[13] Mangiavacchi, L., Rapallini, C.: Allocazione del benessere all'interno della famiglia in Italia: un approccio collettivo basato sulla soddisfazione economica. In Romano, M.C.,, Mencarini, L., Tanturri, M.L. (eds.) Uso del tempo e ruoli di genere, pp.149-146. Istat, Roma, (2012)

[14] Robinson, J.P., Godbey, G.: Time for Life: The Surprising Ways Americans Use Their Time. University Park, Pennsylvania State University Press (1999)

[15] Zellner A.: An efficient method of estimating seemingly unrelated regression and tests for aggregation bias. J. Am. Stat. Assoc. 57(298), 348-368 (1962)


DRAFT

# Εἰλείθυια comes from afar: The foreigners' contribution to fertility by Italian provinces

Eleonora Miaci[a], Cristina Giudici[a], Eleonora Trappolini[b], Marina Attili[c], Cinzia Castagnaro[c], Antonella Guarneri[c]

[a] University of Rome, Sapienza; eleonora.miaci@uniroma1.it
[b] University of Milano-Bicocca; eleonora.trappolini@unimib.it
[c] Italian National Institute of Statistics- ISTAT; maattili@istat.it

**Abstract**

The fertility differential of foreign women in Italy has decreased considerably over the years and it is expected to decrease even more over time but, despite this, the role of foreign children in slowing the decline in births remains crucial.

This study aims to further the debate on the fertility of migrants, providing an estimate of the fertility rates among foreign nationals at the provincial level over the past two decades. We intend to analyse the evolution over time and space of the contribution to fertility by foreign female citizens, investigating the determinants of their fertility behaviour and identifying differences and similarities with Italian female citizens.

**Keywords:** Migrants' Fertility, Italy, Foreign woman, Fertility, Employment level.

DRAFT

## 1. Introduction

Although fertility gaps within the same territory are frequent in many countries (Frejka, Sobotka 2008), there are few analyses exploring the subnational level in Europe and especially in Italy (Vitali, Billari 2017; Campisi et al. 2020).

Italy is constituted by very heterogeneous territories in terms of geographic, socioeconomic and demographic characteristics (Reynaud et al., 2020) and fertility levels differ considerably both at the level of rural or urban territories and by geographical distribution (Zambon et al. 2020).

Despite the extensive research in the field of migrant fertility examining the relationship between migration and family and reproductive dynamics (Landale 1997; Toulemon, Mazuy 2004; Cooke 2008; Milewski 2010), to our knowledge, there is no analysis to date that explores the subnational level with regard to the fertility of foreign female citizens.

The contribution of this paper is precisely to fill this gap in the literature.

We aim to estimate the age-specific and total fertility rates of foreign women in Italy, disaggregated by citizenship (for the five main citizenships), at a provincial level.

## 2. Fertility of foreign women in Italy

On 1st January 2022, more than 50% of the foreign female population resident in Italy came from Romania, Albania, Morocco, China and Ukraine. These five groups are characterised by different levels of maturity in the migration cycle. Foreign female residents are on average a younger population than the Italian female one and have contributed to slowing the decline in births, due to a significant presence at fertile age, a younger average age at childbirth and a fertility rate almost close to replacement level.

The share of foreign children in the Italian birth rate remains crucial. Due to the composition of the presence, it is not surprising that the highest number of foreign children recorded in the national registry is Romanian (14,248 born in 2020), followed by Moroccan (9,991) and Albanian (8,082). These citizenships represent about 40% of births to foreign mothers residing in Italy. As regards other citizenships, a non-negligible contribution comes from other African communities (in particular Egyptian and Tunisian nationals) and several Asian communities (in particular from India, Bangladesh, China and Pakistan).

As for Italian nationals, the fertility of foreigners shows territorial differences: the fertility rate is 2 children per woman in the North, 1.65 in the Centre and 1.86 in the South. (ISTAT, 2020)

Although the gap with the fertility rate of Italian women (1.17 in 2020) remains considerable, the reduction experienced by foreign female citizens in the last ten years has been sharper (-18% for the TFR of foreign women and -12% for the TFR of Italian women since 2010).

Numerous studies underline how the convergence of Italian and foreign women's reproductive behaviour is determined by social, political, cultural and employment conditions in the host society (Alba, Nee 1997; Carter 2000). In this context, several works stress the importance of the growing propensity of foreign women to work, mostly in the family service sectors with great variability linked to nationality and migratory background (Gabrielli et al. 2007; Lübke 2015; Sobotka 2008).

## 3. Data and methods

This research is conducted as part of a curricular internship at ISTAT, carried out in the context of the EMOS (European Master in Official Statistics Workshop) programme, on "Estimating the fertility of the cohorts of women resident in Italy by citizenship in the years 2002-2020", within the framework of the doctorate in the School of Statistical Sciences, Demography curriculum, at La Sapienza University of Rome. Starting from the ISTAT births database, which contains data on all women who have had at least onechild registered in the registry office since 1999, total and age-specific fertility rates were constructed for Italians and foreigners and for the five main citizenships in Italy (Romanians, Albanians, Moroccans, Chinese, Ukrainians). These indicators were then validated to provincial detail.

## 4. Preliminary results

Figure 1 displays age-specific fertility rates for the total, Italian and foreign female population at national level and by macro-areas. Foreign nationals realize their fertility behaviour earlier than Italians and with higher rates but, compared to 2010, in 2020 they experienced a greater reduction in number of births and a stronger postponement. The underlying area between the age specific rate of the total population and that of the Italian population reveals that the contribution of foreign female citizens, both in 2010 and 2020, is lower in the South than in the rest of Italy. This depends on several factors: the lower fertility rate experienced by foreign nationals in the South compared to North and Centre, their older age structure (explained by looking at the prevailing nationalities in the different macro-areas) and also on the fact that in the south, as we have already pointed, there are precisely fewer foreigners present and therefore their contribution is lower when calculating the rate of the total population.

The provincial detail of the total fertility rate (figure 2) reveals how despite a certain heterogeneity between the geographic areas, there is a clear homogeneity between the various levels of fertility of Italian provinces in the different macro-areas.

The provinces reporting a high total fertility rate for foreigners are mainly in the North, with the exception of some Sicilian provinces. The lowest rates are instead concentrated in the Centre, in some provinces of the South and in Sardinia.

Although the fertility of foreigners follows the trend of that of Italians, the data indicate a fertility differential present in almost all the provinces of the North and the Centre but limited in the South.

DRAFT

The five communities examined, although they have all experienced widespread fertility declines, differ in their levels of fertility in their countries of origin. In 2020 TFR ranges from 1.2 in Ukraine to 2.4 in Morocco. Except for the latter case, while remaining higher than that of Italian nationals, the TFRin the countries of origin lies below the replacement level.

One of the factors explaining the low fertility rate of foreign women in Italy is their origin mainly fromlow fertility countries.

Concerning Romanian migrants, we found that they have a lower fertility rate than non-migrant women,confirming previous research (Mussino, Cantalini, 2022).

One of the most original results that emerged from our study is the higher fertility of Albanian and Moroccan citizens in Italy compared to both those who remained in the country of origin (table 1) and to native Italians (figure 3). The analysis of the socio-economic characteristics, labour market integration and attitudes of these communities allowed us to hypothesise explanations for this result.

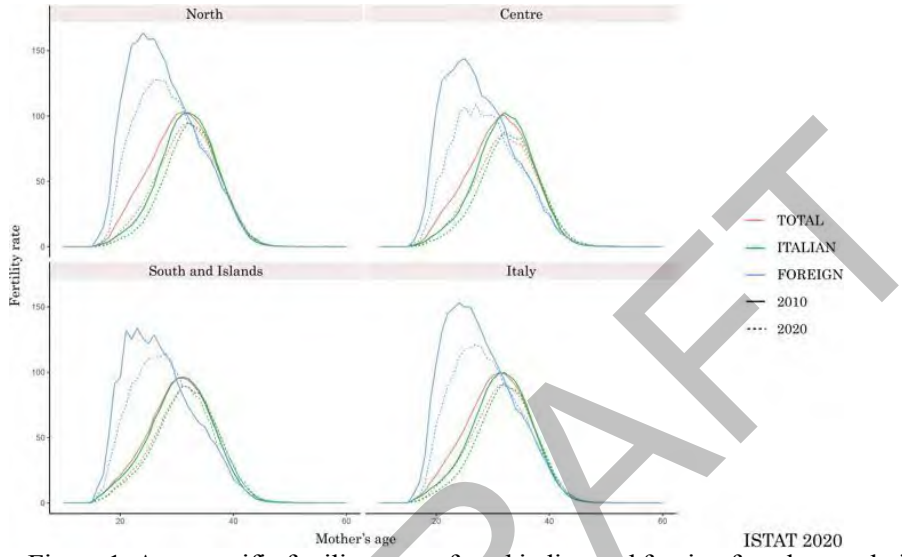



Figure 1: Age-specific fertility rates of total italian and foreign female population for microregion

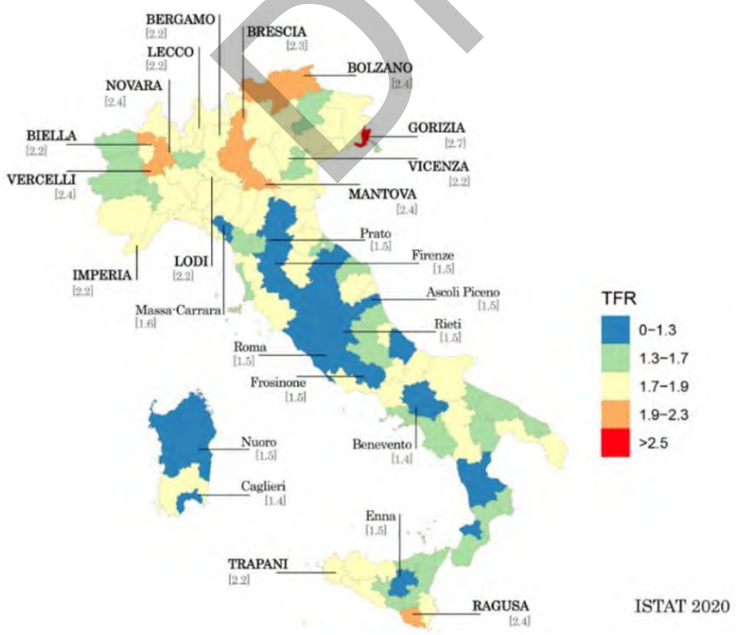



Figure 2: Total fertility rate of foreign population in Italy, 2020

DRAFT

Table 1: TFR of the first five citizenship in Italy and in the country of origin, 2019-2021

| *ITALY* | Romania | Albania | Ukraine | China | Morocco |
|---|---|---|---|---|---|
| **2019** | 1.4 | 2.1 | 1.3 | 1.5 | 3.4 |
| **2020** | 1.4 | 2 | 1.3 | 1.1 | 3 |
| **2021** | 1.5 | 2.1 | 1.2 | 0.9 | 2.9 |

ISTAT

| *COUNTRY OF ORIGIN* | Romania | Albania | Ukraine | China | Morocco |
|---|---|---|---|---|---|
| **2019**[b] | 1.8 | 1.4 | 1.2 | 1.5 | 2.4 |
| **2020**[c] | 1.6 | 1.4 | 1.2 | 1.3 | 2.4 |
| **2021**[d] | 1.7 | 1.4 | 1.2 | 1.2 | 2.3 |

[b] [c] World Bank, [d] World Population Prospects

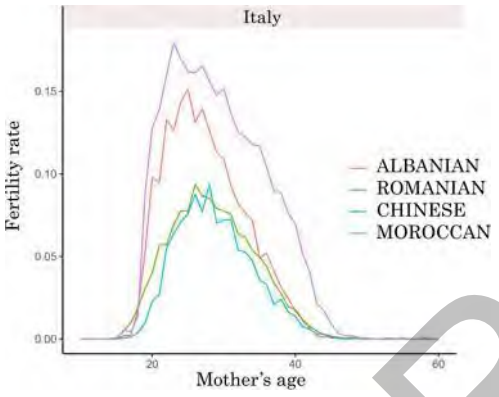



Figure 3: Age-specific fertility rate of foreign female population by citizenship in Italy, 2020

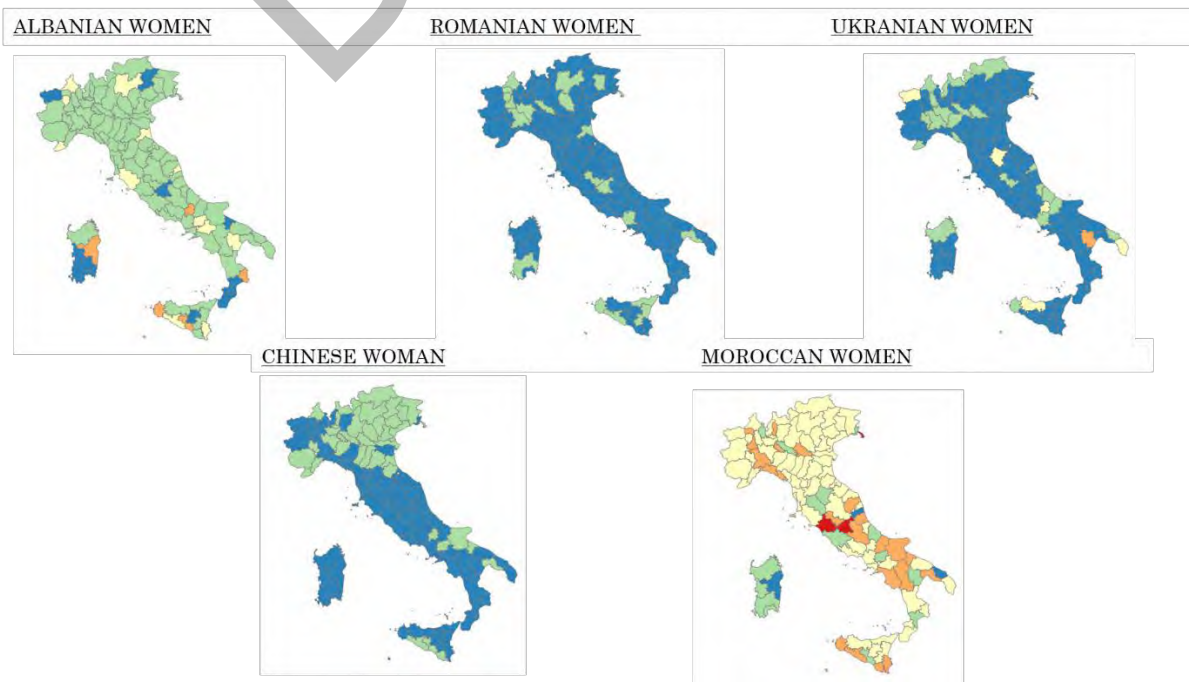



Figure 4: Total fertility rate at provincial level for the five main citizenship in Italy, 2020

DRAFT

## 5. Conclusion and further developments

This preliminary work introduced the relevant literature on migrants' fertility, presented a syntheticpicture of the foreign presence on Italian territory and of the characteristics of the different nationalities, and illustrated a selection of the descriptive results obtained.

The paper will be implemented with linear regression models to analyse the relationship between fertility behaviour and socio-economic and contextual factors, as well as to identify differencesbetween Italian provinces.

### References


Istat. (2020). Censimento della popolazione e dinamica demografica 2020. Retrieved from https://demo.istat.it/ Ministero del Lavoro e delle Politiche sociali. (2020). Le comunità migranti in Italia. Rapporti 2020.

Adserà, A., Ana M. Ferrer. The fertility of recent immigrants to Canada. (2013).

Algan Y., Bisin A., Manning A., Verdier, T. Cultural integration of immigrants in Europe. Oxford University Press, (2012).

Alba, R., V. Nee. Rethinking Assimilation Theory for a New Era of Immigration. International Migration Review 31(1997):826–74.

Alderotti, G., Mussino, E., Comolli, C. L. (2022). Natives' and migrants' employment uncertainty and childbearing during the great recession: a comparison between Italy and Sweden. European Societies, 1-35.

Campisi, N., Kulu, H., Mikolai, J., Klüsener, S., Myrskylä, M. (2020). Spatial variation in fertility across Europe: Patterns and determinants. Population, Space and Place, 26(4), e2308.

Carling, J. Gender dimensions of international migration. Global migration perspectives 35.1 (2005): 1-26.

Carter, M. Fertility of Mexican immigrant women in the US: A closer look. Social science quarterly (2000): 1073-1086.

Colombo, A., Sciortino G. Italian immigration: The origins, nature and evolution of Italy's migratory systems. Journal of Modern Italian Studies 9.1 (2004): 49-70.

Cooke, T.J. Migration in a family way. Population, space and place 14.4 (2008): 255-265.

Fernández, R., Fogli A. Fertility: The role of culture and family experience. Journal of the European economic association 4.2-3 (2006): 552-561.

Frejka, T., Sobotka, T. (2008). Overview Chapter 1: Fertility in Europe. Demographic Research, 19(3), 15-46.

Gabrielli, G., Paterno, A., White, M. (2007). The impact of origin region and internal migration on Italian fertility. Demographic Research, 17, 705-740. https://doi.org/10.4054/DemRes.2007.17.24

Garssen, J., Han N. Fertility of Turkish and Moroccan women in the Netherlands: Adjustment to native level within one generation. Demographic Research 19 (2008): 1249-1280.

Giannantoni, P., Strozza, S. (2016). Foreigners' contribution to the evolution of fertility in Italy: A re-examination on the decade 2001-2011. Rivista Italiana di Economia, Demografia e Statistica, 69(3), 129-140.

Kulu, H., González-Ferrer, A. (2014). Family dynamics among immigrants and their descendants in Europe: Current research and opportunities. European journal of population, 30, 411-435.

Landale, N. S. (1997). Immigration and the family: An overview. Immigration and the family: Research and policy on US immigrants, 281-291.

Lübke, C. How migration affects the timing of childbearing: The transition to a first birth among Polish women in Britain. European Journal of Population 31 (2015): 1-20.

Milewski, N. (2010). Immigrant fertility in West Germany: Is there a socialization effect in transitions to second and third births? European Journal of Population/Revue européenne de Démographie 26 (3), 297-323

Mussino, E., Cantalini, S. (2022). Influences of origin and destination on migrant fertility in Europe. Population Space and Place, https://doi.org/10.1002/psp.2567.

Mussino, E., Ortensi, L. E. (2018). The same fertility ideals as in the country of origin? A study of the personal ideal family size among immigrant women in Italy. Comparative Population Studies, 43, 243–274. https://doi.org/10.12765/CPoS-2019-03.

Reynaud, C., Miccoli, S., Benassi, F., Naccarato, A., & Salvati, L. (2020). Unravelling a demographic 'Mosaic': Spatial patterns and contextual factors of depopulation in Italian Municipalities, 1981–2011. Ecological Indicators, 115, 106356.

Sobotka, T. (2008). Overview Chapter 7: The rising importance of migrants for childbearing in Europe. Demographic research, 19, 225-248.

Toulemon, L., Mazuy, M. (2004). Comment prendre en compte l'âge à l'arrivée et la durée du séjour en France dans la mesure de la fécondité des immigrants (p. 34). France: INED.

Vitali, A., Billari, F. C. (2017). Changing determinants of low fertility and diffusion: A spatial analysis for Italy. Population, Space and Place, 23(2), e1998.

Zambon, I., Rontos, K., Reynaud, C., & Salvati, L. (2020). Toward an unwanted dividend? Fertility decline and the North–South divide in Italy, 1952–2018. Quality & Quantity, 54, 169-187.


DRAFT

# ESG, sustainability and stock market risk

Michele Costa[a]

[a]Department of Economics, University of Bologna; michele.costa@unibo.it

**Abstract**

In this paper we aim to investigate the relationship between ESG score and assets characteristics, focusing on volatility. We classify stocks on the basis of high/low ESG score and we evaluate ESG effects by measuring the distance between the 2 group distributions. The analysis of stocks in the STOXX 600 Index from 2017 to 2021 suggests that companies with higher ESG outperform companies with lower ESG, also allowing to highlight COVID related effects.

***Keywords:*** Sustainable finance, ESG, Stock market risk, Volatility

DRAFT

## 1. Introduction

Environmental, Social and Governance (ESG) score represents the most widespread and used indicator to evaluate the sustainability dimension of a company and it is the key factor to assess the sustainability impact of an investment in a company. Over the last years we are facing a significant shift in capital markets perception toward corporate sustainability, which has fueled a growing empirical and theoretical literature.

In this paper we aim at investigating the relationship between the ESG score and the risk of a stock, with the purpose to contribute to the debaate on the relevance and the effectiveness of ESG indicators and to explore and assess the role of ESG.

The relationship between ESG and assets characteristics is multi-faceted and not straightforward (1; 2; 4; 5). On one hand, companies with high ESG could have stronger risk mangement practices, which could contribute to a lower volatility, on the other hand, ESG score represents only one element of the overall risk profile, which certainly depends on a variety of factors. There are also different ESG metrics and their effect on asset characteristics may not be uniform.

We add inequality decomposition methods to the tools used to evaluate the effects of ESG scores, so as to combine distribution-based assessments with traditional assessments based on synthetic indicators such as averages. An analysis of the companies included in the STOXX Europe 600 Index points to a link between higher ESG and lower volatility, with, however, companies with high ESG mostly affected by effects related to the COVID19 pandemic.

## 2. Methodology

We address che relationship between ESG score and asset characteristics as a classification problem. Given $n$ assets, in order to classify them into 2 groups, we introduce a $n \times 1$ vector $Z$

$$Z = \begin{cases} z_1 & \text{with } z_1 = 1 \text{ if } ESG < p_1 \\ z_2 & \text{with } z_2 = 0 \text{ if } ESG \geq p_2 \end{cases} \tag{1}$$

where, for $p_1 = p_2$, e.g. when using the median of ESG score, all $n$ assets are considered, while for $p_1 \neq p_2$, e.g. when the ESG quartiles are considered, it is possible to refer to a part of the assets only.

With the aim to evaluate the relevance of ESG score and its effects on volatility, we thus define two groups of assets on the basis of a high / low ESG score level, thus tracing the assessment of ESG-related effects to the comparison between two distributions.

Our traditional starting point is given by the synthetic evaluation provided by the difference between the group means. Our purpose is to add to the information derived from the group means, the information provided by the comparison of the group distributions.

We compare the volatility distribution of groups High and Low: when the two distributions perfectly overlap, the ESG score does not affect the assets returns, while, for decreasing overlapping levels, the influence of ESG score on asset volatility increases. When groups High and Low do not overlap, the influence of ESG score reaches its maximum and assets are perfectly classified on the basis of their ESG score.

In order to evaluate the distance between the distributions of groups High and Low, as well as to take into account their overlapping, we resort to the decomposition of an inequality indicator, thus allowing to effectively compare different groups.

We refer to one of the most used and widespread inequality measures, the Gini index, for which many different decompositions have been proposed. Among the many contributions we use the Dagum's Gini index decomposition (3), which allows to explicitly take into account the overlap between groups.

For the case of $n$ assets disaggregated into 2 groups of size $n_1$ and $n_2$, with $n_1 + n_2 = n$, the Gini index can be expressed as

$$G = \frac{1}{2n^2\bar{y}} \sum_{j=1}^{2} \sum_{h=1}^{2} \sum_{i=1}^{n_j} \sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \tag{2}$$

where $Y$ is the assets characteristic being investigated, in our case volatility, $\bar{y}$ is its arithmetic mean, $y_{ji}$ is its value in the $i$-th asset of the $j$-th group and, accordingly, $y_{hr}$ is its value in the $r$-th asset of the $h$-th group. For the case $j \neq h$, let us define

$$(y_{ji} - y_{hr})^+ = max\{(y_{ji} - y_{hr}), 0\}$$

and

$$(y_{ji} - y_{hr})^- = max\{-(y_{ji} - y_{hr}), 0\}$$

such as

$$|y_{ji} - y_{hr}| = (y_{ji} - y_{hr})^+ + (y_{ji} - y_{hr})^-.$$

It is therefore possible to derive the expressions for the inequality between groups $G_b$ and for the overlapping component $G_o$ as

$$G_b = \frac{1}{2n^2\bar{y}} \left( \sum_{i=1}^{n_1} \sum_{r=1}^{n_2} (y_{1i} - y_{2r})^+ + \sum_{i=1}^{n_2} \sum_{r=1}^{n_1} (y_{2i} - y_{1r})^+ \right) \tag{3}$$

and

$$G_o = \frac{1}{2n^2\bar{y}} \left( \sum_{i=1}^{n_1} \sum_{r=1}^{n_2} (y_{1i} - y_{2r})^- + \sum_{i=1}^{n_2} \sum_{r=1}^{n_1} (y_{2i} - y_{1r})^- \right). \tag{4}$$

Inequality between $G_b$ and overlapping component $G_o$ allow to evaluate the contribution to total inequality attributable to the differences between the groups, that is, in our case, to the effect of ESG score.

The role of the two components is quite different. On one hand, an high (low) $G_b$ indicates a relevant (slight) ESG effect, as total inequality is (is not) strongly influenced by inequality between. On the other hand, as pointed out above, an high (low) $G_o$ suggests a slight (relevant) ESG effect, since complete overlapping corresponds to the absence of ESG effect, while $G_0 = 0$ (High and Low groups are perfectly separated) indicates a total stratification.

DRAFT

On the basis of the different meaning of $G_b$ and $G_o$ it is possible to derive an indicator of the relevance of ESG as $I_G = (Gb - G_o)/G$. A further indicator of ESG effects can be obtained by calculating the ratio between the means of the standard deviations in the two groups: $I_{\bar{\sigma}} = (\bar{\sigma}_L - \bar{\sigma}_H)/\bar{\sigma}_H$. When the two indicators show a similar trend, the information provided by the means is supported by the indications derived by the group distributions. In the opposite case, when $I_G$ differs from $I_{\bar{\sigma}}$, the group means are not fully informative and it is important to also exploit information from group distributions.

Our method is extremely flexible and can be adopted to investigate different assets characteristics more than volatility. We can generalize our analysis, moving from the $n \times 1$ vector $Z$ to a $n \times k$ matrix $Z$, where $k$ is the number of characteristics analyzed and each column of $Z$ is a $0, 1$ vector classifying the $n$ asstes into two groups with respect a specific characteristic. Further possible generalizations concern employing more than two groups and weighing the different characteristics differently, always remaining within the same framework.

## 3. Data and results

We investigate the relationship between ESG and volatility of the stocks included in the STOXX Europe 600 Index, measured by means of the standard deviation of their returns. We employ monthly data ranging from 2017 to 2021, thus allowing us to also evaluate the presence of effects related to the COVID19 pandemic. Among the different ESG metrics available, we refer to Refinitiv methodology, fully aware that the use of a different metric could influence the results. An interesting future development could involve the joint use of multiple ESG indicators and the comparison of the deriving stock classifications, so as to analyze their effects on results. In order to classify the $n$ assets into 2 groups we use both the median of ESG score with $p_1 = p_2$, and the case $p_1$ equal to the first quartile and $p_2$ equal to the third quartile.

Table 1 illustrates the results related to the two groups obtained by classifying the assets on the basis of the median of their ESG score (the results for the first and last quartile are quite similar). First we report the mean of the standard deviations of the assets in the two groups. We can observe how, from 2017 to 2021, the average of the standard deviations increases, with a peak in 2020 in correspondence with the COVID. The values of volatility in the group with lower ESG are steadily higher.

We assess the relevance of our results by developing a bootstrap procedure resampling assets from the original list according to an i.i.d. sampling with replacement. Table 1 includes the mean and the standard deviation of 10000 replicates of the bootstrap. The bootstrap mean is always equal or very close to the observed values; this element, together with the low values of the standard deviation, strongly supports our results.

The second part of Table 1 shows the ratios $G_b/G$ and $G_o/G$ which, by evaluating the weight of inequality between groups and of overlapping component on total inequality, allow us to obtain further information on the role of ESG scores. From 2017 to 2021 the relevance of $G_b$ increases slightly, with the exception of the negative peak of 2020, suggesting a stronger ESG effect and a deep impact of COVID pandemic. The overlapping component also indcates, by decreasing, a stronger ESG effect, still except in 2020. Also for $G_b$ and $G_o$ the bootstrap results support our findings: the means of 10000 bootstrap replicates are always equal or very close to the observed values, with an almost negligible standard deviation.

The last part of Table 1 shows indicators $I_{\bar{\sigma}}$ and $I_G$ which agree in suggesting an increase in the ESG effect over the last 5 years, together with detecting a strong impact of COVID pandemic. An interesting element arising from the comparison between $I_{\bar{\sigma}}$ and $I_G$ highlights the dynamic after the pandemic: from 2020 to 2021, $I_G$ shows a stronger recovery and, with respect to 2019, $I_G$ indicates a more pronounced ESG effect, while $I_{\bar{\sigma}}$ suggests a slight decline.

DRAFT

Table 1: Mean of the observed standard deviations of the assets in group with High / Low ESG score; observed ratios $G_b/G$ and $G_o/G$; mean and standard deviation of 10000 bootstrap replicates

| | | 2017 | 2018 | 2019 | 2020 | 2021 | 2018-19 | 2020-21 |
|---|---|---|---|---|---|---|---|---|
| High ESG | $\bar{\sigma}$ | 5.26 | 6.31 | 6.54 | 11.71 | 6.65 | 6.77 | 9.68 |
| | mean boot | 5.26 | 6.31 | 6.54 | 11.71 | 6.65 | 6.77 | 9.69 |
| | std.dev.boot | 0.12 | 0.13 | 0.17 | 0.30 | 0.14 | 0.13 | 0.21 |
| Low ESG | $\bar{\sigma}$ | 5.63 | 7.01 | 7.17 | 12.09 | 7.23 | 7.67 | 10.22 |
| | mean boot | 5.64 | 7.02 | 7.18 | 12.07 | 7.21 | 7.67 | 10.20 |
| | std.dev.boot | 0.19 | 0.18 | 0.18 | 0.32 | 0.17 | 0.17 | 0.25 |
| Ineq. between | $G_b/G$ | 0.29 | 0.32 | 0.30 | 0.27 | 0.31 | 0.34 | 0.28 |
| | mean boot | 0.29 | 0.32 | 0.30 | 0.27 | 0.31 | 0.34 | 0.28 |
| | std.dev.boot | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| Overlapping | $G_o/G$ | 0.21 | 0.19 | 0.20 | 0.23 | 0.19 | 0.17 | 0.22 |
| | mean boot | 0.21 | 0.19 | 0.20 | 0.23 | 0.20 | 0.17 | 0.22 |
| | std.dev.boot | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 |
| | $I_{\bar{\sigma}}$ | 0.07 | 0.11 | 0.10 | 0.03 | 0.09 | 0.13 | 0.06 |
| | $I_G$ | 0.08 | 0.13 | 0.10 | 0.04 | 0.12 | 0.17 | 0.06 |

## 4. Conclusions

The analysis between ESG score and assets characteristics can be addressed in the framework of classification and can greatly benefit from using inequality decompositioni methods, able to also take into account the overlapping between group distributions. On the basis of these methods it is possible to exploit the information related to the entire distribution of the groups and not rely only on the group means.

We analyze the stocks included in the STOXX Europe 600 Index and refer to Refinitiv ESG score from 2017 to 2021. Our findinds support ESG as a positive driver of assets returns, with a relationship between higher ESG score and lower volatility. Covid pandemic strongly affected ESG effect, which was greatly weakeened in 2020, while a strong recovery can already be observed in 2021. The size of the recovery is stronger on the basis of $I_G$, that is when evaluating the entire group distributions.

The extreme flexibility of our proposal makes it possible to add other asset characteristics to the analysis, also considering them jointly, with numerous further developments, such as the possibility of more than two groups, capable of providing promising results and effectively including sustainability issues in stock market risk evaluation.

DRAFT

## References


[1] Baker M., Egan M.L., Sarkar S.K.: How do investors value ESG? NBER WP n. 30708 (2022)

[2] Bertelli B., Torricelli C.: ESG compliant optimal portfolios. CEFIN WP n. 88 (2022)

[3] Dagum C.: A new approach to the decomposition of the Gini income inequality ratio. Emp. Eco. **22**, 515–531 (1997)

[4] Friede G., Busch T., Bassen A.: ESG and financial performance: aggregate evidence from more than 2000 empirical studies. J. Sust. Fin. & Inv. **5**, 210–233 (2015)

[5] Hartzmark S.M., Sussmann A.B.: Do investors value sustainability? A natural experiment examining ranking and fund flows. J. Fin. **74**, 2789–2837 (2019)

# Exploring the effect of consumer motivation and perception of sustainability on food choices with a Discrete Choice Experiment

Gloria Solano-Hermosilla [a], Jesus Barreiro-Hurle[b], and Ilaria Lucrezia Amerise[c]

[a] Department of Business Organisation and Marketing, University Pablo de Olavide; Sevilla, SPAIN,
gmsolher@upo.es
[b] European Commission - Joint Research Centre; Sevilla, SPAIN,
jesus.barreiro-Hurle@ec.europa.eu
[c] Department of Economics, Statistics and Finance; University of Calabria; ITALY,
ilaria.amerise@unical.it

DRAFT

**Abstract**

In shifting towards a more sustainable food system, agri-food companies increasingly integrate sustainability into their innovation decisions and practices. However, their success depends on consumers' willingness to pay (WTP) for more sustainable food products and actual purchases. Key to this are government policies and companies' marketing and pricing strategies, for which understanding the interaction of factors driving food choices is essential. We use conditional logit in this work because the objective of the Discrete choice experiments (DCE) is to relate choice to the attribute levels used to define each profile in the choice task.

**Keywords:** Sustainability, Discrete choice experiment, Food, Consumers

## 1. Introduction

In shifting towards a more sustainable food system, agri-food companies increasingly integrate sustainability into their innovation decisions and practices. However, their success depends on consumers' willingness to pay (WTP) for more sustainable food products when they buy [1]. Abundant literature suggests that consumers pay significant attention to and declare WTP for sustainability-related attributes of packaged food products, such as eco-labels, fair trade, health and other claims and visual attributes, but these may not be the main drivers of consumer choices [2], [3]. There is a value-action gap (also called the attitude-behavior gap) that the complex interactions between different purchase priorities and perceptions may explain [3], [4]. To close this gap, government policies and companies' marketing and pricing strategies are critical, for which understanding the interaction of factors driving choices is essential.

In this study we used a Discrete choice experiments (DCE) that is a quantitative technique for eliciting preferences that can be used in the absence of revealed preference data. The method involves asking individuals to state their preference over hypothetical alternative scenarios, goods or services. Each alternative is described by several attributes and the responses are used to determine whether preferences are significantly influenced by the attributes and also their relative importance. DCEs have become a method increasingly used in food research to uncover trade-offs when choosing alternatives, particularly when exploring credence attributes, such as those related to sustainability [5].

## 2. Motivation

The paper makes an original contribution to the field of sustainable food systems by examining the factors that influence consumers' choices and willingness to pay for more sustainable food products. We argue that for agri-food companies to successfully integrate sustainability into their innovation decisions and practices, it is important to understand how consumers' choices are influenced by government policies and marketing and pricing strategies.

To achieve this, the paper uses a Discrete Choice Experiment (DCE) and conditional logit analysis to assess consumers' preferences and WTP for different attributes of sustainable food products. The results of the study provide insights into the relative importance of various attributes, such as eco-labels, product origin, and production methods, in driving consumers' choices.

The paper's original contribution lies in its use of DCE and conditional logit analysis to understand the factors that influence consumer choices and WTP for sustainable food products. This approach allows for a more nuanced analysis of the complex interactions between various attributes and their impact on consumers' choices. The results of the study can inform the development of effective marketing and pricing strategies for sustainable food products that are more likely to appeal to consumers. In addition, the original contribution comes from a unique, multi-national and representative database, providing important insights and empirical findings on how to bridge the gap between consumers valuing sustainable food and being willing to pay for it.

In summary, we analyzed the factors that influence consumers' choices and WTP for more sustainable food products. The study's results can inform the development of effective marketing and pricing strategies that promote sustainable food choices and contribute to the shift towards a more sustainable food system.

DRAFT

## 3. Methods

The analysis of DCE data typically involves regression models that have a dichotomous or polychotomous categorical dependent variable, such as a probit, logit, or multinominal logit specification. In its simplest form, the observed sources of utility can be defined as a linear expression in which each attribute is weighted by a unique parameter to account for that attribute's marginal utility. It is usual to specify the regression model in terms of differences in attribute levels between the choices being analyzed:

$$\Delta Y = \beta_0 + \beta_1(X_{1i} - X_{1j}) + \beta_2(X_{2i} - X_{2j}) + \cdots + \beta_k(X_{ki} - X_{kj}) + (\varepsilon_i - \varepsilon_j)$$

Moreover, as respondents are asked to consider multiple choice pairs, it cannot be assumed that the error terms are independent and panel data estimation techniques are required. The estimated parameters represent the marginal utility associated with a change in the attribute level in moving from one alternative to the other.

This work will assist researchers in evaluating and selecting among alternative approaches to conducting statistical analysis of DCE data.
We first present a DCE example and a simple method for using the resulting data.
In this study, we use a survey of more than 20 000 individuals evenly split across ten European countries to explore whether purchase motives are reflected in food choice decisions. Consumers undertook a Discrete Choice Experiment with packaging elements as varying attributes in six food products (instant coffee, crisps, baby food, fish fingers, chocolate and yoghurt) and an online experiment where they identified packaging elements that led them to believe products were different concerning sustainability aspects (i.e. environmental sustainability and healthiness) among others. Then consumers identified the three most important motives when purchasing food. In the analysis, we combine these three information sources for the same individual to see whether consumers declaring sustainability is the most important factor when buying food attach higher utility to packaging elements informing them of significant differences in sustainability between products (i.e. packaging elements conveying product sustainability).
For this, the experiment presented product pairs as front-of-pack pictures of hypothetical branded product versions. They differed across eight potential design elements, including claim or logo-type attributes (taste, quality, origin, recipe) which can be either present or absent, and package design variants (product description, background colour, picture and positioning of the picture) which can take one of two options. The purchase decision (DCE) choice set followed an efficient design, resulting in 12 choice situations. As each choice card had two options, this yielded 24 different images per product. The 12 choice cards were blocked into six groups of two, and each individual saw a group for each of the five products. The choice set for eliciting perceived product differences (i.e. environmental sustainability and healthiness) was created manually to capture choice pairs with variations in the number of elements differing between options in the choice cards. In both experiments, the choice set block and order of products were randomised across individuals.

DRAFT

### 3.1 Conditional Logit

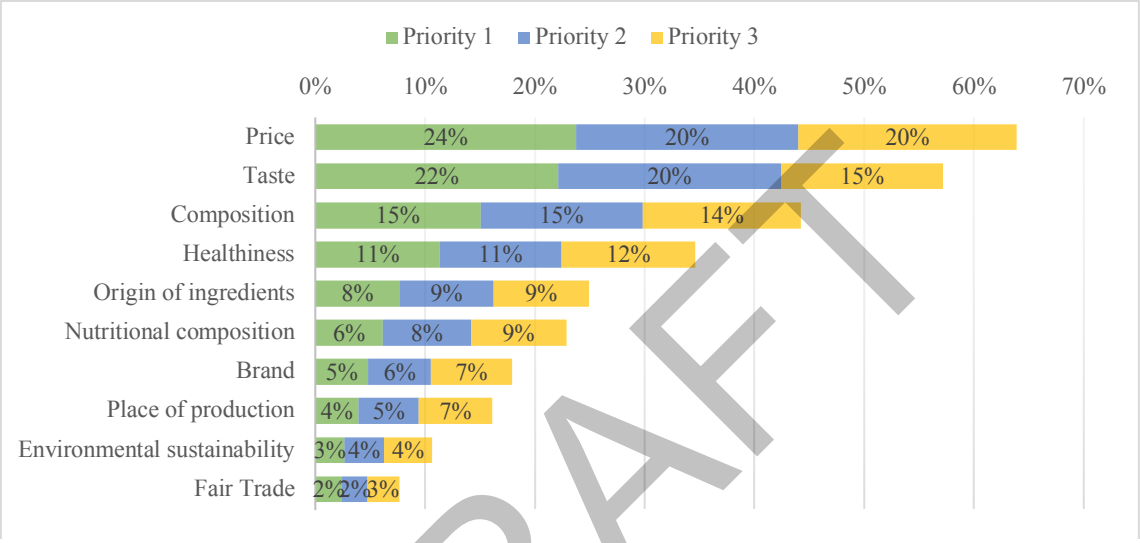
Choice data from a two-alternative, forced-choice DCE as described in several examples are most often analyzed using a limited dependent-variable model because the left-hand-side variable in the regression is typically a 1 for the alternative that was chosen in a given choice task or a 0 for the alternative that was not chosen in that choice task. The basic limited dependent-variable method used to analyze data generated by this type of experiment is conditional logit. Conditional logit relates the probability of choice among two or more alternatives to the characteristics of the attribute levels defining those alternatives. In a DCE, the elements describing the alternatives are the attribute levels used to define each profile in the choice task. Conditional logit was shown by [7] to be consistent with random utility theory. The novelty in McFadden's use of the logit model is that he applied this model to choice behavior that was consistent with economic theory and derived a regression model that relates choices to the characteristics of the alternatives available to decision makers. McFadden used the term "conditional logit" to describe this innovation.
McFadden originally applied this framework to observed transportation choices. His work laid the foundation for what is now known as conjoint analysis [8] involving hypothetical or stated choices. Using random utility theory, the utility associated with an alternative or profile is assumed to be a function of observed characteristics (attribute levels) and unobserved characteristics of the alternative. This theoretic framework also assumes that each individual, when faced with a choice between two or more alternatives, will choose the alternative that maximizes his or her utility.
We use conditional logit in this work because the objective of the DCE is to relate choice to the attribute levels used to define each profile in the choice task. The conditional logit model can be executed by using the function clogit() that is included in the survival package. Frequently,

in order to consider the effects of individual characteristics on the valuation of attributes, the interaction between individual characteristics and attribute variables are included in the model.

## 4. Results

The results in Figure 1 show that, consistent with other surveys, consumers prioritise price (64%) and taste (57%) over sustainability concerns when purchasing food. Regarding sustainability dimensions, health (34%) is more important than the environment (11%) when shopping for food.

***Table 1. Declared top three consumers' food purchase motivations.***




DRAFT

We run logistic regressions for the two dichotomous variables capturing whether consumers perceive differences in product environmental sustainability and healthiness [6]. Our independent variables capture whether there is a difference in each of the design elements. We also include controls for product, country and socio-demographics.

*Table 2 Effects of variations in packaging attributes on consumers believing product versions differ in sustainability-related product characteristics.*

| | Front-of-pack elements | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Product characteristics** | **Recipe claim** | **Colour** | **Image position** | **Origin of ingredients** | **Taste claim** | **Quality logo** | **Image** | **Product description** |
| Healthiness | -0.073 | -0.18*** | 0.042 | 0.068** | 0.034 | 0.051 | 0.038 | 0.016 |
| Sustainability | -0.088 | -0.055 | -0.069** | 0.25*** | 0.095 | -0.090* | 0.14* | -0.073* |

Robust standard errors in parentheses clustered at the level of respondent ID
*** p<0.01, ** p<0.05, * p<0.1
Note: A positive (negative) and significant coefficient means that consumers perceive the attribute signals (does not signal) the specific product characteristic. A non-significant coefficient means that the attribute does not affect the perception of the specific product characteristic. We retain the information from those coefficients significant at $p < 0.05$
Source: Solano-Hermosilla et al. (2023)

The DCE results show that consumers' purchasing behaviour aligns with their stated motives, as they are willing to pay more for attributes that signal their motivations and less for those that do not. However, interaction with the importance of prices partly masks this for environmental

sustainability but not healthiness. It partly explains the lack of correspondence between declared motives and actual choices.

*Table 3. Estimated coefficients and WTP for the impact of package elements, prices and interactions on consumer choices.*

| VARIABLES | Choice | WTP |
|---|---|---|
| buy | 0.98*** | 8.936*** |
| | (0.018) | |
| price | -0.11*** | |
| | (0.0034) | |
| ORIGIN | 0.064*** | 0.586*** |
| | (0.0079) | |
| TASTE | -0.21*** | -1.879*** |
| | (0.0066) | |
| QUALITY | -0.13*** | -1.213*** |
| | (0.0071) | |
| RECIPE | -0.28*** | -2.565*** |
| | (0.0064) | |
| COLOUR_1 | 0.023*** | 0.210*** |
| | (0.0058) | |
| PICTURE_1 | -0.065*** | -0.590*** |
| | (0.0061) | |
| POSITION_1 | 0.23*** | 2.107*** |
| | (0.0055) | |
| NAME_1 | 0.059*** | 0.542*** |
| | (0.0066) | |
| ***Interaction priorities - price*** | | |
| Prioritise product sustainability x price | -0.021*** | -0.189*** |
| | (0.0075) | |
| Prioritise product healthiness x price | 0.014** | 0.129*** |
| | (0.0057) | |
| ***Interaction sustainability priority x attributes signalling sustainability*** | | |
| Prioritise sustainability x origin claim | 0.13*** | 1.142*** |
| | (0.020) | |
| ***Interaction healthiness priority x attributes signalling healthiness*** | | |
| Prioritise healthiness x origin claim | 0.033** | 0.299** |
| | (0.013) | |
| ***Interaction sustainability priority x attributes not signalling sustainability*** | | |
| Prioritise sustainability x image position 1 | -0.042** | -0.386** |
| | (0.017) | |
| Prioritise sustainability x quality logo | 0.025 | 0.232 |
| | (0.020) | |
| Prioritise sustainability x product descr 1 | -0.017 | -0.156 |
| | (0.020) | |
| ***Interaction sustainability priority x attributes not affecting sustainability perception*** | | |
| Prioritise sustainability x recipe claim | 0.071*** | 0.649*** |
| | (0.019) | |
| Prioritise sustainability x colour 1 | 0.012 | 0.105 |
| | (0.018) | |
| Prioritise sustainability x taste claim | 0.060*** | 0.547*** |
| | (0.020) | |
| Prioritise sustainability x image 1 | 0.012 | 0.107 |
| | (0.019) | |
| Observations | 603,990 | |

DRAFT

The study results are expected to help policymakers and agri-food companies design labelling and marketing and communication strategies, as well as possible interventions to affect what consumers value and moderate the relationship between what they value, perceive and buy based on consumer responses and decisions. Further research can explore heterogeneity of results across products and

and human-centric design principles in their AI development strategies. New business models will also emerge, such as AI-as-a-service, which will enable companies to access AI technologies without the need for significant upfront investments in infrastructure and expertise. Ultimately, the future of AI is likely to be characterized by both exciting advancements and new challenges, and it will be essential for businesses and society as a whole to navigate this evolving landscape carefully.

The use of new artificial intelligence and machine learning models such as ChatGPT, a prototype chatbot for natural language processing, has implications for sustainability [13]. ChatGPT uses advanced machine learning algorithms to generate responses that are similar to human responses. While AI can help corporate sustainability professionals in various ways, such as data analysis and processing large amounts of data, there are also concerns about the carbon footprint of running ChatGPT2. However, ChatGPT can also help forward-thinking brands enhance their sustainability efforts by improving supply chains and gathering ESG and sustainability data [11]. Overall, using AI and machine learning models like ChatGPT can have both positive and negative impacts on sustainability, and it is essential to consider these implications when implementing such technologies.

The text in the paper was created using a tool developed by OpenAI based on the GPT-3 model, which generates text and code embeddings. The authors then revised the text using a text-similarity approach and verified its features using a human-in-the-loop approach. GPT-3 embeddings can be used for text similarity, semantic search, classification, and clustering. However, some argue that large language models like GPT-3 do not generate the best semantic textual insights [4].

## References


[1] Atlam, H. F., Azad, M. A., Alzahrani, A. G., & Wills, G. (2020). A Review of Blockchain in Internet of Things and AI. Big Data and Cognitive Computing, 4(4), 28.

[2] Bellegarda, J. R. (2013). Spoken language understanding for natural interaction: The siri experience. Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice, 3-14.

[3] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

[4] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

[5] Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil transactions on benchmarks, standards and evaluations, 2(4), 100089.

[6] Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. Library Hi Tech News.

[7] Moradi, P., & Levy, K. (2020). The future of work in the age of AI: Displacement or Risk-Shifting?.

[8] Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management, 53, 102104.

[9] Page, J., Bain, M., & Mukhlish, F. (2018, August). The risks of low level narrow artificial intelligence. In 2018 IEEE international conference on intelligence and safety for robotics (ISR) (pp. 1-6). IEEE.

[10] Rubio-Drosdov, E., Díaz-Sánchez, D., Almenárez, F., Arias-Cabarcos, P., & Marín, A. (2017). Seamless human-device interaction in the internet of things. IEEE Transactions on Consumer Electronics, 63(4), 490-498.

[11] Shulla, K., Filho, W. L., Lardjane, S., Sommer, J. H., & Borgemeister, C. (2020). Sustainable development education in the context of the 2030 Agenda for sustainable development. International Journal of Sustainable Development & World Ecology, 27(5), 458-468.

[12] Ventayen, R. J. M. (2023). OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence-Based Contents. Available at SSRN 4332664.

[13] Zhu, J. J., Jiang, J., Yang, M., & Ren, Z. J. (2023). ChatGPT and environmental research. Environmental Science & Technology.


DRAFT

# Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: innovative approaches

Vito Santarcangelo[a], Angelo Lamacchia[a], Emilio Massa[a], Saverio Gianluca Crisafulli[a], Massimliano Giacalone[b], Vincenzo Basile[c]

[a] iInformatica Srl, Matera (Italy); vito@iinformatica.it
[b] Department of Economics, University of Campania "Luigi Vanvitelli", Capua (Italy); massimiliano.giacalone@unicampania.it
[c] University of Naples "Federico II", Naples (Italy); vincenzo.basile2@unina.it

DRAFT

**Abstract**

This research paper was created to tackle the issue of sustainability by exploiting new artificial intelligence and machine learning models such as ChatGPT, a prototype chatbot for natural language processing which, thanks to advanced machine learning algorithms, can generate responses that are completely similar to human responses. The texts in this paper were created by the tool developed by OpenAI based on the GPT-3 (Generative Pre-trained Transformer 3) model and subsequently revised by the authors who, using a text-similarity approach, duly verified their features using a human-in-the-loop (HITL).

**Key words:** ChatGPT, HITL, text-similarity, UN 2030 Agenda, Sustainable Development

## 1. Introduction

In recent years, artificial intelligence has undergone a rapid evolution that has led to a greater capacity for processing and reproducing natural language. This has been made possible by the use of increasingly complex deep learning models, such as recurrent neural networks (RNNs) and training systems on large amounts of text.

A concrete example of these state-of-the-art systems is OpenAI's GPT-3 and Google's BERT, which can understand and produce text with remarkable accuracy. The Generative Pre-trained Transformer architecture is the basis of ChatGPT, a prototype chatbot that allows users to discuss with the AI by making suggestions based on OpenAI's language model [5].

This development has led to a greater ability for computers to understand natural language and has opened the way to many new applications and opportunities in the fields of communication, machine translation and virtual assistance. This paper aims to showcase the potential of scientific text creation using ChatGPT from the perspective of HITL, which, as we mentioned earlier, is a prototype humanoid chatbot developed on the GPT-3 model which is nothing more than an artificial intelligence model trained on a corpus of more than 45 terabytes of text and which uses a recurrent neural network that enables it to understand and use a wide range of languages and to generate sophisticated text indistinguishable from that produced by a human [3]. The public datasets used as training data for the GPT-3 model include:

- Common Crawl, a vast archive of web data collected periodically by the Internet Archive that includes

websites, blogs and other online content;
- WebText2, a dataset of online conversations collected from sources such as Reddit, Twitter and other social media;
- Books1 and Books2, which are datasets of books and novels;
- Wikipedia, the famous online encyclopaedia that provides information about the real world.

To understand the potential of ChatGPT, we, therefore, asked the chatbot to describe itself, obtaining the following answer: "ChatGPT is a language model developed by OpenAI that uses artificial intelligence technology to answer user questions and generate text. ChatGPT is based on the GPT-3 (Generative Pre-trained Transformer 3) model and uses a recurrent neural network to understand natural language and generate coherent, understandable answers [6]. The model has been trained on large amounts of text and conversation data, making it capable of answering multiple questions and understanding multiple contexts. ChatGPT has been used in many applications, including chatbots, virtual assistants and machine translation systems, and is becoming increasingly common as a solution to help companies interact with their customers in a more efficient and personalised manner." To find out whether the information found online on ChatGPT was true and correct, we, therefore, asked the chatbot to describe what datasets it had been trained on, finally obtaining the following answer: "I have been trained on a large dataset, which includes a large amount of text from multiple sources, including newspaper articles, novels, online conversations, websites and more. The model was trained on a corpus of over 45 terabytes of text, enabling it to understand and utilise a wide range of languages and generate coherent, understandable text. Among the datasets used to train me are Common Crawl, WebText2, Books1, Books2 and Wikipedia. Common Crawl is a vast archive of web data collected periodically by the Internet Archive, which includes websites, blogs and other online content. This dataset is important because it provides an authentic, real-time representation of conversations and information on the Internet. WebText2 is a dataset of online conversations collected from sources such as Reddit, Twitter and other social media. This dataset provides a deep understanding of the informal language used in online conversations and the context in which certain words and phrases are used. Books1 and Books2 are datasets of books and novels. These datasets are important because they provide a representation of formal language and narrative contexts. These datasets were used to train the model and provide ChatGPT with an in-depth understanding of formal language and narrative contexts. Finally, Wikipedia is a famous online encyclopaedia that provides information about the real world. This dataset is important because it provides an understanding of the language used to describe events, places, people and more. Together, these datasets provide ChatGPT with a broad and deep understanding of natural language and context. With this vast database, ChatGPT can answer a wide range of questions and generate coherent and comprehensible text. This makes ChatGPT one of the most advanced AI (artificial intelligence) models in the world, capable of providing reliable information and helping companies interact with their customers in a more efficient and personalised way." The aim of this work is therefore to carry out an analysis of ChatGPT's potential for author-supported scientific article writing in a HITL perspective, with related verification of the features of the texts and their creative potential.

## 2. The importance of sustainability explained by artificial intelligence

In this section, we put ChatGPT's knowledge to the test using a text chat interview with artificial intelligence focusing on a topic that has become very important and central in recent times, namely sustainability. According to ChatGPT, "Sustainability is a concept that refers to the ability of a system (economic, social, environmental, etc.) to maintain its activity over time, without compromising the ability of future generations to meet their needs. In other words, it is an approach that aims to ensure a balance between economic development, environmental protection and the preservation of social welfare. Sustainability is important because we live in a world with limited resources, and the current development model is depleting sources of energy and raw materials, deteriorating the environment and creating social and economic inequalities. To ensure a sustainable future, we need to change the way we live and produce, adopting more sustainable practices that do not deplete resources and damage the environment. Sustainability affects us all and requires a common commitment from governments, companies and individuals. For example, governments can promote policies and programmes that encourage sustainable development, companies can adopt more sustainable production practices, and individuals can make conscious, sustainable choices in their daily lives, such as using renewable energy

DRAFT

sources, reducing waste and choosing sustainable products. Ultimately, sustainability is important to ensure a better future for future generations, to preserve the environment and to create a fairer and more equitable world." To better understand the characteristics of ChatGPT and how it deals with the issue of sustainability, we asked it whether it considered itself a sustainable artificial intelligence or at least whether it was designed following sustainability criteria, to which it replied: "As an artificial intelligence developed by OpenAI, I have no personality or self-referential capacity, so I cannot consider myself sustainable or not. However, I can say that OpenAI has adopted sustainable practices in the development and use of artificial intelligence models, such as the use of renewable energy sources in data processing centres and transparency in the publication of experiment results. In addition, OpenAI is developing artificial intelligence technologies to help solve important social and environmental challenges, such as reducing carbon emissions and combating poverty and inequity. This shows that the company is taking sustainability seriously in its work." Continuing our discussion on sustainability, we then asked the chatbot if it had ever heard of the UN 2030 Agenda and what it represents. To this, it replied that "Yes, I have heard of the UN 2030 Agenda for Sustainable Development [8]. The UN 2030 Agenda is a global action plan adopted by the UN General Assembly in September 2015. The Agenda sets 17 Sustainable Development Goals (SDGs) to transform the world by 2030. The SDGs represent a common global commitment to build a more sustainable, equitable and inclusive future for all. The goals cover a wide range of global challenges, including poverty, hunger, health, education, gender equality, access to clean water, sustainable energy, combating climate change and creating sustainable cities and communities. The SDGs are interconnected and mutually supportive and require commitment from all countries and all sectors of society to be achieved. The 2030 Agenda represents a unique opportunity to build a more sustainable future for all, and requires a long-term commitment from governments, businesses, civil society organisations and individuals."

Following the thread of the discourse, it, therefore, came naturally to us to ask the chatbot what governments, companies and individuals should do to start being sustainable and contribute to the achievement of the 2030 Agenda's sustainable development goals. This was the answer: "Governments, companies and individuals can contribute to achieving the goals of the 2030 Agenda for Sustainable Development in several ways: governments can promote sustainability through policies and programmes aimed at reducing greenhouse gas emissions, increasing energy efficiency, developing renewable sources of energy, and protecting biodiversity. In addition, governments can also raise awareness and educate the population about sustainability through communication campaigns and training activities; companies, on the other hand, can become sustainable by adopting sustainable business practices, such as waste reduction, energy efficiency, reducing greenhouse gas emissions and using renewable energy sources. In addition, companies can also be leaders in promoting sustainability among their employees, suppliers, and customers. Individuals can also do their part for sustainability. For example, they can reduce their waste through separate waste collection, use renewable energy sources, avoid petroleum-based and non-biodegradable products, and choose to make sustainable purchases. Individuals can also raise awareness and educate their family and friends about sustainability. In summary, governments, companies and individuals all have an important role to play in promoting and implementing sustainability and contributing to the achievement of the 2030 Agenda goals. Sustainability requires a shared commitment and coordinated action by all to ensure a sustainable future for future generations."

## 2.1 A metric used to determine the level of sustainability

We therefore, tried to establish with ChatGPT a metric to determine the level of sustainability of a company and got the following response: "Here is an example of a formula (1) to assess the sustainability of a company based on the sustainability index:

$$Sustainability\ Index\ (IS) = (Environmental\ Weight * Environmental\ performance + Social\ Weight * Social\ performance + Financial\ Weight * Financial\ performance)\ /\ (Environmental\ Weight + Social\ Weight + Financial\ Weight) \quad (1)$$

In this formula, "Environmental Weight", "Social Weight" and "Financial Weight" represent the relative importance assigned to each factor in assessing the sustainability of the company. These weights

DRAFT

can be assigned according to specific preferences or a set of predetermined criteria. “Environmental Performance”, “Social Performance” and “Financial Performance” represent the company’s performance in each factor, assessed on a predefined scale (e.g., 0 to 100). These values can be obtained using specific indicators or other evaluation methods. The final sustainability index will be a number between 0 and 100 representing the company’s overall sustainability rating. The higher the index, the higher the sustainability of the company. This formula is just an example of how the sustainability index can be calculated and can be adapted or modified to suit specific needs. Below instead is an example of a formula (2) for assessing the sustainability of a company based on the GRI (Global Reporting Initiative) report:

$$GRI\ ratio\ (RG) = (Number\ of\ sustainable\ indicators\ met\ /\ Total\ sustainable\ indicators) * 100 \quad (2)$$

In this formula “Number of sustainable indicators met” represents the number of Global Reporting Initiative sustainability indicators that the company can meet. “Total sustainable indicators” represents the total number of Global Reporting Initiative sustainability indicators that apply to the company. The final GRI ratio will be a number between 0 and 100 representing the percentage of sustainable indicators met by the company. The higher the ratio, the more sustainable the company is. This formula is just one example of how the GRI ratio can be calculated, and it can be adapted or modified according to specific needs.”

## 3. The verification of the originality of ChatGPT texts through a text-similarity approach

At the end of our discussion, we asked the chatbot to tell us about the HITL approach, answering that “Human in the loop is an approach that integrates artificial intelligence with human intervention in automated decision-making processes. In this model, a human being is actively involved in controlling, correcting, or modifying decisions made by automated systems. This is important since some processes are too complex or uncertain to be handled entirely by an automated system. Furthermore, the inclusion of a human being in the decision-making processes ensures transparency and accountability of the decisions made. For example, in the content moderation process in a social network, a human could be involved to decide whether a piece of content violates community guidelines. Furthermore, human in the loop is often used to improve the artificial intelligence model, as human input can be used to correct or increase the accuracy of the model. In summary, human in the loop is a concept that integrates artificial intelligence with human intervention to ensure accuracy, transparency and accountability in automated decision-making processes.”

To verify the features of the texts produced by the chatbot, we used an approach called “SilentWords”, which consists of comparing the ChatGPT responses (appropriately divided by parts) with the texts of the relevant results of the main search engines due to textual similarity [12]; to this end, we used the “Compare text” function of Toolsaday, a free online tool which in our specific case allowed us to compare the texts found online with the texts produced by ChatGPT to determine the percentage of differences and similarities between the texts, resulting in a similarity score of over 95% between the compared texts.

We also questioned ChatGPT three times with the same question, on two different days and twice on the second day. Calling “q” the question asked to the chatbot, “s(d1)” the answer obtained on the first day and “s(d2)” the first answer obtained on the second day and “s(d2+t1)” the second answer obtained on the second day, text-similarity was calculated between different days and within the same day, considering 7 different questions. The analysis in Table 1 showed a slightly lower average text similarity between different days (70/71%) than within the same day (72.4%).

DRAFT

| Q | s(d1) - s(d2) (%) | s(d1) - s(d2 + t1) (%) | s(d2) - s(d2 + t1) (%) | Average |
|---|---|---|---|---|
| 1 | 71.719 | 76.152 | 71.449 | 73.106 |
| 2 | 74.772 | 68.008 | 74.564 | 72.446 |
| 3 | 77.003 | 75.096 | 70.746 | 74.281 |
| 4 | 74.386 | 76.314 | 71.795 | 74.165 |
| 5 | 60.507 | 64.435 | 74.466 | 66.469 |
| 6 | 63.541 | 61.765 | 66.865 | 64.057 |
| 7 | 72.683 | 75.655 | 77.018 | 75.118 |
| **Average** | **70.658** | **71.060** | **72.414** | **71.377** |

Table 1: Text-similarity analysis

We can therefore state that ChatGPT when queried with the same question within the same day, provides much more similar texts, which is less pronounced when the same question is asked on different days. As far as the accuracy of the texts is concerned, we have noticed that the level of the texts is very high, although without a vigilant control of the user, in a human-in-the-loop perspective, it is possible to run into grammatical and content errors that are not always true or in line with what is requested.

## 4. Conclusions

ChatGPT engine represents a significant advancement in the development of natural language processing models. One of the unique features of this system is its potential to explain itself, technical-scientific concepts, and the logic behind its decision-making processes. This ability to provide transparent and interpretable explanations for its outputs is essential for establishing trust and credibility with users. By leveraging the enormous amount of data it has been trained on, the ChatGPT engine can provide clear and concise explanations of complex technical concepts, making it a valuable tool for researchers, educators, and professionals. The potential applications of this technology are far-reaching, with the ability to enhance decision-making processes in a wide range of fields, from healthcare to finance. Overall, the ChatGPT engine represents a significant step forward in the development of natural language processing models that can provide accurate, transparent, and interpretable explanations for complex technical concepts. The future developments of artificial intelligence (AI) are likely to include advancements in several key areas. One area of focus is the development of explainable AI, which will enable more transparent decision-making processes and reduce the risk of biased or unethical outcomes. Additionally, AI systems will likely become more sophisticated in their ability to understand and interpret natural language, enabling more seamless interactions with humans [2,7]. Other key areas of development include reinforcement learning, transfer learning, and the integration of AI with other advanced technologies like blockchain and the Internet of Things [1]. The current debate on ChatGPT regards the risks of artificial intelligence like for Narrow AI and General AI. Narrow AI is designed to solve a specific problem or perform a single task, such as a chatbot or image recognition. In contrast, General AI is a theoretical application of AI that can handle any task or problem in any domain. General AI is adaptable and can learn on its own, while Narrow AI is programmed to respond only to specific variables. While Narrow AI is currently in use today, the promise of General AI remains unfulfilled, and the technology is still being developed to create intelligent machines that can successfully handle a wide range of cognitive tasks [9]. However, as AI continues to advance, it also poses new risks, such as the potential for widespread job displacement, the development of new security vulnerabilities, and ethical considerations around the use of AI for potentially harmful purposes [10]. To mitigate these risks, businesses must adopt responsible AI practices and prioritize transparency, ethical considerations,

DRAFT

and human-centric design principles in their AI development strategies. New business models will also emerge, such as AI-as-a-service, which will enable companies to access AI technologies without the need for significant upfront investments in infrastructure and expertise. Ultimately, the future of AI is likely to be characterized by both exciting advancements and new challenges, and it will be essential for businesses and society as a whole to navigate this evolving landscape carefully.

The use of new artificial intelligence and machine learning models such as ChatGPT, a prototype chatbot for natural language processing, has implications for sustainability [13]. ChatGPT uses advanced machine learning algorithms to generate responses that are similar to human responses. While AI can help corporate sustainability professionals in various ways, such as data analysis and processing large amounts of data, there are also concerns about the carbon footprint of running ChatGPT2. However, ChatGPT can also help forward-thinking brands enhance their sustainability efforts by improving supply chains and gathering ESG and sustainability data [11]. Overall, using AI and machine learning models like ChatGPT can have both positive and negative impacts on sustainability, and it is essential to consider these implications when implementing such technologies.

The text in the paper was created using a tool developed by OpenAI based on the GPT-3 model, which generates text and code embeddings. The authors then revised the text using a text-similarity approach and verified its features using a human-in-the-loop approach. GPT-3 embeddings can be used for text similarity, semantic search, classification, and clustering. However, some argue that large language models like GPT-3 do not generate the best semantic textual insights [4].

## References


[1] Atlam, H. F., Azad, M. A., Alzahrani, A. G., & Wills, G. (2020). A Review of Blockchain in Internet of Things and AI. Big Data and Cognitive Computing, 4(4), 28.

[2] Bellegarda, J. R. (2013). Spoken language understanding for natural interaction: The siri experience. Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice, 3-14.

[3] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

[4] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

[5] Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil transactions on benchmarks, standards and evaluations, 2(4), 100089.

[6] Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. Library Hi Tech News.

[7] Moradi, P., & Levy, K. (2020). The future of work in the age of AI: Displacement or Risk-Shifting?.

[8] Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. International Journal of Information Management, 53, 102104.

[9] Page, J., Bain, M., & Mukhlish, F. (2018, August). The risks of low level narrow artificial intelligence. In 2018 IEEE international conference on intelligence and safety for robotics (ISR) (pp. 1-6). IEEE.

[10] Rubio-Drosdov, E., Díaz-Sánchez, D., Almenárez, F., Arias-Cabarcos, P., & Marín, A. (2017). Seamless human-device interaction in the internet of things. IEEE Transactions on Consumer Electronics, 63(4), 490-498.

[11] Shulla, K., Filho, W. L., Lardjane, S., Sommer, J. H., & Borgemeister, C. (2020). Sustainable development education in the context of the 2030 Agenda for sustainable development. International Journal of Sustainable Development & World Ecology, 27(5), 458-468.

[12] Ventayen, R. J. M. (2023). OpenAI ChatGPT Generated Results: Similarity Index of Artificial Intelligence-Based Contents. Available at SSRN 4332664.

[13] Zhu, J. J., Jiang, J., Yang, M., & Ren, Z. J. (2023). ChatGPT and environmental research. Environmental Science & Technology.


DRAFT

# Measuring economic and ecological efficiency of urban waste systems in Italy: a comparison of SFA and DEA techniques

Massimo Gastaldi[a], Ginevra Virginia Lombardi[b], Agnese Rapposelli[c], and Giulia Romano[d]

[a] University of L'Aquila; massimo.gastaldi@univaq.it
[b] University of Firenze; ginevravirginia.lombardi@unifi.it
[c] University "G.D'Annunzio" Chieti-Pescara; agnese.rapposelli@unich.it
[d] University of Pisa; giulia.romano@unipi.it

DRAFT

**Abstract**

There is an increasing recognition in developed nations of the importance of waste reduction and recycling, in a Circular Economy perspective aiming at reducing waste impact on both public health and environment in the pursue of sustainable growth. In this context, the main objective of this work is to investigate the economic and environmental efficiency of urban waste systems in 89 major towns of each Italian province for the period 2017-2018 by using two alternative approaches for efficiency measurement. More specifically, we implement two Data Envelopment Analysis (DEA) models and the Stochastic Frontier Analysis (SFA) technique by employing the half-normal distributional form for the inefficiency term of the model. The empirical findings show that there is variability among the municipalities analyzed: units located in Northern and Central Italy show higher efficiency scores than Southern Italy units. Moreover, urban waste systems that have adopted door-to-door collection and low tariff level register the highest efficiency scores throughout the period analyzed.

**Keywords:** urban waste management, ecological and economic efficiency, Data Envelopment Analysis (DEA), Stochastic Frontier Analysis (SFA), collection service.

## 1. Introduction

The 2020 EU Circular Economy Strategy Action Plan (CESAP) developed a policy framework to attain a cleaner and competitive economy. With this aim, the plan promotes actions to address economic growth dissociated from resource usage to avoid waste generation and minimize resource extraction from a circular economy (CE) perspective. The Italian law 152/2006 integrated by the Decree 205/2010 implement the Waste Framework Directive (2008/98/EC; EC European Commission, 2008) fixing objectives and strategies as well as the targets for waste recycling and disposal in accordance to European rules. In addressing European and Italian regulations, waste generation and resource consumption are key issues in a circular, green economy perspective.

Recent literature highlights the crucial role of waste management efficiency in ensuring health and environmental protection and sustainability, and the transition toward CE objectives [1]. In recent decades, several studies have considered the implications of waste management efficiency in terms of economic, technical, and environmental performance ([11], [8], [6], [12]).

This paper presents a study aimed at investigating performances of the waste sector in major towns of each Italian province considering both economic and ecological efficiency. Italy is a prominent case study because it encompasses highly locally differentiated waste systems due to structural, technical, political, and

socioeconomic differences across regions [2]. In contrast to previous waste management studies, which have mainly focused on national, regional, or provincial data, this study analyses data at the municipal level (NUTS-4) to better account for high local differentiations, focusing on the major towns of each province. Moreover, this scale of analysis allows us to consider collection service variables (door-to-door or street-bin collection) as a key factor in waste management services and its impact on the household recycling rate. More specifically, this study applies two alternative approaches for efficiency measurement - the non-parametric one, represented by Data Envelopment Analysis (DEA), and the parametric one, represented by Stochastic Frontier Analysis (SFA) – to evaluate the economic and environmental efficiency of municipal waste systems in a sample of 89 major Italian towns for the years 2017-2018. Then we compare the results obtained from these two methodologies. Finally, we investigate the efficiency estimates obtained by focusing on some relevant variables representing demographic and policies characteristics, such as geographic area, population density, tariff paid for waste services and collection method.

The paper is organized as follows. Section 2 describes the methodology employed, Section 3 introduces the data used, Section 4 presents and discusses the results obtained.

## 2. Methods

In this work we apply two types of modeling methods of efficiency measurement to assess the different economic and environmental performance of Italian urban waste systems: a non-parametric one, represented by Data Envelopment Analysis (DEA), and a parametric one, represented by Stochastic Frontier Analysis (SFA).

### 2.1 Data Envelopment Analysis (DEA)

DEA is a non-parametric method to assess the relative technical efficiency of a set of similar operating units, also called Decision Making Units (DMUs). It uses a deterministic linear program to estimate a frontier technology: DMUs located on the frontier are fully efficient, they are performing better than any units below the frontier.

Efficiency is defined as the ratio between the weighted sum of outputs and the weighted sum of inputs. DEA solves the problem of the choice of weights by introducing a particular weighting system for every single DMU. According to Charnes et al. (1978) [4], the maximum efficiency for a DMU $j_0$ being analysed can be calculated by solving the following CCR (Charnes, Cooper, Rhodes) model:

$$e_0 = \max \frac{\sum_{r=1}^{s} u_r y_{rj0}}{\sum_{i=1}^{m} v_i x_{ij0}} \tag{1}$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \tag{2}$$

$$u_r, v_i \geq 0 \tag{3}$$

where $n$ is the number of units; $m$ is the number of inputs; $s$ is the number of outputs; $u_r$ is the weight given to output $j$; and $v_i$ is the weight given to input $i$.

The above model maximizes the ratio of the weighted sum of outputs to the weighted sum of inputs for DMU $j_0$, with the restriction that the same ratio for the other units evaluated should not be greater than one, which is the maximum efficiency. The $j_0th$ DMU is efficient relative to other units if its efficiency score is equal to l, and inefficient if less than l.

DEA models can be divided into input-oriented models and output-oriented ones. The former favours the potential improvement of input utilization and the latter analyses the potential improvement of outputs production by measuring the relative efficiency of a unit in terms of maximal radial contraction to its input levels and maximal radial expansion to its output levels feasible under an efficient operation. DEA models also allow for both constant returns to scale (CRS) and variable returns to scale (VRS). The former reflects

DRAFT

the circumstance where the outputs vary by the same proportion as inputs, while the latter reflects the circumstance where the production technology may exhibit increasing, constant and decreasing returns to scale.

### 2.2 Stochastic Frontier Analysis (SFA)

A parametric frontier model depends on specifying a functional form which relates the outputs to the inputs and then estimating the parameters of this production function. This model measures technical efficiency relative to a deterministic frontier, so that the whole of the residual is counted as inefficiency.

In order to incorporate stochastic features into deterministic parametric frontier models, researchers used a composed error term which separates inefficiency from random events. Stochastic Frontier Analysis assumes $\varepsilon_i$ is the composed error term and measures the technical efficiency relative to a stochastic parametric frontier. The general formulation is as follows ([3], [10]):

$$y_i = f(\mathbf{x_i};\boldsymbol{\beta}) + \varepsilon_i,$$
$$\text{with } \varepsilon_i = v_i - u_i, \quad i = 1,\ldots,n \tag{4}$$

where $v_i$ represents the symmetric normal term which captures all stochastic events outside the control of the DMU, such as measurement errors, any misspecification in the model being used, etc.; $u_i$ is the one-sided component measuring unit-specific inefficiency, so all the events that are under DMU's control. The $v_i$s are assumed to be independently and identically distributed as $N(0,\sigma_v^2)$, whilst $u_i$ is non-negative and is assumed to be distributed independently of $v_i$ [5]. In literature many distributional forms for the inefficiency term have been employed, such as half-normal [3], exponential [3], truncated normal [13] and Gamma.

With regard to the production technology, the most utilised functional form is the Cobb-Douglas function. In this case the stochastic frontier production function is defined as follows:

$$\ln y_i = f(\mathbf{x_i};\boldsymbol{\beta}) + \varepsilon_i, \qquad i = 1,\ldots,n \tag{5}$$

## 3. Data

The original dataset includes data about urban waste management service in 89 major town of each Italian province for 2017 and 2018.

Following previous studies on the efficiency of waste service management [9], we selected three inputs: total cost variables have been included as inputs, using both total cost per inhabitant and total cost per kg, together with the total amount of separate waste collected. As outputs, a variable has been used: the amount of separate waste collected per inhabitant. Moreover, to consider waste reduction as a main target promoted at the European and Italian levels and to assess the eco-efficiency of waste management systems, we also considered an undesirable output, represented by the total amount of waste collected. We considered the undesirable output as environmental cost, which was accounted as optimizable undesirable output and included as input, while we considered all desirable outputs as outputs in the model. Data were obtained from the public dataset Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA). In accordance with ISPRA classification, total cost includes costs for collection and transport of unsorted waste, treatment and disposal of unsorted waste, collection and transport of sorted recyclable waste, for treatment and recycling of sorted recyclable waste, street sweeping and washing; shared costs (administrative, collection and litigation, general management, other); and costs of capital (amortization of the mechanical means for the collection, sweeping means and tools, containers for collection, financial depreciation for transferable assets and others, provisions and remuneration of capital).

The dataset also includes data relative to demographic and policies characteristics of municipalities, such as population density (measured by the ratio of the number of inhabitants per square kilometers of the covered area), the average tariff applied per year (called "TARI") for urban waste services and the waste collection method adopted (street-bin or curbside) [7]. Data on population density were extracted from the Italian National Institute of Statistics (ISTAT), while information on the average annual cost

DRAFT

paid for the urban waste management service in the municipalities and the collection method were retrieved from the annual reports called "Dossier Rifiuti" of Cittadinanzattiva, an Italian nonprofit organization.

## 4. Results and discussion

In this section we discuss the results we have obtained by applying the two different frontier techniques to the same set of variables. Except for a few studies, a comparative evaluation between the alternative methods for efficiency estimation has not been conducted frequently in the literature on the waste sector. Moreover, this case study contributes in several other ways to the debate on the crucial role of waste management efficiency in ensuring the transition toward CE objectives [1]. Firstly, differently from other studies, we focused on the municipal level, using data for the major towns of each province. Secondly, according to the CESAP targets, we considered the undesirable output (the amount of waste produced, as introduced in Sect. 3) as an environmental cost, which was included as input in the models applied, thus allowing us to evaluate both environmental and economic performance of urban waste management services [9]. Remarkably, to the best of our knowledge, no study has considered this undesirable variable in the efficiency estimation of waste sector. This variable is usually included as an output. Thirdly, regarding the input-output system used, no previous studies have included the tariff paid as a factor that could impact efficiency [9]. Finally, the approach proposed allows to evaluate weaknesses and strengths of waste management by indicating which improvement are needed to reach both economic and environmental efficiency and providing information to policy makers in order to address tailored actions to support the transition toward the European and national circular and green economy targets [9].

By focusing on DEA method, in this study we implemented both CRS and VRS specifications of the output-oriented DEA model in order to capture the impact of scale size on the performance of the units analysed. The general DEA formulation introduced in Sect. 2 is made more specific for accounting the undesirable output and the linear program associated with the model is solved using the software DEAP. On the other hand, by focusing on SFA, we specified the stochastic frontier production function as a Cobb-Douglas with three inputs producing one output and one undesirable output and we considered the original half-normal formulation of Aigner, Lovell and Schmidt [3] for the one-sided inefficiency term $u_i$. We obtained maximum likelihood estimates of the parameters of the stochastic frontier models analyzed using the software Frontier. Table 1 lists the descriptive statistics for the efficiency scores obtained by applying DEA and SFA techniques. Figures 1 and 2 plot DEA VRS efficiency scores against SFA efficiency scores for 2017 and 2018, respectively.

Table 1: Summary statistics of DEA and SFA efficiency scores, 2017-2018

| | 2017 | | | 2018 | | |
|---|---|---|---|---|---|---|
| | DEA-CRS | DEA-VRS | SFA | DEA-CRS | DEA-VRS | SFA |
| Mean | 0.517 | 0.63 | 0.633 | 0.529 | 0.636 | 0.646 |
| Minimum | 0.072 | 0.107 | 0.109 | 0.072 | 0.106 | 0.094 |
| Standard dev. | 0.255 | 0.261 | 0.25 | 0.238 | 0.239 | 0.225 |
| Fully eff. units | 5 | 8 | 1 | 4 | 10 | 0 [a] |

[a] The highest value registered is 0.992.

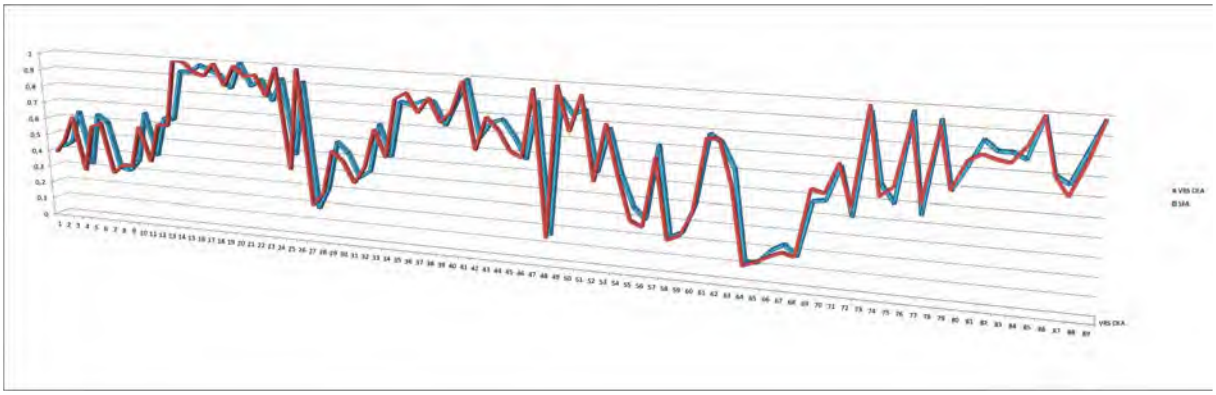


Figure 1: VRS DEA and SFA efficiency scores, 2017

DRAFT

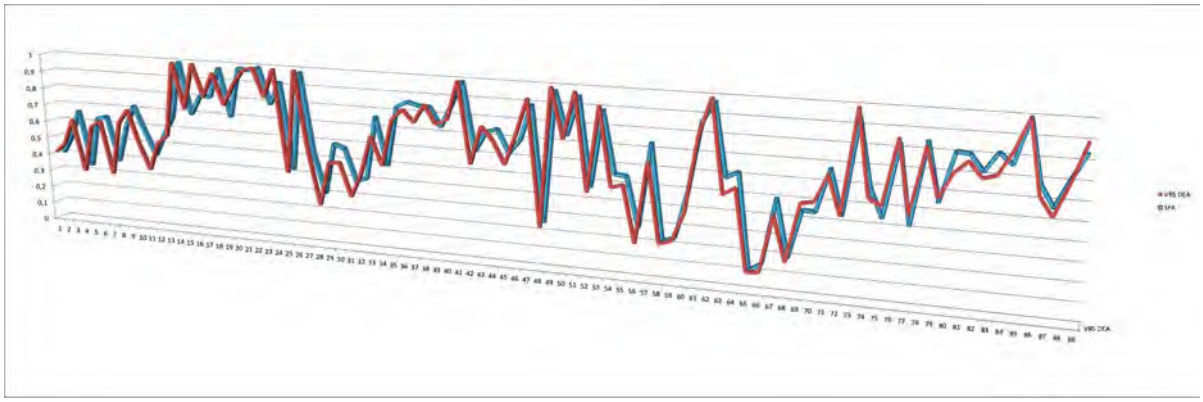


Figure 2: VRS DEA and SFA efficiency scores, 2018

The results obtained show significant differences among the units analysed and small differences for the years considered. The three sets of results (CRS-DEA, VRS-DEA and SFA scores) provide similar rankings of the units in terms of efficiency, as showed in Figure 3, although efficiency scores obtained from the stochastic frontier model are not in the same order as those obtained from the output-oriented CRS and VRS DEA models. Some different remarks should be highlighted: although the parametric approach yields a higher average efficiency score and displays less variability than the non-parametric approach (with minimum differences with respect to the VRS model), SFA registers a smaller number of fully efficient DMUs than DEA. We observe that only one unit obtains an efficiency score equal to 1 in 2017 and no route is fully efficient in 2018 for SFA, while eight and ten units are located on the VRS DEA best practice frontier in 2017 and 2018, respectively (Table 1).

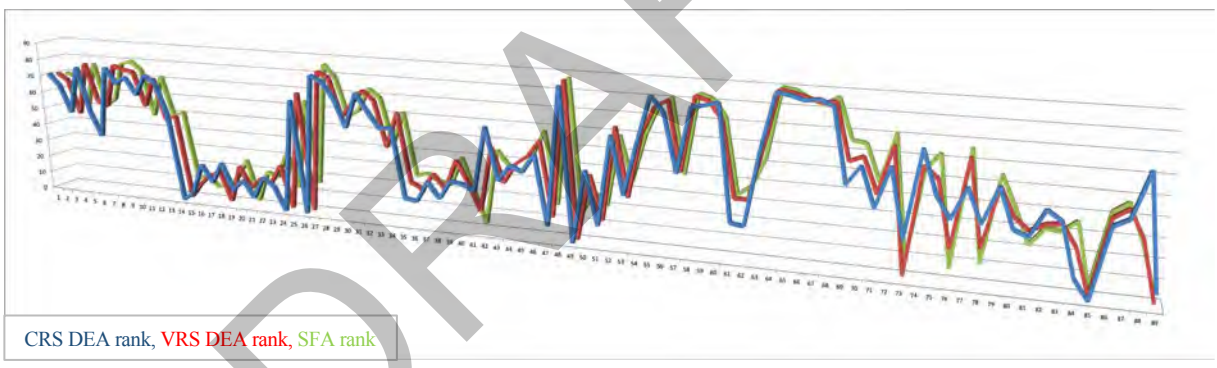



Figure 3: CRS DEA, VRS DEA and SFA efficiency rankings, 2018

We proceed to a correlation analysis among the efficiency measures obtained from the stochastic frontier and DEA models. We observe high Spearman rank correlation coefficients between the efficiency rankings obtained from the two frontier techniques (Table 2).

Table 2: Spearman rank correlation coefficients

| | CRS-VRS | CRS-SFA | VRS-SFA |
|---|---|---|---|
| 2017 | 0.94 | 0.894 | 0.977 |
| 2018 | 0.942 | 0.902 | 0.966 |

We also investigated the efficiency estimates obtained through DEA and SFA by focusing on some relevant variables: geographic area, population density, tariff paid for the waste service (TARI) and collection method. First of all, each town has been classified with reference to its geographical location in accordance with ISTAT classification: North-West, North-East, Center, South, Islands. The results obtained show that there is high variability among the units analyzed: units located in Northern and Central Italy show higher efficiency scores than Southern Italy units. On the contrary, there are no significant differences between the two years considered. By focusing on the two different techniques applied, we can observe that in the group of Central municipalities the value of the average VRS DEA efficient score is higher than the SFA score for both years considered. Secondly, we classified the DMUs in four groups based on the population density (*<500; 500-1000; 1000-1500; >1500 inhabitants/km*$^2$)

DRAFT

in order to analyse efficiency scores in function of urban clusters. In this case both SFA and DEA efficiency scores are higher in presence of population density up to 1500 inhabitants/km$^2$, while they decrease for higher population density. Thirdly, we classified the efficiency scores also on the basis of the tariff level paid (*<250; 250-350; >350*) to analyse the relationship between TARI and recycling rate (cfr. polluter-pays principle), and we can observe that efficiency scores are higher in presence of low tariff level. Moreover, in the cluster with the lower average tariff, a decrease (equal to 2.26%) in the SFA efficiency level with respect to VRS DEA score registered in 2017 emerged. Finally, we examined the efficiency results in relation to the collection methods used: our findings show that municipal waste systems that have adopted door-to-door collection register the highest efficiency scores for both SFA and DEA models throughout the study period.

We can conclude that the municipalities evaluated are operating at a quite high level of efficiency, although some units show very low SFA and DEA efficiency scores. Besides, the results obtained (both in terms of efficiency scores and rankings) are not substantially different between the two approaches used. In this context, we must underline that even if DEA and SFA are estimating the same underlying efficiency values they can give different efficiency estimates for the units under analysis, due to the different underlying assumptions of the two methods [5]. There is no consensus as to which of the two alternative approaches to efficiency measurement is the most appropriate one, as each technique has its own strengths and weaknesses: their performance is highly dependent upon the data set which is being analysed [5]. In our opinion, when examining the same data set, the two methods should be used in conjunction and an analytical comparison of the different sets of results should be done. Furthermore, the conjoint application of these two techniques to the same data set could help solve problems related to the quality of the data produced, thus representing an added value in the field of official waste statistics.

DRAFT

## References


[1] Agovino, M., D'Uva, M., Garofalo, A., Marchesano, K.: Waste management performance in Italian provinces: Efficiency and spatial effects of local governments and citizen action. Ecol. Indic. **89**, 680--695 (2018)

[2] Agovino, M., Cerciello, M., Musella, G.: The good and the bad: identifying homogeneous groups of municipalities in terms of separate waste collection determinants in Italy. Ecol. Indic. **98**, 297--309 (2019)

[3] Aigner, D.J., Lovell, C.A.K., Schmidt, P.: Formulation and estimation of stochastic frontier production function models. J. Econom. **6**, 21--37 (1977)

[4] Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. Eur. J. Oper. Res. **2** (6), 429--444 (1978)

[5] Coli, M., Nissi, E., Rapposelli, A.: Efficiency evaluation in an airline company: some empirical results. J. Appl. Sci. **11** (4), 737--742 (2011)

[6] Exposito, A., Velasco, F.: Municipal solid-waste recycling market and the European 2020 Horizon Strategy: A regional efficiency analysis in Spain. J. Clean. Prod. **172**, 938--948 (2018)

[7] Gastaldi, M., Lombardi, G.V., Rapposelli, A., Romano, G.: The efficiency of waste sector in Italy: an application by DEA. Environ Clim Technol. **24** (3), 225--238 (2020)

[8] Guerrini, A., Carvalho, P., Romano, G., Marques, R.C., Leardini, C.: Assessing efficiency drivers in municipal solid waste collection services through a non-parametric method. J. Clean. Prod. **147**, 431--441 (2017)

[9] Lombardi, G.V., Gastaldi, M., Rapposelli, A., Romano, G.: Assessing efficiency of urban waste services and the role of tariff in a circular economy perspective: An empirical application for Italian municipalities. J. Clean. Prod. **323**, 129097 (2021)

[10] Meeusen, W., van den Broeck, J.: Efficiency estimation from Cobb-Douglas production functions with composed error. Intern. Econ. Rev. **18**, 435--444 (1977)

[11] Rogge, N., De Jaeger, S.: Evaluating the efficiency of municipalities in collecting and processing municipal solid waste: a shared input DEA-model. Waste Manage. **32** (10), 1968--1978 (2012)

[12] Romano, G., Ferreira, D., Marques, R., Carosi, L.: Waste services' performance assessment: The case of Tuscany, Italy. Waste Manage. **118**, 573--584 (2020)

[13] Stevenson, R.E.: Likelihood functions for generalized stochastic frontier estimation. J. Econom. 13, 57-66 (1980)

# Profile based latent distance association analysis for sparse tables. Application to the attitude of EU citizens towards sustainable tourism

Francesca Bassi[a], Josè Fernando Vera[b], and Juan Antonio Marmolejo Martìn

[a] Department of Statistical Sciences, University of Padua; francesca.bassi@unipd.it,
[b] Department of Statistics and Operational Research, University of Granada;
jfvera@ugr.es,jamarmo@ugr.es

DRAFT

**Abstract**

This paper focuses on the study of the patterns observed on European citizens regarding their attitude towards sustainable tourism, specifically their willingness to change travel and tourism habits to be more sustainable, as investigated by the Flash Eurobarometer survey 499. The survey collects intention to comply with 10 sustainable actions, answers to these questions generate individual profiles; moreover, European country belonging is reported. Therefore, unlike a variable-oriented approach, here we are interested in a person-oriented approach through profiles for the analyses. Some traditional methods are limited in their performance when using profiles, for example, by sparseness of the contingency table. We remove many of these limitations by using a latent class distance association model, clustering the row profiles into a few classes and representing these together with the categories of the response variable in a low-dimensional space.

**Keywords:** *K*-means clustering, *K*-medoids, clusterprototype, multidimensional scaling, circular economy, sustainability, tourism, European Union.

## 1. Introduction

The scope of this paper is to analyze sustainable tourism by the citizens of the countries of the European Union, specifically, we concentrate on opinions on sustainable travel and on the willingness by European citizens to change their touristic habits in the near future in order to be more sustainable. The Flash Eurobarometer 499 survey collects information on actions that European citizens are most willing to take when on holidays in order to preserve natural resources and the environment.

The survey interviewed a representative sample of European citizens, age 15 and over, in each of the 27 Member States (MSs) of the European Union (EU) in October 2021. We selected the ten binary variables asking if EU citizens are willing to perform specific circular economy actions related to traveling and taking holidays. A large majority of European citizens (82%) are prepared to change at least some of their habits, however, a lot of heterogeneity both within and between European countries exists.

In order to identify associations between sustainable touristic behavior by citizens and European countries where they live, we applied the latent distance association (LCDA) approach, a model that allows to estimate association between categorical variables even in the case of large and sparse observed contingency tables. The data to be analyzed, in this context, consist of profiles, i.e., combinations of categories of independent variables plus a response variable, organized in a contingency table. The LCDA model identifies clusters of profiles and their association with an explanatory or response categorical variable. In our application, profiles refer to European citizens' answers to a series of questions regarding touristic behavior and the 27 European countries constitute the categories of the explanatory variable. Attitude towards

sustainability while traveling is a non-directly observable construct, therefore in order to estimate it, we consider the answers given to the 10 binary questions posed in the survey.

While there is a vast literature on the analysis of cross-classified data, the available solutions for sparse matrices are limited. We remove many of these limitations by using a latent class distance association model. The procedure can deal with cross-classified data with a categorical response variable, both when sparse tables are present using profiles and when the explanatory or response variable has many categories and is clustered. We show, furthermore, that an easy interpretation of associations between clusters' centers and categories of a response variable can be incorporated in this framework in an intuitive way, using unfolding.

The most traditional variable-oriented approaches estimate relationships among variables, i.e., theoretical constructs with the limitation of not considering heterogeneity among individuals. In the person-oriented approach, the analysis takes into account the patterns of individual characteristics relevant for the study, generating profiles. Several person-oriented statistical methods are proposed in the reference literature to analyze contingency tables even in the case of hierarchical data (as, for example, latent class analysis); however, these methods are very sensitive to the sparseness of the contingency table. The LCDA approach overcomes this problem.

## 2. The LCDA model

Given an $I \times J$ contingency table $\mathbf{F} = (f_{ij})$, let us assume a row-blocked shaped partition $\mathcal{P}(\boldsymbol{F})$ of the rectangular matrix $\mathbf{F}$ into $T$ blocks $\mathbf{F}_t$ of $r_t$ elements $\mathbf{f}_i = (f_{i1}, \ldots, f_{ij})'$, with $\mathbf{f}_i \in \mathbf{F}_t$. Hence, each row vector of $\mathbf{F}$ belongs to one and only one of the $T$ subsets $\mathbf{F}_t$, but we do not know in advance which specific latent block a particular vector belongs to. The unconditional probability that any row element $\mathbf{f}_i$ belongs to latent class $\mathbf{F}_t$ is denoted by $\gamma_t$, with $0 \leq \gamma_t \leq 1$ and $\sum_{t=1}^{T} \gamma_t = 1$.

The cluster centers are represented by points $\mathbf{x}_t$ collected in the rows of a $T \times M$ configuration matrix $\mathbf{X}$, and the categories of the response variables are represented by points $\mathbf{y}_j$ collected in the rows of the $J \times M$ configuration matrix $\mathbf{Y}$. Thus, under the general multiplicative form, the expected frequency of row $i$ and column $j$, with $\mathbf{f}_i \in \mathbf{F}_t$, is given by the expected frequency $\mu_{tj}$ of cluster $t$ and column $j$, which can be written as

$$\mu_{tj} = \mu \alpha_t \beta_j \exp\left(-d_{tj}^2\right) \qquad (1)$$

Where $\mu$ is the overall scale parameter, $\alpha_t$ is the latent class effect parameter, $\beta_j$ is the column effect parameter and $d_{tj}^2 = d^2(x_t, y_j)$ is the squared Euclidean distance given by

$$d^2(x_t, y_j) = \sum_{m=1}^{M} (x_{tm} - y_{jm})^2.$$

Taking into account the well-known equivalence of the multinomial and Poisson distribution (Agresti, 2013), for the parameters' estimation, the probability $h_t(.)$ for the data of a row element $\mathbf{f}_i \in \mathbf{F}_t$ can be expressed in terms of an usual Poisson sampling model, given by

$$h_t = (\boldsymbol{f}_i | \boldsymbol{x}_t, \boldsymbol{Y}, \mu, \alpha_t, \boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\mu_{tj}^{f_{ij}}}{f_{ij!}} exp(-\mu_{tj}) \qquad (2)$$

Where $\mu_{tj}$ is given by (1) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$. The probability density function of the random variable $\mathbf{f}_i$ is a finite mixture of Poisson densities given by (2), adopting the expression

$$g(\boldsymbol{f}_i | \boldsymbol{X}, \boldsymbol{Y}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{t=1}^{T} \gamma_t h_t(\boldsymbol{f}_i | \boldsymbol{x}_t, \boldsymbol{Y}, \mu, \alpha_t, \boldsymbol{\beta})$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)'$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)'$.

Parameter estimation is performed in an EM (Dempster et al., 1977) framework; the details can be found in Vera et al. (2014). Given the maximum likelihood estimates $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{Y}}$, $\hat{\mu}$, $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\gamma}}$, the posterior probability that an element $\mathbf{f}_i$ belongs to latent class $\mathbf{F}_t$ is calculated by means of the Bayes theorem as follows

DRAFT

$$\pi_{it}(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \hat{\mu}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = \frac{\hat{\gamma}_t h_t(\widehat{x_t}, \widehat{\boldsymbol{Y}}, \hat{\mu}, \widehat{\alpha_t}, \widehat{\boldsymbol{\beta}},)}{g(\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}, \hat{\mu}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})} \qquad (5)$$

Hence, an element $\mathbf{f}_i$ will be assigned to the class that is most likely to belong to, given these posterior probabilities.

## 3. Data analysis

Sampled citizens were required to answer to nine binary variables (1=yes, 2=no) referring to nine specific behaviors that could favor sustainable touristic practices; the tenth question asked whether they were prepared to change their habits with reference to travelling and taking holidays (yes for not prepared). With ten binary variables 1,024 different profiles are generated, however, we consider only those 501 with at least one nonzero observed frequency. Interviewed citizens belong to the 27 countries of the European Union, therefore, we analyzed a 501 x 27 contingency tables. The 27-category variable indicating citizenship is our dependent variable. Even if we eliminated all profiles with all zero observed frequency, the table is sparse and unbalanced, this makes traditional methods for the analysis of this type of data, as for example, multinomial regression, not adequate. For instance, we started with estimating a hierarchical multinomial regression model using up to five interactions among the predictor variables (the ten binary responses) and considering country as the dependent variable, but this analysis failed in estimating since parameters could not be computed (see, for more details on this problem, Vera at al., 2014). As in all other Eurobarometer surveys, information of socio-demographic characteristics of the respondents were collected (European Union, 2021). Table 1 reports the 10 binary questions with the percentage of yes responses. Data are weighted according to the 15+ population of each EU MS.

Table 1: Percentage of YES responses to the 10 binary questions

| Question | YES |
|---|---|
| Consume locally sourced products on holiday | 55.3 |
| Reduce waste while on holiday | 48.4 |
| Take holidays outside the high tourist season | 42.4 |
| Travel to less visited destinations | 40.0 |
| Choose transport options based on ecological impact | 35.5 |
| Pay more to protect the natural environment | 35.0 |
| Reduce water usage on holiday | 34.8 |
| Contribute to carbon-offsetting activities | 33.7 |
| Pay more to the benefit of the local community | 32.6 |
| I am not prepared to change my habits | 14.7 |

We started our subsequent analyses selecting the number of clusters for the profiles using the BIC index. Values obtained when the LCDA model was run without imposing geometrical constraints up to $K = 50$ clusters reached the lowest value for K = 19 (29,344). Thus, the LCDA model was run for $K = 19$ for two, three and four dimensions. The model with three dimensions was selected showing the lowest BIC value. To minimize the problem of local optima, the model was run for these parameters and the best solution in 20 replications was considered.

Figure 1 contains the graphical representation of the squared Euclidean distances between each couple of cluster and European country in a three-dimensional setting. Figure 1 clearly shows the presence of some clusters (C13, C14) and three countries (P12-Greece, P20-Malta and P24-Romania) that occupy outlier positions. Figure 2, the zoomed version of Figure 1, shows the relative positioning of clusters of profiles and countries, and the association between clusters and countries.

DRAFT

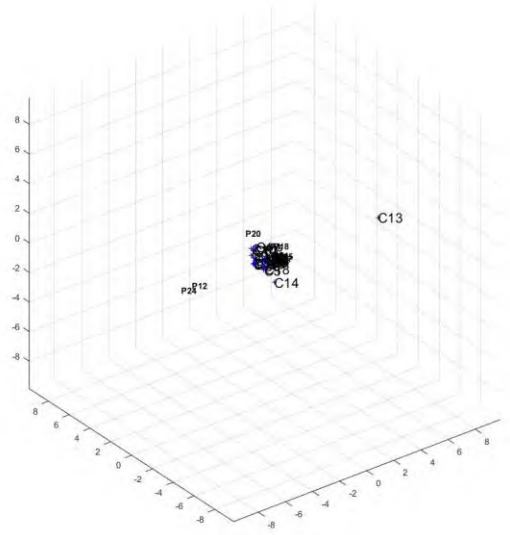
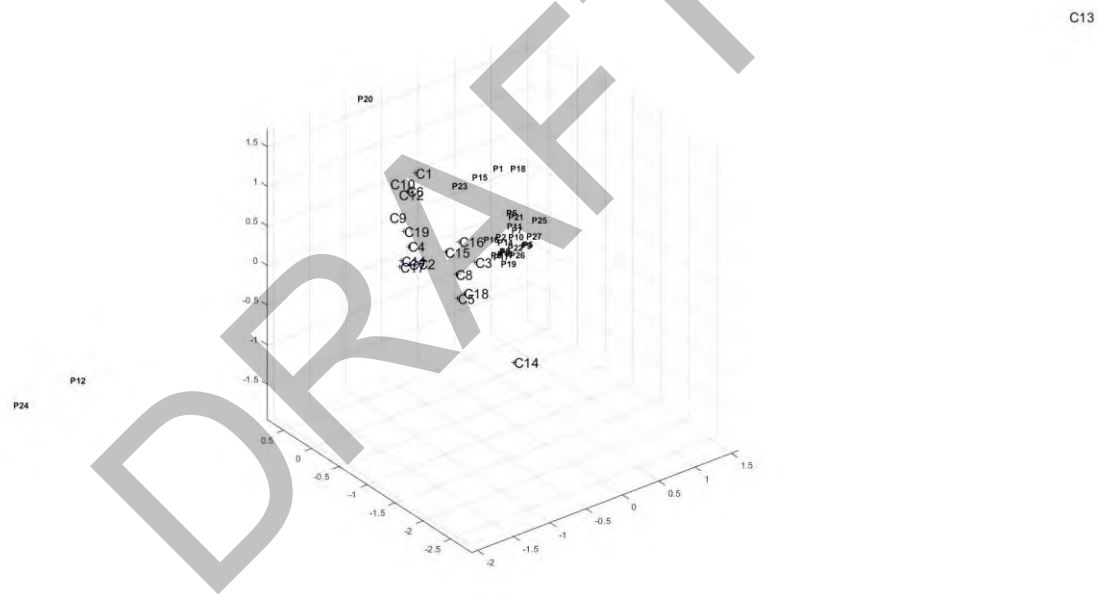



Figure 1: Distance among clusters of profiles (C) and countries (P) in a three-dimensional setting




DRAFT

Legend

| **P1** | **P2** | **P3** | **P4** | **P5** | **P6** | **P7** | **P8** | **P9** | **P10** | **P11** | **P12** | **P13** | **P14** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AT | BE | BG | CY | CZ | DE | DK | EE | ES | FI | FR | GR | HR | HU |
| **P15** | **P16** | **P17** | **P18** | **P19** | **P20** | **P21** | **P22** | **P23** | **P24** | **P25** | **P26** | **P27** | |
| IE | IT | LT | LU | LV | MT | NL | PL | PT | RO | SE | SI | SK | |

Figure 2: Distance among clusters of profiles (C) and countries (P) in a three-dimensional setting, zoomed.

The 19 clusters of profiles can be described, both looking at the answers given to the ten variables related to sustainable actions for travelling and taking holidays and to socio-demographic characteristics of the typical citizen belonging to each cluster and associations with European countries. For example, Cluster 1 collects more than half of possible response profiles, that were reported by 8.95% of the interviewed sample; for all investigated sustainable actions, the probabilities of a yes answer are very similar to the probabilities of the no answer and near to 50%. This cluster represents European citizens that declare to be prepared to change their habits but then are uncertain about performing the specific sustainability actions referred in the survey. The typical respondent in this group is female, works as employee, has an age between 15 and 24, single, lives in a large town, and before the pandemic used to travel several times a year. This cluster is

associated with Austria and Luxemburg. This result means that in these two countries, citizens are developing a positive attitude towards the topic of sustainability in tourism, but still have not transformed attitude in practice. Cluster 1, summarizes the largest number of different profiles, these profiles are all related to citizens with a good disposition to all circular economy practices investigated with the survey.

13.68% of European citizens are assigned to Cluster 2; they all declared that are not yet prepared to change their habits with reference to this behavior, they do not intend to perform any of the CE actions proposed in the questionnaire. The typical respondents in this cluster is male, age over 64, self-employed, single, living in a small or medium/sized town, never travelling before the pandemic. This cluster is associated to many European countries, mostly located in the Eastern area: Bulgaria, Cyprus, Denmark, Estonia, Finland, Croatia, Hungary, Poland, Slovenia. These citizens appear the most distant to sustainable tourism; this is the largest cluster in terms of citizens but it refers only to one specific profile.

The 6.39% of citizens belonging to Cluster 3 declared that they are prepared to change all their habits with reference to traveling and holidays in order to be more sustainable. The typical respondent is female, between 25 and 34, living in a large town, in a household with three components over 14, working as employee, travelling several times a year before Covid-19. This cluster is associated to Sweden. Cluster 3 refers to those European citizens that show the best attitude towards sustainability practices.

All other 16 clusters refer groups of respondents with a mild approach to this topic, mostly prepared to adopt one or more of the proposed circular economy actions.

Looking at Figure 2, some interesting evidences emerge with reference to some European countries, specifically Greece (P12), Malta (P20) and Romania (P24), who are located very distance from the other EU MSs. The result of Romania is associated to Cluster 3 (tourists very keen to sustainability practices) and this is a quite new result, since there are many evidences in the recent literature on the fact that in Eastern European countries CE practices are not adopted (Bassi and Dias, 2020). Malta and Greece have a similar position in the three-dimensional map; however, the smallest distance for Greece is with Cluster 18, for Malta with Cluster 12. Almost all citizens of Malta (99%) declare that they are willing to change their habits, however, the outlier position of this country in the map clearly shows that this positive attitude is not followed by an adherence to the specific sustainability actions proposed in the survey.

Cluster 13 and 14 that occupy outlier positions in Figure 2 are both small (1.82% and 5.62%). All citizens belonging to cluster 13 are prepared to choose transport options based on ecological impact, they are not prepared to perform the other eight sustainable actions. The typical citizen is between 45 and 54 years, living in a household with four or more members over 14 and in a small or medium/sized town, not occupied, never travelling before the pandemic. This cluster is associated with Denmark, but this is the profile that shows the largest distances with all EU countries. In Cluster 14, all citizens are prepared to pay more to protect the natural environment, pay more to the benefit of the local community, consume locally sourced products on holiday and reduce waste while on holiday; the majority of them (80% and over) are also prepared to perform all other five sustainable actions. The typical citizen is female, between 35 and 44 years, living in a household with two members over 14 and in a small or medium/sized town, working as employee, travelling many times in a year before the pandemic. This cluster is associated with Slovakia.

With the parameters estimated by the best fitting LCDA model and Euclidean squared distances, it is possible to calculate ratios and odds ratios. As an example, we calculate the log odds for a citizen of living in Austria rather than in Belgium given belonging to Cluster 1.

$$log\left(\frac{\mu_{1AU}}{\mu_{1BE}}\right) = log(\beta_{AU}) - log(\beta_{BE}) - d^2_{1AU} + d^2_{1BE} = log(0.1394) - log(0.1191) - 1.03 + 2.24 = 0.28$$

therefore, the odds ratio of being a citizen of Austria with respect to Belgium, given assignment to latent Cluster 1 is equal to 1.32. In general, odds ratios represent the relationship between the clusters (independent variables) and the European countries (multinomial categorical dependent variable). Calculating odds ratios gives a relative measure of the probability for citizens belonging to a certain cluster to be associated with a specific European country with reference to another country chosen as benchmark. With reference to the above odds ratios, we may see that it is less probable for a citizen, with a respondent profile classified in Cluster 1, to leave in Austria than in Belgium.

## 3. Conclusions

The latent class distance association model identified 19 clusters of profiles, corresponding to 19 groups

DRAFT

of European citizens. These clusters describe people with different levels of commitment towards the environment and specifically with different levels of preparedness to perform actions related to travel and tourism that could preserve natural resources. These groups vary from that of citizens who are prepared to change their habits with reference to all sustainable actions proposed in the survey to a group of citizens who do not wish to change at all. In between the model identified 17 other clusters that gather European citizens committed to one or more specific sustainable actions related to traveling and taking holidays.

The latent class distance association model estimated also the associations between each cluster and each one of the 27 EU countries. Citizens more committed to an environmentally friendly behavior live in Sweden, Austria, Luxemburg, Germany; citizens less willing to change their habits towards a more sustainable behavior live in Bulgaria, Cyprus, Denmark, Estonia, Croatia, Hungary, Latvia, Poland, Slovenia.

Citizens preparedness to change habits however depends also on their socio-demographic characteristics such as gender, age, occupation, type of community where living, household size and the frequency of travelling before the Covid-19 pandemic. Respondents who did not use to travel are, as it is quite obvious, less interested in the topic of sustainable tourism. Female and youngest respondents are keener to change to adopt a more sustainable behavior, as well as those who live in large towns. Youngest citizens, however, do not like those practices that increase the prices.

Implementation of sustainability in the touristic sector can, as it is obvious, not only preserve the natural environment, but, as well attract tourists and increase their satisfaction. It is strategic for touristic destinations and firms to know customers' attitude towards circular economy and specific sustainability actions.

## References


[1] Agresti, A.: Categorical data analysis. Wiley, New York (2013).

[2] Bassi, F.; Dias, J.G.: Sustainable development of the small and medium-sized enterprises in the European Union: a taxonomy of circular economy practices, Buss Strat Environ (2020) doi: 10.1002/bse.2518.

[3] Dempster, A. P., Laird, N. M., Rubin, D. B: Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B Stat Methodol, 39(1), 1–38. (1977).

[4] European Union: Flash Eurobarometer 499. Report. (2021).

[5] Vera, J.F., de Rooij, M., Heser, R.: A latent class distance association model for cross-classified data with a categorical response variable. Br J Math Stat (2014) doi: 10.1111/bmsp.12038.


DRAFT

# Sustainable tourism: a survey on the propensity towards eco-friendly accommodations

Claudia Furlan[a] and Giovanni Finocchiaro[b]

[a] Dept. of Statistical Sciences, University of Padova; claudia.furlan@unipd.it
[b] ISPRA - Italian Institute for Environment Protection and Research, 00144 Rome; giovanni.finocchiaro@isprambiente.it

**Abstract**

Considering that the environmental dimension of tourism sustainability is the primary lever to ensure healthy tourism for future generations in a welcoming area for both tourists and residents, this work is trying to probe awareness of the importance of sustainable tourism in all its dimensions, but above all a tourism that respects the environment.

In fact, this work presents the results of a survey on propensity towards environmentally friendly choices when planning the holidays, with a focus on bookings for sustainable accommodation. The people who consider the problem of limiting the environmental damages when planning a holiday are more frequently females, working in the tourism industry, interested in climate changes, persuaded that tourism makes damage, attracted by sustainable tourism and available to pay more for a sustainable tourism. It was found that, what turns a tourist into a *repeater* in the sustainable tourism is how much he is involved in adopting a sustainable lifestyle and the willing to pay more for a sustainable tourism.

**Keywords:** sustainable tourism, eco-friendly accommodation, customer satisfaction

DRAFT

## 1. Introduction

Sustainable Tourism takes account of environmentally sustainable practices used in and by the tourism industry. It is defined by the UN Environment Program and UN World Tourism Organization (UNWTO) as "tourism that takes full account of its current and future economic, social and environmental impacts, addressing the needs of visitors, the industry, the environment and host communities". Sustainable tourism also refers to "the environmental, economic, and socio-cultural aspects of tourism development, and a suitable balance must be established between these three dimensions to guarantee its long-term sustainability" [1].

Globally, tourism, although it has been explicitly linked to 3 specific Goals of the 2030 Agenda, namely Goals 8, 12 and 14, has the potential to contribute, directly or indirectly, to all the Sustainable Development Goals (SDGs). Indeed, the UNWTO (United Nations World Tourism Organization) has defined how each objective corresponds to a specific response from the tourism sector [2].

Although the environment represents one of the main attractions of the tourist destinations themselves, what often attracts tourists to travel and leave their usual environment is the motivation to see and experience different places and natural phenomena including countryside, beaches, mountains, islands, environmental assets, that become a fundamental part of the tourist offer of the various destinations. The importance of protecting the environmental dimension of tourism sustainability is not always perceived and then measured in a systematic way [3, 4].

This work was born within the Three-year Degree Course in Planning and Management of Cultural Tourism of the Padua University where through a series of thesis and with the constant technical-scientific comparison with ISPRA's researchers, an attempt is being made to sound out the awareness of the

importance of a tourism sustainable in all its dimensions but, above all, of environmentally friendly tourism. The investigation at the basis of this work, has tried to indirectly investigate the perception of environmental sustainability in the choice of tourism and, in particular, in the choice of tourist accommodation, and has also probed the propensity for eco-compatible behavioural approaches, trying to outline a profile of a typical tourist attracted by environmentally friendly tourism.

In fact, in this paper, we first depict the profile of people, based on the whole dataset ($n$=474), who are attracted by doing sustainable tourism and that consider the problem of limiting the environmental damages when planning a holiday. Then, by restricting the analysis to the $n$=66 individuals that had already used an eco-friendly accommodation in the past we find out which aspects a) contributes to the satisfaction of the experience and b) are relevant to transform tourists in *repeaters* in tourism sustainability.

The outline of the paper is as follows: the questionnaire is described in Section 2, the Results are shown in Section 3 and Conclusions follow at Section 4.

## 2. The Questionnaire

A questionnaire on eco-friendly accommodation propensity was posted on social networks (Instagram, Facebook, LinkedIn) between March and April 2022: 474 answers were in total collected, and 66 were given by people who already had chosen an eco-friendly accommodation in the past.

The questionnaire is divided in 4 sections, and we essentially report here the main questions which will be used in the analysis shown in Section 3. The first one deals with the general information of the individuals, such as age, gender, educational level, profession, and investigates if the individuals are studying (or have studied) on tourism-related disciplines, and if they are working (or have worked) in the tourism industry.

The second section regards the sustainability with a question on how much the individuals inquire themselves about the climate emergency and at which level they adopt a sustainable lifestyle. Both questions were recorded from 0 to 10, where 0 means "not at all" and 10 "completely".

The third section is dedicated to the sustainability in tourism. Among all, we report here 2 questions. The first is a statement to which the respondent must indicate the level of agreement between 1 and 5, where 1 indicates "strongly disagree" and 5 "strongly agree: "Nowadays in Italy there is an emergency for the damages that the tourism may cause". The second question is "How much are you attracted by a sustainable tourism? Give a value between 0 (not at all) to 10 (completely)".

The fourth section is dedicated to the personal choices that individuals make on sustainable tourism. Individuals were asked if they would be willing to pay more for a sustainable tourism (yes, no), and if they consider, at the moment of planning a holiday, the problem of limiting the environmental damages (yes, no). The individuals who replayed "yes" to the former question, were asked if limiting the environmental damages is a need a) that always have, and it was constant through lifetime b) that always have, and it has grown a lot in the last years, or c) that was born in the last years. To those who search information, during the holiday planning, whether an accommodation facility is environmentally friendly, were asked: "In your opinion, when an accommodation facility is environmentally friendly?" It was possible to choose one or more items among, for example, "photovoltaic panels, bio architecture, plastic free, organic/veggie/vegan/zero km cuisine, waste sorting". To those who had already used an eco-friendly accommodation in the past ($n$=66), where asked the type of the accommodation facility, through a multiple answer question (hotel, B&B, camping, farmhouse, etc), if they think they have paid more if compared with a non-sustainable accommodation (yes, no, I don't know), the satisfaction level of having used a sustainable accommodation (from 0 to 10, where 0 indicates "not at all" and 10 "completely satisfied"), and if they are going to book again in the future a sustainable accommodation (not at all, little, enough, much).

## 3. Results

The correlation between being attracted by a sustainable tourism and the degree of a sustainable adopted lifestyle is very highly correlated (0.96), based on the whole dataset (n=474): it means that the

DRAFT

attention towards the tourism sustainability grows in people that are already generally involved in environmental sustainability.

A logistic regression analysis was performed, on the whole dataset, for the question "at the moment of planning a holiday, do you consider the problem of limiting the environmental damages? (yes, no)". The results of the backward stepwise are shown in Table 1.

Table 1. Backward logistic regression for the question "at the moment of planning a holiday, do you consider the problem of limiting the environmental damages? (yes, no)". Percentage of correctly classified cases: 77.6.

| Parameters | Estimate | St. Error | p-value | OR |
|---|---|---|---|---|
| Age | 0.020 | 0.008 | 0.013 | 1.020 |
| Gender (f vs m) | 0.906 | 0.247 | <0.001 | 2.474 |
| Working in tourism industry (yes vs no) | 0.654 | 0.309 | 0.034 | 1.923 |
| Inquire about the climate emergency | 0.459 | 0.069 | <0.001 | 1.583 |
| damages that the tourism may cause (yes vs no) | 0.409 | 0.120 | <0.001 | 1.505 |
| Attraction by sustainable tourism | 0.362 | 0.109 | <0.001 | 1.436 |
| Willing to pay more (yes vs no) | 1.844 | 0.388 | <0.001 | 6.323 |
| Constant | -10.258 | 1.226 | <0.001 | 0.000 |

With respect to the demographical variables, the odds ratio shows that every unit increase in age is associated with a 2% increase in the odds of considering the problem of limiting damages, and the females have a 147.4% increase in the odds.

The personal experience of working (or have worked) in the tourism industry is associated with a 92.3% increase in the odds.

Considering the issue of sustainability, every unit increase in the level of how much the individuals inquire themselves about the climate emergency is associated with a 58.3% increase in the odds.

The next two variables belong to the section of the tourism sustainability: every unit increase in the level agreement with the statement "Nowadays in Italy there is an emergency for the damages that the tourism may cause" is associated with a 50.3% increase in the odds of considering the problem of limiting damages; moreover, every unit increase in the level of being attracted by sustainable tourism is associated with a 43.6% increase in the odds.

The last considered variable regards the economical aspect of the sustainable tourism, which is a very important issue beyond the personal belief: in fact, who wills to pay more for a sustainable tourism has a 532.5% increase in the odds.

The following two analysis were performed on the n=66 individuals that had already used an eco-friendly accommodation in the past.

The first regard the satisfaction level of having used a sustainable accommodation (Table 2), by a backward multiple regression. With respect to the demographical variables, the score of satisfaction decreases on average of 0.016 points for every unit increase in age, and it is larger of 0.798 points for females. The personal background of working (or have worked) in the tourism industry and of studying (or have studied) in tourism disciplines is associated with a decrease of about 0.72 points in the satisfaction.

Considering the issue of sustainable tourism, the satisfaction is positively related to the degree of being attracted by a sustainable tourism ($\hat{b}$ =0.767), and negatively related to the need of limiting the environmental damages if it was born in the last years ($\hat{b}$ =-0.743), compared to when if it was present through lifetime.

The economical topic is crucial in this analysis as well, since the score of satisfaction of those tourists who would like to pay more for a sustainable tourism is 2.376 points higher.

The last two variables are concerned to the accommodation. Hotels are the only accommodation facility which remains significant in the model and is positively associate to satisfaction. For those who express that an accommodation facility is considered environmentally friendly if there is an organic/veggie/vegan/zero km cuisine, the score of satisfaction decreases of 0.582 points. Perhaps their expectations were not satisfied.

DRAFT

Table 2. Backward multiple regression of the level of satisfaction. $R^2$=0.513.

| Parameters | Estimate | St. Error | p-value |
|---|---|---|---|
| Constant | 1.165 | 1.391 | 0.406 |
| Age | -0.016 | 0.007 | 0.039 |
| Gender (f vs m) | 0.798 | 0.258 | 0.003 |
| Working in tourism industry (yes vs no) | -0.728 | 0.293 | 0.016 |
| Studying in tourism disciplines (yes vs no) | -0.714 | 0.331 | 0.035 |
| Attraction by sustainable tourism | 0.767 | 0.126 | <0.001 |
| Limiting the environmental damages is a need: c) last years | -0.743 | 0.235 | 0.003 |
| Willing to pay more (yes vs no) | 2.376 | 0.482 | <0.001 |
| Eco-friendly: cuisine | -0.582 | 0.249 | 0.023 |
| Hotel | 0.491 | 0.228 | 0.036 |

The second analysis regards the will of booking again in the future a sustainable accommodation (Table 3), by a backward logistic regression. It is clear that, what really matters to turn tourists into *repeaters* in sustainable tourism is, first of all, that the tourists must already have adopted a sustainable lifestyle, and that the willing to pay more for a sustainable facility.

Table 3. Backward logistic regression of the will of booking again in the future a sustainable accommodation. Percentage of correctly classified cases: 84.6.

| Parameters | Estimate | St. Error | p-value | OR |
|---|---|---|---|---|
| Sustainability | 0.915 | 0.367 | 0.013 | 2.497 |
| Willing to pay more (yes vs no) | 5.494 | 1.641 | <0.001 | 243.265 |
| Constant | -10.395 | 3.802 | 0.006 | 0.000 |

DRAFT

## 4. Conclusions

In conclusion, the results presented here show that those individuals with a greater awareness of what sustainability means, with a concomitant sustainable lifestyle, are more inclined and even satisfied to choose "eco-friendly" hotels to spend their holidays even at the cost of paying more.

The greater awareness is also confirmed by the results that emerged with respect to posing the problem of making "tourist" choices that limit the damage to the environment. In fact, greater attention to these environmental aspects in tourist choices is guaranteed for "insiders" or for those who work in the tourism sector, for those who are informed about climate change, which represents one of the environmental problems with the greatest global impact and is increasingly evident in everyday life, for those who are aware that tourism, as a demographic pressure, causes damage to the surrounding environment. In terms of structural variables, this awareness is more pronounced among women and with increasing age.

## References

[1] UNEP & UNWTO, 2005: 11-12. Making Tourism More Sustainable – A Guide for Policy Makers
[2] tourism4sdgs.org
[3] ETC/ULS. Tourism and the environment Towards a reporting mechanism in Europe, Report, 2018. Available at https://www.eionet.europa.eu/etcs/etc-uls/products/etc-uls-report-01-2018-tourism-and-the-environment-towards-a-reporting-mechanism-in-europe
[4] European Parliament. Research for Tran Committee–From responsible best practices to sustainable tourism development, 2016. Available at www.europarl.europa.eu/RegData/etudes/STUD/2015/573421/IPOL_STU(2015)573421_EN.pdf

# A comparison of computational approaches for posterior inference in Bayesian Poisson regression

Laura D'Angelo[a]

[a]Department of Economics, Management and Statistics, University of Milano-Bicocca;
laura.dangelo@unimib.it

**Abstract**

Poisson regression is a well-established tool to study the effect of a set of covariates on a count variable of interest. While in the frequentist paradigm estimation of the coefficients has long been studied and refined, in the Bayesian framework it is customary to use general strategies, not explicitly targeted for this problem. Moving from the analysis of a dataset on the bicycle traffic on the Brooklyn Bridge in New York, where the variable of interest is the number of bikes crossing the bridge, we investigate the computational challenges of posterior inference for the Poisson log-linear model. We compare the performances of the well-known random-walk Metropolis-Hastings algorithm with those of the recently introduced algorithm of [2] implemented in the R package "bpr".

***Keywords:*** Count data; Log-linear model; Metropolis-Hastings

DRAFT

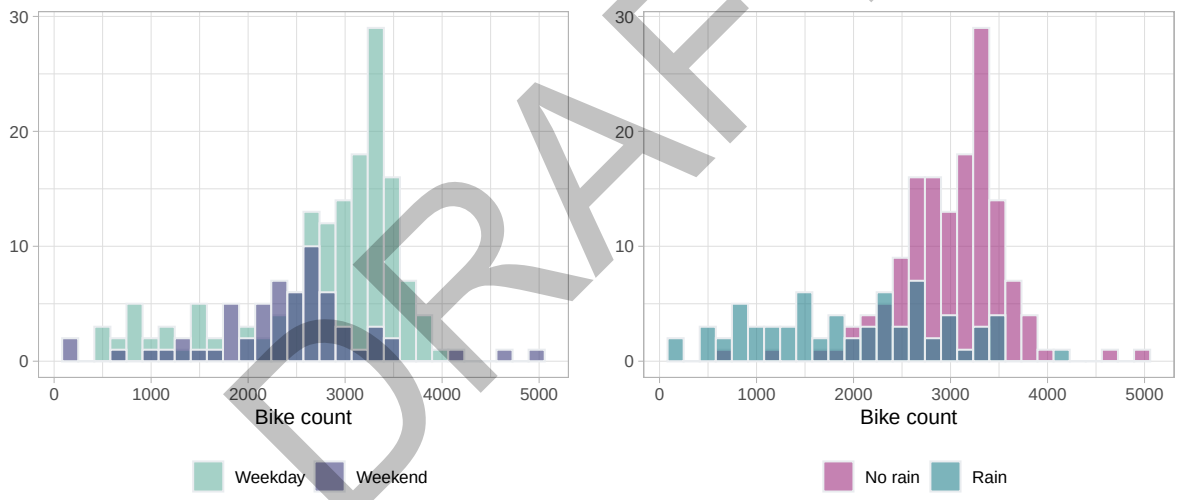
## 1. Introduction

Regression models stand at the core of the analysis of multivariate data sets and of the study of the relationship between several variables. The simplicity of their formulation, combined with the straightforward interpretation of the results, have contributed to their enduring success. When the variable of interest consists of counts, a typical choice is certainly Poisson regression, where the mean parameter of the distribution is linked with the explanatory variables via a logarithmic link function. In the frequentist framework, computation of the maximum likelihood estimates of the regression coefficients via iteratively reweighted least squares algorithm is a well-established procedure. In the Bayesian context, however, since the posterior distribution is not available in closed form, it is customary to approximate it using, for example, some type of Markov chain Monte Carlo method. For this reason, researchers have focused on improving computation for specific classes of models for count data: for example, count-valued time series [5] or hierarchical models [6].

In general, to perform Bayesian inference for Poisson log-linear models, there are mainly two viable options. The first is to rely on libraries such as Jags [3] or Stan [11], which are extremely flexible but require learning their particular programming language; the second is to implement a general algorithm as, for example, a Metropolis-Hastings (MH) [7] algorithm with a suitable proposal density (e.g., a random-walk, or an independent sampler). More recently, [2] developed a Metropolis-Hastings algorithm based on a proposal density which is an approximation of the posterior distribution, and hence leads to superior performances in terms of efficiency and mixing of the chains compared to general proposals. In their work, [2] compared the efficiency of their approach with that of the Hamiltonian Monte-Carlo algorithm [9] implemented using Stan, and they showed that their method held superior performances in scenarios with a moderate number of covariates.

Moving from the analysis of a real dataset of the number of bikes crossings a bridge in New York [10], we delve into the challenges of posterior inference in the context of Bayesian Poisson regression. The dataset we considered is part of a large public data repository provided by the city of New York aimed at studying the road traffic of different types of vehicles to improve transportation planning. In particular, we considered the data collected in 2017 on the Brooklyn Bridge. Each observation corresponds to the daily number of bicycles that crossed the bridge in the period that goes from April to October. Moreover, for each day it is also recorded the maximum and minimum temperature, and the rain intensity, when present. To improve the analysis, we also introduced two additional dummy variables: one that indicates if the day was a weekday or weekend, and a dichotomized version of the precipitation variable (hence indicating only the presence or absence of rain). The dataset hence contains 214 observations of 11 covariates, plus the number of crossings. The bikes count ranges between 151 and 4960; Figure 1 shows its distribution in the two groups identified by the new dummy variables "weekend" and "rain". The number of bicycles crossing the bridge varies greatly depending on the conditions, thus justifying the need to analyze the traffic more thoroughly to improve the management of the city mobility. Regarding the specification of the prior distributions, we selected independent Gaussian priors centered at zero for all the coefficients, since we did not have specific indications or prior knowledge.

The purpose of this paper is twofold: investigating the computational challenges of estimating the parameters of the regression model under consideration, and deriving inferential results on the specific dataset. In Sec. 2, we compare the efficiency of the algorithm of [2] as implemented in the R package `bpr` [1] with that of a MH algorithm with uniform random walk proposals; then, in Sec. 3, we illustrate how the output of the simulation can be used to perform inference on the dataset we considered.




DRAFT

Figure 1: Distribution of the bikes count during weekends/weekdays (left panel), and in days with rain/no rain (right panel).

## 2. Comparison of posterior sampling algorithms

The Metropolis-Hastings algorithm is a well-known Markov chain Monte Carlo method to simulate a sample from a generic target distribution that is known only up to a normalizing constant. The reasons for its success are clear, as it provides a general and extremely simple strategy to approximate any density, and it is especially useful when the number of dimensions is high and other approaches would be too complex or inaccurate. Briefly, the algorithm works as follows: a tentative sample is proposed from a known density, and the sample is accepted or rejected according to a probability that will allow the chains to have as stationary distribution the target density. In addition to the original formulation, other variants have been proposed to improve particular aspects, for example, the ease of finding a suitable proposal density and the consequent efficiency. For this application, as comparison with the algorithm

of [2], we used a modification of the MH that is more suited for high dimensional settings: in our case the parameter lives in a 12-dimensional space, hence finding an adequate proposal density can be tricky. In particular, we implemented the so-called "component-wise" random walk, where one variable at a time is updated, rather than performing a single full-dimensional update [8]. If, on the one hand, this approach allows defining distributions that ensure an adequate acceptance rate also in high-dimensions; on the other hand, it means selecting 12 tuning parameters (in this case, the "length" of the step), which can be burdensome. This is one of the advantages of the algorithm proposed by [2]: being based on an approximation of the posterior distribution of the coefficients of Poisson regression models, the only tuning parameter is the "distance" of the approximation from the target, hence requiring minimal tuning.

Each algorithm was run for 10000 iterations with 5000 of them discarded as burn-in. The convergence of the chains was assessed by graphical inspection of the trace plots; the tuning parameters were selected by running the algorithms several times on a grid of values to optimize the resulting acceptance rate and effective sample size. Both algorithms are written in C++ language exploiting the Rcpp package [4], and they were executed on a Linux machine running R 4.2.2.

As a metric to compare the results we used the time necessary to generate one independent sample (to take into account the autocorrelation of the chains). It can be computed as the total computation time over the effective sample size. Figure 2 shows the logarithm of the time (in seconds) per independent sample for the two algorithms, for each variable (the logarithmic transformation was used only for graphical purposes). Values larger than 0 means that more than 1 second was necessary to simulate one independent sample. The plot clearly shows the superior performances of the approach of [2], even if the random-walk MH allowed for a "variable-wise" tuning.

DRAFT

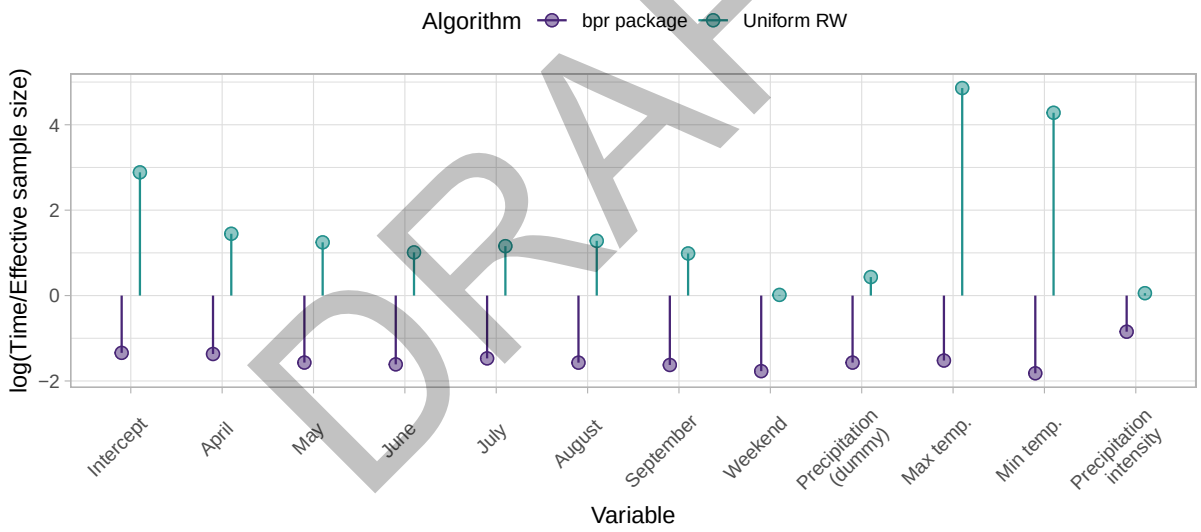



Figure 2: Logarithm of the time per independent sample for the two algorithms, for each variable.

## 3. Bikes count data analysis

To perform posterior inference on the considered dataset we used the output of the *bpr* R package since the chains have a larger effective sample (thus they convey more information). Table 1 summarizes the estimated coefficients: for each variable, it shows the posterior estimate of the median, mean and standard deviation, and the lower and upper bound of the 95% highest-posterior-density (HPD) interval. The only coefficient for which the credible interval contains 0 is the dummy variable associated with the month of May: hence the effect of this month is the same as the reference month, which is October (we introduced the intercept and removed the associated dummy). All other variables have instead a

Table 1: Summary of the estimated regression coefficients: median, mean and standard deviation; lower and upper bound of the 95% HPD interval.

| | Intercept | April | May | June | July | August | September |
|---|---|---|---|---|---|---|---|
| Median | 7.5365 | -0.1013 | -0.0011 | 0.0483 | 0.0146 | 0.1167 | 0.0414 |
| Mean | 7.5364 | -0.1010 | -0.0008 | 0.0480 | 0.0144 | 0.1168 | 0.0416 |
| Std. dev. | 0.0073 | 0.0052 | 0.0050 | 0.0052 | 0.0057 | 0.0056 | 0.0051 |
| HPD lower | 7.5226 | -0.1117 | -0.0093 | 0.0380 | 0.0037 | 0.1058 | 0.0321 |
| HPD upper | 7.5506 | -0.0918 | 0.0104 | 0.0573 | 0.0246 | 0.1279 | 0.0516 |

| | Weekend (dummy) | Precipitation (dummy) | Max temperature | Min temperature | Precipitation intensity |
|---|---|---|---|---|---|
| Median | -0.1165 | -0.1304 | 0.0399 | -0.0291 | -0.6441 |
| Mean | -0.1166 | -0.1304 | 0.0400 | -0.0291 | -0.6442 |
| Std. dev. | 0.0028 | 0.0035 | 0.0005 | 0.0007 | 0.0073 |
| HPD lower | -0.1219 | -0.1368 | 0.0389 | -0.0304 | -0.6564 |
| HPD upper | -0.1109 | -0.1235 | 0.0410 | -0.0278 | -0.6292 |

non-null effect on the mean number of bikes. Looking at the bottom table it is possible to notice how, as expected, the variables associated with a worsening of the atmospheric conditions (presence and intensity of rain, lowering temperature) have a negative effect, with the strongest being the presence of rain. More interesting is the coefficient of the day of the week: the traffic of bikes is more intense during weekdays, hence it is possible that many of the cyclists moving between Brooklyn and Manhattan use their bikes to go to work rather than for pleasure.

To ease the interpretation of the coefficients associated with the months, in Figure 3 it is shown the distribution of the parameters. From the plot it is immediate to see what months are associated with an increased number of bikes: as expected, spring and summer have a positive effect with respect to October. Surprisingly, April has a strong negative effect: however, if we further analyze the characteristics of that particular month, it is possible to see that it was much colder than all other months, even than October. Hence the coefficient of the temperature could not fully capture the exceptional cold of that period compared to the rest of spring and summer.

Although the dataset we considered is quite simple, with only a few recorded meteorological variables, the results are still interesting in showing a clear pattern of the bike traffic moving between Brooklyn and Manhattan. This kind of analysis could turn extremely useful for city planning to improve mobility and safety of the people, for example, by reducing or increasing the space reserved for bicycles according to the period of the year.

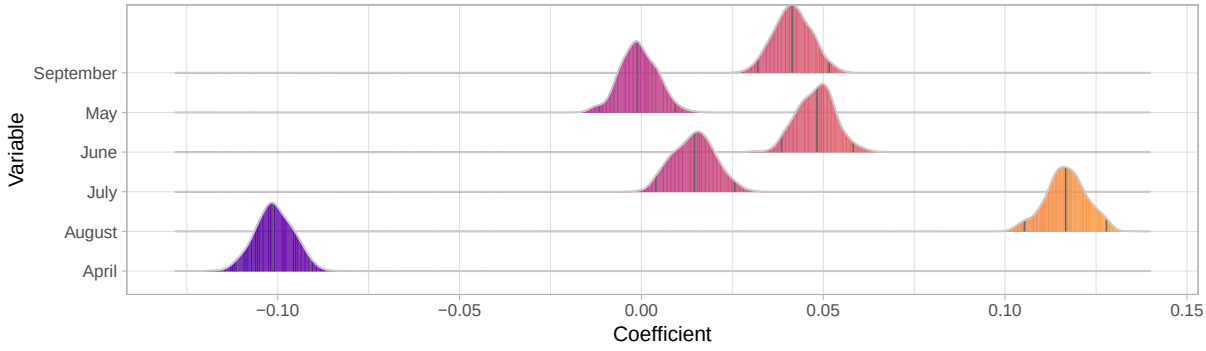



Figure 3: Posterior distribution of the coefficients associated with the month variables. The vertical dark lines corresponds to the lower bound of the HPD interval, median, and upper bound of the HPD interval.

DRAFT

## References


[1] D'Angelo, L. *bpr: Bayesian Poisson regression*. URL: https://CRAN.R-project.org/ package=bpr. 2021.

[2] D'Angelo, L. and Canale, A. "Efficient Posterior Sampling for Bayesian Poisson Regression". *J. Comput. Graph. Stat.* (2022), 1–10. DOI: 10.1080/10618600.2022.2123337.

[3] Denwood, M. J. "runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS". *J. Stat. Softw.* 71(9) (2016), 1–25.

[4] Eddelbuettel, D. and Francois, R. "Rcpp: Seamless R and C++ Integration". *J. Stat. Softw.* 40(8) (2011), 1–18. ISSN: 1548-7660.

[5] Frühwirth-Schnatter, S. and Wagner, H. "Auxiliary Mixture Sampling for Parameter-Driven Models of Time Series of Counts with Applications to State Space Modelling". *Biometrika* 93(4) (2006), 827–841.

[6] Frühwirth-Schnatter, S. et al. "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data". *Stat. Comput.* 19(479) (2009).

[7] Hastings, W. K. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika* 57(1) (1970), 97–109.

[8] Johnson, A. A., Jones, G. L., and Neath, R. C. "Component-wise Markov Chain Monte Carlo: uniform and geometric ergodicity under mixing and composition". *Stat. Sci.* 28(3) (2013), 360–375.

[9] Neal, R. M. "MCMC using Hamiltonian dynamics". *Handbook of Markov chain Monte Carlo* 2(11) (2011), 2.

[10] NYC Open Data. *Department of Transportation (DOT) - Dataset: Bicycle Counts for East River Bridges (Historical)*. URL: https://opendata.cityofnewyork.us/. 2022.

[11] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.21.8. 2023. URL: https://mc-stan.org/.


DRAFT

# Bias-reduction methods for Poisson regression models

Luca Presicce[a], Tommaso Rigon[a], and Emanuele Aliverti[b]

[a]Department of Economics, Management and Statistics, University of Milano–Bicocca, Milano 20126, Italy; l.presicce@campus.unimib.it, tommaso.rigon@unimib.it
[b]Department of Statistics, University of Padova, Padova 35121, Italy; emanuele.aliverti@unipd.it

**Abstract**

Poisson regression models for count responses are routinely used in statistical practice. However, some difficulties may arise when there is limited information within the data: the maximum likelihood estimate may not exist or it could be highly biased. We address this issue by considering a suitable penalty for the likelihood function. Such a penalization term has a clear Bayesian interpretation and important connections with the reduced-bias approach of Firth (1993) [5]. We show that the associated penalized estimates may be obtained by maximizing a genuine likelihood in which pseudo-counts are used in place of the original data. Compared to the method of [5], our approach speeds up computations by a significant margin, as it leverages well-established algorithms for Poisson models. This is achieved while maintaining comparable inferential performance, as illustrated through an application to real data.

***Keywords:*** Bias-reduction, Conjugate priors, Penalized generalized linear models, Poisson regression.

DRAFT

## 1. Introduction

Generalized linear models [18] (GLM) are classical tools for the modeling of the relationship between a response variable and a set of predictors. A canonical overview is the book by [16]. Despite their popularity, GLMs have some well-known limitations. One aspect of potential concern is the presence of bias in the maximum likelihood estimates, which, beyond the Gaussian case, may impact the inferential conclusions. To address this issue, a variety of methods have been proposed in the literature, including the bias-reduction technique of [5] and subsequent developments; see for example [12; 13; 14; 15; 10; 9; 2]. Unfortunately, albeit theoretically appealing, most of these methods may face computational bottlenecks.

In a recent paper, [19] presents a fast alternative to Firth's method in the logistic regression case, which is based on a penalized likelihood function. The key idea relies on the notion of conjugate priors for exponential family models, due to Diaconis and Ylvisaker (1979) [4] and well experienced by several authors; see, for example, [1; 7; 8]. Indeed, [19] penalizes the likelihood by the conjugate prior defined in [4; 1], showing that, under suitably specified parameters, the results can be seen as an approximation of Firth's method [5]. Such an approach is a milestone in the bias-reduction literature, being routinely employed in applied studies thanks to various software implementations [e.g., 11].

Actually, most works on bias reduction for GLM regard Binomial responses. Indeed, binary regression is widely used in several fields but is also more vulnerable to issues, such as data *separability*,

which leads to infinite MLE. Even if this issue is mostly recurrent in logistic regression, it also occurs in other models, such as the Poisson regression. Hence, we will borrow ideas from [19] to present a bias-reduction method for Poisson regression models that brings computational advantages. Similar to this approach, we penalize the likelihood with the Diaconis-Ylvisaker (DY) conjugate prior, obtaining a convenient formulation. The resulting posterior law is a genuine likelihood function in which *pseudo-counts* are employed rather than the original response variables. In particular, each pseudo-count is a linear combination of the original data and the parameters of the DY prior.

The demand for bias-reduction methodologies in Poisson models descends from their broad range of applications. Indeed, the Poisson regression is an essential tool in applied research, see for further details [6; 17]. Moreover, the growing complexity of the contexts in which these models are employed justifies the necessity of faster alternatives.

In this manuscript, we will focus on count response data. More precisely, we assume that the response variables $Y_1, \ldots, Y_n$ are such that

$$Y_i \overset{\text{ind}}{\sim} \text{Pois}(\lambda_i), \qquad g(\lambda_i) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tag{1}$$

independently for $i = 1, \ldots, n$. The vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ contains the covariate information of the $i$th observation, whereas the $p$-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ corresponds to the regression coefficients. The function $g(\cdot)$ is the *link function*, which in this paper is assumed to be $g(\cdot) = \log(\cdot)$, that is, the so-called canonical link for Poisson models [16].

## 2. Diaconis and Ylvisaker conjugate priors

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ be realizations of independent Poisson random variables with means $\lambda_1, \ldots, \lambda_n$, as in (1). Thus, the likelihood function assumes the following form

$$\mathscr{L}(\boldsymbol{\beta}; \boldsymbol{y}) \propto \exp\left[\sum_{i=1}^{n} \{y_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}\right]. \tag{2}$$

Following [1; 19], we consider the DY conjugate prior for the regression coefficients $\boldsymbol{\beta}$, with density proportional to

$$\pi(\boldsymbol{\beta}) \propto \exp\left[\tau \sum_{i=1}^{n} \{\psi_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\}\right], \tag{3}$$

where $\tau > 0$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)^\intercal$ are prior hyperparameters such that $\psi_i > 0$, for $i = 1, \ldots, n$. The hyperparameters $\tau$ and $\boldsymbol{\psi}$ can be interpreted as scale and location parameters, respectively. Specifically, $\tau$ controls the variability and the strength of the prior belief on $\boldsymbol{\psi}$, with larger values for $\tau > 0$ representing strong prior information and smaller $\tau$ less confidence in $\pi(\boldsymbol{\beta})$. Instead, $\boldsymbol{\psi}$ is the mode of the prior distribution, thus defining a location parameter for $\pi(\boldsymbol{\beta})$.

The posterior distribution can be obtained combining prior information (3) with the likelihood function (2) as $\pi(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \mathscr{L}(\boldsymbol{\beta}; \boldsymbol{y})$, leading to

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \exp\left[(\tau + 1)\left\{\sum_{i=i}^{n} \left(\frac{\tau\psi_i + y_i}{\tau + 1}\right) \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})\right\}\right]. \tag{4}$$

In equation (4), it is easy to see that the posterior is in the same family of the prior, leading to a posterior distribution which is again in the DY family with updated parameters. In addition, the results in [1] guarantee that (4) has a finite normalizing constant, and therefore can be regarded as a proper posterior distribution. The prior parameters $\boldsymbol{\psi}$ are updated, obtaining the pseudo-counts $\tilde{y}_i = (y_i + \tau\psi_i)/(\tau + 1)$ for $i = 1, \ldots, n$. We will later tune the prior hyperparameters for bias reduction purposes. Similarly to [19] the penalized likelihood is equivalent to the original one minus a constant term depending only on the hyperparameter $\tau$, that is

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \exp\{(\tau + 1)\ell(\boldsymbol{\beta}; \tilde{\boldsymbol{y}})\}, \tag{5}$$

DRAFT

where $\ell(\boldsymbol{\beta}; \tilde{\boldsymbol{y}})$ is the log-likelihood and $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^\intercal$. From (5), it follows that the maximum a posteriori (MAP) of $\pi(\boldsymbol{\beta} \mid \boldsymbol{y})$ corresponds to the MLE obtained from $\ell(\boldsymbol{\beta}; \tilde{\boldsymbol{y}})$. This result allows us to rely on standard algorithms for Poisson regression, obtaining important computational benefits.

## 3. Penalized score equations

We now introduce the score function, defined as the vector of derivatives $U(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$. Maximum likelihood estimates are obtained as the solution of the system of *score equations* $U_r(\boldsymbol{\beta}) = 0$, for $r = 1, \ldots, p$. A penalization for the likelihood leads to penalized score equations, so that

$$U_r(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu_i) x_{ir} \ , \quad U_{r,\text{DY}}(\boldsymbol{\beta}) = C(\tau) \sum_{i=1}^{n} (\tilde{y}_i - \mu_i) x_{ir} \ , \quad U_{r,\text{FI}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i^* - \mu_i) x_{ir}, \quad (6)$$

with $r = 1, \ldots, p$ and $\mu_i = \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})$, where $U_r(\boldsymbol{\beta})$ is the score equation for the original model, $U_{r,\text{DY}}(\boldsymbol{\beta})$ relies on the score equation of our penalized model with $C(\tau) = (\tau + 1)$, and $U_{r,\text{FI}}(\boldsymbol{\beta})$ is the score equations obtained with [5]. In particular, Firth's pseudo-counts correspond to $y_i^* = y_i + h_{ii}/2$ for $i = 1, \ldots, n$, where $h_{11}, \ldots, h_{nn}$ represent the diagonal elements of the *leverage matrix* $H(\boldsymbol{\beta}) = W(\boldsymbol{\beta})^{1/2} X (X^\intercal W(\boldsymbol{\beta}) X)^{-1} X^\intercal W(\boldsymbol{\beta})^{1/2}$, with $W(\boldsymbol{\beta}) = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

The score equations in (6) are similar and involve different pseudo-observations. Our approach and the one by [5] provide a penalized system of score equations depending on a different set of pseudo-counts, respectively defined as

$$y_i^* = y_i + \frac{h_{ii}}{2} \ , \quad \tilde{y}_i = \frac{y_i + \tau \psi_i}{\tau + 1} \ , \quad i = 1, \ldots, n. \quad (7)$$

Firth's pseudo-counts $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^\intercal$ depend on the diagonal elements of $H(\boldsymbol{\beta})$, while $\tilde{\boldsymbol{y}}$ depends only on the prior hyperparameters $\tau$ and $\boldsymbol{\psi}$. We search for the values $\tau$ and $\boldsymbol{\psi}$ that introduce as much similarity as possible between $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}^*$. Indeed, it is well-known that [5] removes the first-order term of the asymptotic bias.

Henceforth, we exploit the concept of *quadratic balance approximation*, discussed in [3]. This is a computational expedient that avoids the computational drawbacks due to the computation of $H(\boldsymbol{\beta})$ at each iteration of the optimization procedure. It can be shown that $H(\boldsymbol{\beta})$ is a projection matrix with a suitable algebraic property: if the design matrix has full rank, the trace of $H(\boldsymbol{\beta})$ is equal to its rank, which is $p$. Thus, one could approximate each $h_{ii}$ for $i = 1, \ldots, n$ with their mean, obtained as the ratio between the trace of $H(\boldsymbol{\beta})$ and the dimension of its diagonal $n$, that is $p/n$. We call this procedure *mean approximation*, in order to discern it from the concept discussed in [3].

The mean approximation is the tool that allows us to reach the similarity between the pseudo-counts. We set the hyperparameters that define $\tilde{\boldsymbol{y}}$ with the purpose of emulating the pseudo-observations $\boldsymbol{y}^*$. In particular, we can substitute the terms $h_{ii}$ with $p/n$. To mimic this idea under the DY penalty, we allow $\tau = \tau(\alpha)$ and each $\psi_i = \psi_i(\alpha)$ to depend on a fixed constant $\alpha > 0$, chosen so that $\tau(\alpha)\psi_i(\alpha) = p/(2n)$, obtaining

$$\tilde{y}_i(\alpha) = \frac{y_i + \tau(\alpha)\psi_i(\alpha)}{\tau(\alpha) + 1} = \frac{y_i + p/(2n)}{\tau(\alpha) + 1}, \qquad i = 1, \ldots, n. \quad (8)$$

For example, one could set $\tau(\alpha) = p/(\alpha n)$ and $\psi_i(\alpha) = \alpha/2$ for $i = 1, \ldots, n$. The pseudo-counts in (8) do not correspond exactly with the mean approximation of Firth's pseudo-counts, which would be $y_i + p/(2n)$. However, note that the mean approximation of the pseudo-counts can be seen as the result of an improper prior, occurring when the precision $\tau(\alpha) \to 0$. The pseudo-counts $\tilde{\boldsymbol{y}}$ defined in (8) have two important implications: as desired, we recover a quantity that resembles the mean approximation of $\boldsymbol{y}^*$. Secondly, we have an increasing coincidence between $U_{\text{DY}}(\boldsymbol{\beta})$ and $U_{\text{FI}}(\boldsymbol{\beta})$, as the following study suggests.

DRAFT

## 4. Illustration

We follow the structure of the *birthweight* example presented in [15] and [19], considering the *infert* data. The data regard $n = 248$ women in a case-control study, in which the number of spontaneous aborts is considered as a count outcome. The models consider a regression of the response variable on an intercept and six covariates, leading to $p = 7$. In this example, maximum likelihood estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_7)^\intercal$ of the regression coefficients $\boldsymbol{\beta}$ exists finite. In this illustration, we simulate 10000 datasets from a Poisson regression model with parameter $\hat{\boldsymbol{\beta}}$ obtained as MLE from the original data. This allows us to evaluate estimators' behavior for all the models presented in Section 3. Moreover, we compare their performances in terms of bias and root mean squared error (RMSE). Results are presented in Table 1.

| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ |
|---|---|---|---|---|---|---|---|---|
| BIAS | $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ | -0.20 | 0.21 | 0.21 | 0 | 0.00 | -0.01 | 0.00 |
| | $\hat{\boldsymbol{\beta}}_{\text{DY}}$ | -0.09 | 0.09 | 0.09 | 0 | -0.01 | 0.02 | -0.02 |
| | $\hat{\boldsymbol{\beta}}_{\text{FI}}$ | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 |
| RMSE | $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ | 1.57 | 1.49 | 1.49 | 0.02 | 0.07 | 0.16 | 0.17 |
| | $\hat{\boldsymbol{\beta}}_{\text{DY}}$ | 0.78 | 0.60 | 0.60 | 0.02 | 0.07 | 0.15 | 0.17 |
| | $\hat{\boldsymbol{\beta}}_{\text{FI}}$ | 0.73 | 0.53 | 0.53 | 0.02 | 0.07 | 0.16 | 0.17 |

Table 1: Simulation study results on the regression coefficients estimates obtained for spontaneous abort counts problem from the three models considered.

The estimator $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ performs significantly worse than all the reduced-bias methodologies in terms of bias and RMSE. Moreover, we observe a striking empirical similarity in terms of bias between the performance of the proposed penalized estimator $\hat{\boldsymbol{\beta}}_{\text{DY}}$ and the one of [5]. This finding is in line with the considerations discussed in Section 2. As a matter of fact, the proposed penalized estimator $\hat{\boldsymbol{\beta}}_{\text{DY}}$ has only a slightly worse RMSE compared to $\hat{\boldsymbol{\beta}}_{\text{FI}}$.

In addition, we investigate the computational timing of the estimation process for both our and [5] model, relying on a native R implementation and considering 100 replications.

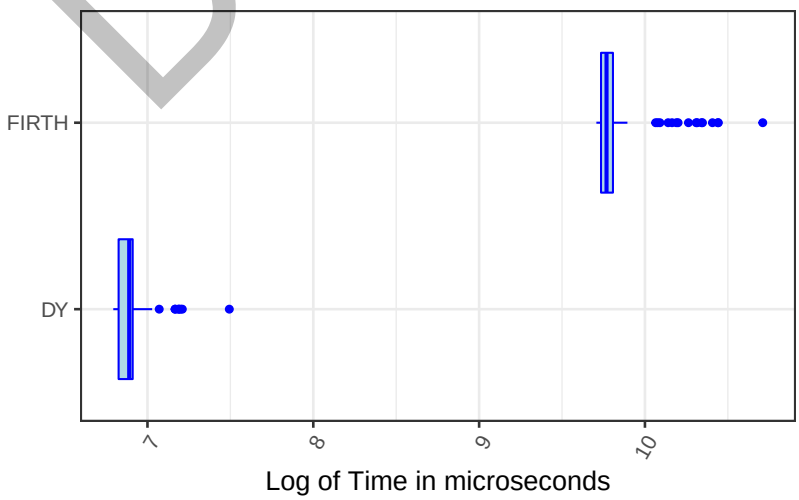



Figure 1: Timing boxplot of the estimation processes for the DY penalized and the Firth's models.

Results are presented in Figure 1, where it is evident that the proposed approach outperforms the competitor. In particular, our model estimation process takes approximately 19 times less than the com-

DRAFT

pared one. This is a further indication that the proposed contribution avoids the computational drawbacks of the milestones in bias-reduction techniques, while retaining comparable performance.

## References


[1] Chen, M.H., Ibrahim, J.G.: Conjugate priors for generalized linear models. Statist. Sinica **13**(2), 461–476 (2003)

[2] Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., Weidman, L.: Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression. J. Am. Statist. Assoc. **86**(413), 68–78 (1991)

[3] Cordeiro, G.M., McCullagh, P.: Bias Correction in Generalized Linear Models. J. Roy. Statist. Soc. Ser. B **53**(3), 629–643 (1991)

[4] Diaconis, P., Ylvisaker, D.: Conjugate Priors for Exponential Families. Ann. Statist. **7**(2), 269 – 281 (1979)

[5] Firth, D.: Bias Reduction of Maximum Likelihood Estimates. Biometrika **80**(1), 27–38 (1993)

[6] Frome, E.L.: The Analysis of Rates Using Poisson Regression Models. Biometrics **39**(3), 665–674 (1983)

[7] Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S.: A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Statist. **2**(4), 1360 – 1383 (2008)

[8] Greenland, S.: Generalized Conjugate Priors for Bayesian Analysis of Risk and Survival Regressions. Biometrics **59**(1), 92–99 (2003)

[9] Greenland, S., Mansournia, M.A.: Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Statist. Med. **34**(23), 3133–3143 (2015)

[10] Kenne Pagui, E.C., Salvan, A., Sartori, N.: Median bias reduction of maximum likelihood estimates. Biometrika **104**(4), 923–938 (2017)

[11] Kosmidis, I.: brglm2: Bias reduction in generalized linear models (2017). R package version 0.1.5

[12] Kosmidis, I., Firth, D.: Bias reduction in exponential family nonlinear models. Biometrika **96**(4), 793–804 (2009)

[13] Kosmidis, I., Firth, D.: A generic algorithm for reducing bias in parametric estimation. Electron. J. Stat. **4**, 1097 – 1112 (2010)

[14] Kosmidis, I., Firth, D.: Multinomial logit bias reduction via the Poisson log-linear model. Biometrika **98**(3), 755–759 (2011)

[15] Kosmidis, I., Pagui, E.C.K., Sartori, N.: Mean and median bias reduction in generalized linear models. Stat. Comput. **30**, 43–59 (2020)

[16] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall / CRC, London (1989)

[17] Muñoz-Pichardo, J.M., Pino-Mejías, R., García-Heras, J., Ruiz-Muñoz, F., González-Regalado, M.L.: A multivariate Poisson regression model for count data. J. Appl. Stat. **48**(13-15), 2525–2541 (2021)

[18] Nelder, J.A., Wedderburn, R.W.M.: Generalized Linear Models. J. Roy. Statist. Soc. Ser. A **135**(3), 370–384 (1972)

[19] Rigon, T., Aliverti, E.: Conjugate priors and bias reduction for logistic regression models. arXiv (2022)


DRAFT

# Finite Mixture Model for Multiple Sample Data

Alessandro Colombi[a], Raffaele Argiento[b], Federico Camerlenghi[a], and Lucia Paci[c]

[a] University of Milan-Bicocca; a.colombi10@campus.unimib.it, federico.camerlenghi@unimib.it

[b] University of Bergamo; raffaele.argiento@unibg.it

[c] Università Cattolica del Sacro Cuore; lucia.paci@unicatt.it

**Abstract**

In this paper, we propose a Bayesian nonparametric level-dependent mixture model for clustering. To achieve this, we employ a vector of species sampling models with shared atoms and level-specific weights. This results in multiple random probability measures with a common support, which we use to perform both inter-level and within-level clustering of the data. This approach enables us to take into account both heterogeneity and common patterns shared across levels in our clustering analysis. Specifically, we study the properties of the group-dependent clustering structure induced by our hierarchical mixture model. We develop both a marginal and a conditional Gibbs sampler to perform Bayesian inference. We evaluate the model's ability to recover the original clustering of the data and assess its goodness of fit through simulated data.

***Keywords:*** Bayesian analysis, clustering, partial exchangeability, density estimation

DRAFT

## 1. Introduction

In several statistical settings there is the need to model multilevel data, allowing for sharing of information across the levels. In the Bayesian framework, this is achieved by hierarchical modeling, where the joint distribution of level-specific parameters accounts for such dependence. For instance, in Bayesian nonparametrics, the seminal work by (11) considered a mixture model within each level, say $j$, where the level-specific parameter is the mixing measure $P_j$ whose joint law is defined by an extra layer of hierarchy, yielding to the hierarchical Dirichlet process. This approach has been extended to the class of Normalized Random Measures with Independent Increments (NRMI) (10) by (6) and (1) and to species sampling models by (3).

In this work, we propose a hierarchical model where the level-specific mixing distribution belongs to the class of almost surely finite dimensional distributions introduced by (2). The following are defined such that they have level-specific weights but shared atoms, i.e they share the same support. The idea of using a common support for the mixing measures reminds the recent works by (4), (8) and (7) with the difference that both previous works build the dependent prior following a nested approach while we focus on the hierarchical construction. By following this approach, we attain both a level-specific and a global clustering of the observations. Specifically, for each individual $i$ and each level $j$, we independently sample a latent parameter $\boldsymbol{\theta}_{ji}$ from $P_j$, which is almost surely discrete. This results in ties within each level, thus producing a level-specific clustering. Furthermore, since the $P_j$'s share the same support, we expect also ties between levels, providing a global clustering.

The paper proceeds as follow, in Section 2. we define the vector of species sampling models and derive the proposed model. Then in Section 3. we shade light on the clustering mechanism induced by the model and in Section 4. we provide a simulation study to assess the ability of recovering the true partitioning of the data as well as to perform density estimation.

## 2. Model

Consider $d$ levels of observations, $\boldsymbol{y}_j = \left(y_{j1}, \ldots, y_{jn_j}\right)$ for $j = 1, \ldots, d$ that we assume to be partially exchangeable. This means that the level membership introduces heterogeneity across subgroups of the population. Traditional modeling approaches such as pooling all observations or performing $d$ level independent analysis fall short in capturing both the differences across levels or the relationships between the observations, which relate to the same phenomenon and therefore may share common patters. To balance these opposing situations, we aim to build a statistical model that accounts for heterogeneity while also allowing for borrowing of strength across units.

For example, referring to Figure 1, we model students within each school to be exchangeable but we do not allow for exchangeability between schools acknowledging that their performances may be influenced by their school membership. At the same time, it would also be advantageous to share information across schools.

Let $y_{ji}$ be the observed variable for level $j$, and individual $i$, $i = 1, \ldots, n_j$. We assume that the data in each level $j$ come from a finite mixture of $M$ components, that is

$$y_{j1}, \ldots, y_{jn_j} \overset{\text{iid}}{\sim} \int_{\Theta} f(y_{ji}|\boldsymbol{\theta}_j) P_j(d\boldsymbol{\theta}_j), \quad \text{for each } j = 1, \ldots, d, \tag{1}$$

where $f(y_{ji} \mid \boldsymbol{\theta}_j)$ is the parametric density kernel over the sampling space, $\boldsymbol{\theta}_j \in \Theta \subset \mathbb{R}^l$ for each $j = 1, \ldots, d$ and $P_1, \ldots, P_d$ are the mixing measures. In the present work, we focus on finite dependent random probability measures having support $\Theta$ defined as follows,

$$P_j(\cdot) = \sum_{m=1}^{M} \frac{S_{jm}}{T_j} \delta_{\tau_m}(\cdot), \tag{2}$$

having denoted by $\delta_{\tau_m}$ the delta-Dirac mass at $\tau_m$, and $T_j = \sum_{m=1}^{M} S_{jm}$ is the total mass. Here $(\tau_1, \ldots, \tau_M)$ are common random atoms across the $d$ random probability measures, which are assumed independent and identically distributed with common distribution $P_0$ i.e, a diffuse probability measure on $\mathbb{R}^l$. The unnormalized weights $(S_{j1}, \ldots, S_{jM})$ are i.i.d. from gamma$(\gamma_j, 1)$, where $\gamma_j$ is a parameter specific for level $j \in \{1, \ldots, d\}$. The induced prior on the normalized mixture weights $(\pi_{j1}, \ldots, \pi_{jM})$ is

$$(\pi_{j1}, \ldots, \pi_{jM}) = \left(\frac{S_{j1}}{T_j}, \ldots, \frac{S_{jM}}{T_j}\right) \sim \text{Dir}_M(\gamma_j, \ldots, \gamma_j), \quad \text{for each } j = 1 \ldots, d,$$

where $\text{Dir}_M(\gamma_j, \ldots, \gamma_j)$ denotes the $M$-dimensional symmetric Dirichlet distribution with parameter $\gamma_j$. Finally, $M$ is supposed to be a positive integer-valued random variable with distribution $q_M(\cdot)$ on $\{1, 2, 3, \ldots\}$.

Summing up, the model can be formulated in the following hierarchical form,

$$\begin{aligned}
y_{j1}, \ldots, y_{jn_j} \mid \boldsymbol{w}, \boldsymbol{\tau}, M &\overset{\text{iid}}{\sim} \sum_{m=1}^{M} w_{jm} f(y_{ji} \mid \boldsymbol{\tau}_m), && \text{for } j = 1, \ldots, d \\
\tau_1, \ldots, \tau_M \mid M &\overset{\text{iid}}{\sim} P_0(d\theta) && \\
S_{j1}, \ldots, S_{jM} \mid \gamma_j, M &\overset{\text{iid}}{\sim} \text{gamma}\left(\gamma_j, 1\right), && \text{for } j = 1, \ldots, d \\
\gamma_1, \ldots, \gamma_M \mid M &\overset{\text{iid}}{\sim} \text{gamma}\left(a_1, b_1\right) && \\
M &\sim q_M(M). &&
\end{aligned} \tag{3}$$

DRAFT

In this work we focus on the particular case where the kernel $f(y \mid \boldsymbol{\theta})$ represents the density of a univariate normal distribution with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)$, hence $l = 2$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$. We choose $P_0(\boldsymbol{\theta})$ as the density of a conjugate normal inverse gamma prior with parameters $\mu_0$, $\kappa_0$, $\nu_0$ and $\sigma_0^2$. Finally, we set $q_M(\cdot)$ to be the p.m.f. of a 1−shifted Poisson distribution with parameter $\Lambda$. To allow further flexibility, we also place a gamma$(a_2, b_2)$ prior on $\Lambda$.

## 3. Model Based Clustering

The hierarchical model in (3) allows to define a level-dependent clustering based on the latent variables $\boldsymbol{\theta}_{ji}$. First, we introduce latent allocation variables $c_{ji}$ such that $c_{ji} = m$ if $\boldsymbol{\theta}_{ji} = \tau_m$. Then, we denote $\mathscr{M}^{(a)}$ the set of couples $(j, m)$ such that exists an index $i$ for which $c_{ij} = m$ and we define the number of allocated components as

$$K = \left| \left\{ m \in \{1, \dots, M\} \ : \ \text{there exists one couple} (j, m) \in \mathscr{M}^{(a)}, j = 1, \dots, d \right\} \right|,$$

where the $|\cdot|$ operator indicates the cardinality of the set. We denote $\mathscr{M}^{(na)}$ the complement of $\mathscr{M}^{(a)}$. Hence, for every pair $(j, m)$, we define $n_{jm} = |\{(j, i) : \ c_{ji} = m\}|$. Note that

$$(j, m) \in \mathscr{M}^{(na)} \Rightarrow n_{jm} = 0$$

$$(j, m) \in \mathscr{M}^{(a)} \Rightarrow n_{jm} \geq 0.$$

Finally, let $c_1^*, \dots, c_K^*$ be the allocated columns, that is, the indexes within $\{1, \dots, M\}$ such that $(j, c_k^*) \in \mathscr{M}^{(a)}$. For each level $j$, we define the partition $\rho_j = \{C_{j1}, \dots, C_{jK}\}$, where $C_{jk} = \{(j, i) : \ (j, c_{ki}^*) \in \mathscr{M}^{(a)}\}$ and $k = 1, ..., K$. In other words, $C_{jk}$ is the set of data points of level $j$ belonging to the $k$-th cluster. By construction, $n_{jk} = |C_{jk}|$. A distinctive feature of our setting, is that some $C_{jk}$ can be an empty sets. Nevertheless, if $C_{jk} = \emptyset$ appears in $\rho_j$, it means that there is at least another group $\tilde{j}$ such that $C_{\tilde{j}k}$ is not empty. In other words, $n_{jk} \geq 0$ but they must satisfy the following constrains

$$\sum_{j=1}^{d} n_{jk} > 0, \text{ and } \sum_{k=1}^{K} n_{jk} = n_j,$$

for $k = 1, ..., K$ and $j = 1, \dots, d$ respectively.

Figure 1 provides a visual representation of the clustering. In this example, we measure data for each student in each of the $d = 3$ schools (levels). To account for heterogeneity between schools, students are assumed to be partially exchangeable, i.e they are exchangeable within each school but not across different schools. Each color represent a cluster, hence $K = 2$. The same color in different schools shows how clusters with the same interpretation can be found in different schools.

## 4. Simulation Study

In this section we aim to assess the performance of our model in terms of ability of recovering the original clustering of the observations as well as to evaluate its goodness of fit. In particular, to evaluate the clustering we monitor the estimated number of clusters as well as the popular Rand Index (RI). Additionally, we perform density estimation and visually inspect the results to gain further insights, see Figure 2.

We generated data from $d = 3$ distinct levels of 200, 100, 100 observations, respectively. The global number of clusters $K$ is 4 with means $-20, -10, 0, 15$ and standard deviations $2, 1, 2.5, 3$. Data in the first level are generated from a mixture of all four clusters with weights equal to $2/10, 3/10, 1/10, 4/10$. This is the only level where all 4 clusters are present, the second level does not include the cluster with mean equal to $-10$ while the third level does not include neither the one centered in $-20$ nor the one

DRAFT

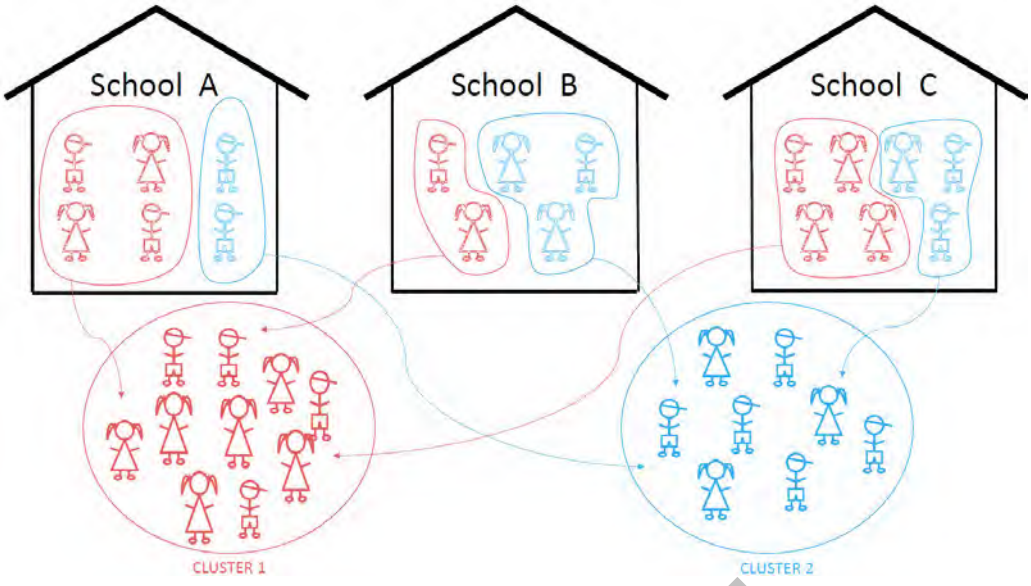



Figure 1: Visualization of local and global clustering. Each color represents a cluster

centered in 0. The mixture weights are equal to $3/10, 3/10, 4/10$ and $2/10, 8/10$, respectively. Given such framework, we generated 50 independent different datasets.

We follow a default non informative choice for both the process and $P_0$ hyperparameters. The first ones are $a_1 = 1$, $b_1 = 1$, $a_2 = 10$, $b_2 = 2$ while the second ones are $(\mu_0, \kappa_0, \nu_0, \sigma_0^2) = (0, 0.1, 10, 1)$. In all simulations, we started the MCMC with the empty partitions, i.e each data point is a cluster by its own. To perform posterior inference, we developed a marginal posterior sampler. Once the posterior sample has been collected, a point estimate for the partition by minimizing the expected posterior loss. Different choices can be made for the loss function, in this work we resort to the variation of information loss function (9) which is preferred to the Binder loss function (5) as tends to avoid clusters of size 1.

Throughout the 50 repetitions, the true partition has almost always been recovered correctly. The mean RI value is 0.985 with very little standard deviation (0.084). The estimated number of cluster has always been the correct one, that is 4, but only one time. Finally, Figure 2 shows the estimated density where the solid red line is the true density that generated the data while the solid blue line is the density estimated by our model, with the corresponding 2.5% and 97.5% credibility bands. In particular, we highlight how in the second level, our estimated density presents a small peak in $x = -10$ even though there are no points in that area. The peak is so small that it does not really influence the result, nevertheless it is a perfect example of the borrowing of information across the levels that is allowed by our hierarchical model.

Going forward, our efforts will be directed towards both theoretical and practical advancements. We intend to delve deeper into the distributional properties of the proposed model, focusing specifically on the distribution of the number of clusters induced by the prior. Additionally, we plan to apply the model to real-world environmental data to demonstrate its practicality.

DRAFT

## References

[1] Argiento, R., Cremaschi, A., Vannucci, M.: Hierarchical normalized completely random measures to cluster grouped data. J. Amer. Statist. Assoc. **115**(529), 318–333 (2020). DOI 10.1080/01621459.2019.1594833

[2] Argiento, R., De Iorio, M.: Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. Ann. Statist. **50**(5), 2641–2663 (2022). DOI 10.1214/22-aos2201

[3] Bassetti, F., Casarin, R., Rossini, L.: Hierarchical species sampling models. Bayesian Anal. **15**(3), 809–838 (2020). DOI 10.1214/19-BA1168

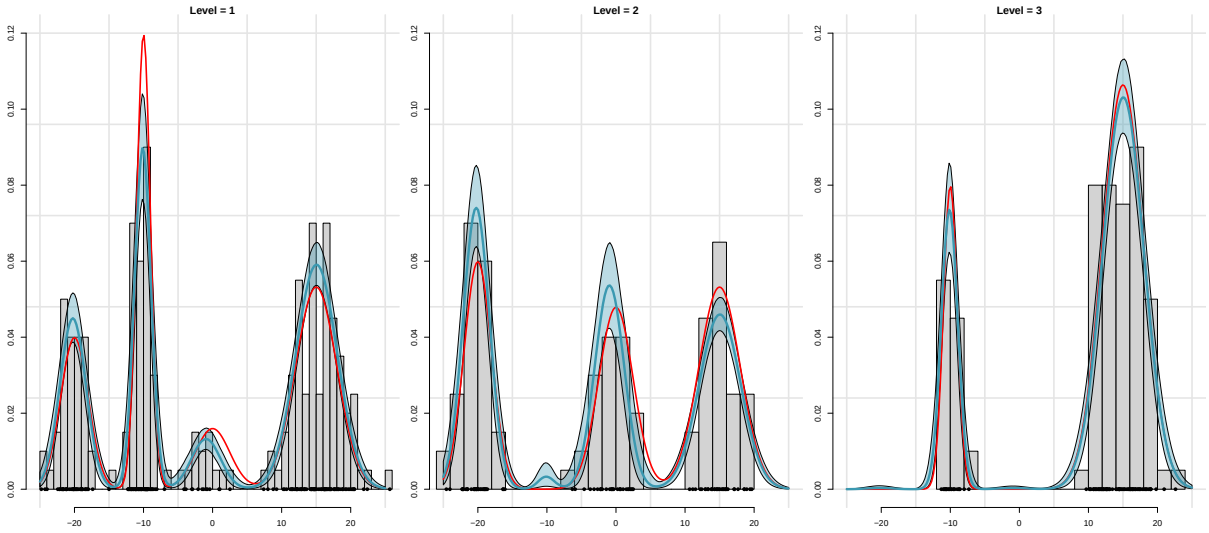



Figure 2: Each panel represents a level. Data points are reported on the bottom of the plot. On the background, histogram of the data (*grey bars*). Density estimation: the true density (*solid red line*), pointwise estimate using our proposed model (*solid blue line*) as well as its 95% credible band (*shaded blue area*).

DRAFT

[4] Beraha, M., Guglielmi, A., Quintana, F.A.: The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. Bayesian Analysis **16**(4), 1187 – 1219 (2021). DOI 10.1214/21-BA1278

[5] Binder, D.A.: Bayesian cluster analysis. Biometrika **65**(1), 31–38 (1978). DOI 10.1093/biomet/65.1.31

[6] Camerlenghi, F., Lijoi, A., Orbanz, P., Prünster, I.: Distribution theory for hierarchical processes. Ann. Statist. **47**(1), 67–92 (2019). DOI 10.1214/17-AOS1678

[7] D'Angelo, L., Canale, A., Yu, Z., Guindani, M.: Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data. Biometrics **0**(0), 1–13 (2022). DOI https://doi.org/10.1111/biom.13626

[8] Denti, F., Camerlenghi, F., Guindani, M., Mira, A.: A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. J. Amer. Statist. Assoc. **0**(0), 1–12 (2021). DOI 10.1080/01621459.2021.1933499

[9] Meilă, M.: Comparing clusteringsan information based distance. J. Multivariate Anal. **98**(5), 873–895 (2007)

[10] Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. pp. 560–585 (2003). DOI 10.1214/aos/1051027881. Dedicated to the memory of Herbert E. Robbins

[11] Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Amer. Statist. Assoc. **101**(476), 1566–1581 (2006). DOI 10.1198/016214506000000302

# On Bayesian power analysis in reliability

Fulvio De Santis[a], Stefania Gubbiotti[a], and Francesco Mariani[a]

[a]Dipartimento di Scienze Statistiche, Sapienza Università di Roma;
fulvio.desantis@uniroma1.it,
stefania.gubbiotti@uniroma1.it,
f.mariani@uniroma1.it

**Abstract**

Evaluation of reliability of a production process is a crucial step in sustainability assessment. In this article we consider the sample size determination problem when time-to-failure is modeled by a Rayleigh distribution. Following a hybrid Bayesian-frequentist approach, the selection of the number of units is based on the so-called probability of success (PoS) of the experiment, that is the expected value of the power function with respect to a design prior distribution for the mean failure time. This method works properly only if PoS is a representative summary of the distribution of the power function induced by the design prior. Therefore we derive and analyze the density of the power function for one-sided tests on the Rayleigh parameter, using conjugate design priors. Numerical examples are discussed.

***Keywords:*** Bayesian analysis, power function, probability of success, Rayleigh model, sample size determination.

DRAFT

## 1. Introduction

Sustainability is of great importance in the development of new technologies. However the life cycle of a product and the failure probability of its components strongly influence its overall environmental impact. Thus, evaluation of sustainability in the production process should be accompanied by the assessment of reliability since the earliest steps of design and production. In this paper we consider the Rayleigh distribution, often used to model time-to-failure in reliability analysis (see for instance (1)). We focus on sample size determination, one of the first issues to be addressed in experimental design, usually based on the power of a statistical test, that is a function of the unknown parameter of a model. Following the Bayesian approach to experimental design, instead of relying on a single design value we specify a design prior distribution that models possible uncertainty on the true value of the parameter of interest $\Theta$, here considered as a random variable. As a consequence, the power function is a random variable as well: its distribution is typically summarized by taking its expectation that is the so-called probability of success (PoS) of the experiment (2; 3; 4): the larger its value, the higher the chances of observing data supporting reliability. The limitations of PoS as a suitable summary of the random power function have been pointed out in (5) with reference to tests on the location parameter of the normal model. The Authors propose to study the whole distribution instead of considering only its expected value. In this paper we follow this idea with application to the scale parameter of a Rayleigh model. Specifically, we derive the closed-form expressions for the cumulative distribution function (cdf) and probability density function (pdf) of the power function, induced by a conjugate prior on $\Theta$.

## 2. Methodology

Let $\boldsymbol{X_n} = (X_1, \ldots, X_n)$ be a random sample from a Rayleigh distribution where $\theta \in \mathbb{R}^+$ is the scale parameter of interest. The density function is

$$f_X(x|\theta) = \frac{x}{\theta^2} \exp\left\{-\frac{x^2}{2\theta^2}\right\}, \quad x \geq 0, \quad \theta > 0. \tag{1}$$

It is easy to check that $T(\boldsymbol{x_n}) = \sum_{i=1}^n x_i^2$ is a sufficient complete statistic and that

$$W(\boldsymbol{X_n}, \theta) = \frac{T(\boldsymbol{X_n})}{\theta^2} \sim \chi^2_{2n} \tag{2}$$

is a pivotal quantity for $\theta$. Without loss of generality we consider the one-sided test $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$. The size-$\alpha$ uniformly most powerful test rejects the null hypothesis if $W(\boldsymbol{X_n}, \theta_0) \leq q(\alpha)$, where $q(\alpha)$ is the $\alpha$-level quantile of the $\chi^2_{2n}$ distribution. The power function is then

$$\begin{aligned} \beta_n(\theta) &= \mathbb{P}\left[W(\boldsymbol{X_n}, \theta_0) \leq q(\alpha) \,\middle|\, \theta\right] \\ &= \mathbb{P}\left[W(\boldsymbol{X_n}, \theta) \leq \left(\frac{\theta_0}{\theta}\right)^2 q(\alpha)\right] \\ &= \mathbb{F}\left[\left(\frac{\theta_0}{\theta}\right)^2 q(\alpha)\right] \end{aligned} \tag{3}$$

where $\mathbb{P}(\cdot)$ and $\mathbb{F}(\cdot)$ are the probability measure and the cdf of the $\chi^2_{2n}$ random variable. According to the Bayesian approach to experimental design (3), we assign a design distribution to $\Theta$ and we denote by $\mathbb{P}_\Theta(\cdot)$, $\mathbb{F}_\Theta(\cdot)$, $f_\Theta(\cdot)$, $\mathbb{E}_\Theta(\cdot)$ the corresponding probability measure, cdf, pdf and expected value. Specifically, we consider a conjugate square root inverse gamma as design prior (see (1)), i.e. $\Theta \sim \text{SqInvGa}(a, b)$, whose prior mode is $\tilde{\theta} = \sqrt{2b/(2a+1)}$ and prior sample size (that controls the precision of the distribution) is $n_d = 2a + 1$. Following (5), we now consider the random variable $\beta_n(\Theta)$ with cdf

$$\begin{aligned} G(y) &= \mathbb{P}_\Theta\left[\beta_n(\Theta) \leq y\right] \\ &= \mathbb{P}_\Theta\left(\mathbb{F}\left[\left(\frac{\theta_0}{\theta}\right)^2 q(\alpha)\right] \leq y\right) = \\ &= \mathbb{P}_\Theta\left[\left(\frac{\theta_0}{\theta}\right)^2 q(\alpha) \leq q(y)\right] = \\ &= 1 - \mathbb{F}_\Theta\left[\theta_0 \left(\frac{q(\alpha)}{q(y)}\right)^{1/2}\right]. \end{aligned} \tag{4}$$

Then, by deriving $G(y)$ with respect to $y$ we obtain the density function of $\beta_n(\Theta)$

$$g(y) = K \cdot \left[\theta_0^2 q(\alpha)\right]^{-a} \left[q(y)\right]^{a-n} \exp\left\{-\frac{q(y)}{2}\left[\frac{b}{q(\alpha)\theta_0^2} - 1\right]\right\}. \tag{5}$$

where $K = b^a \left[\Gamma(a)\Gamma(n)2^n\right]^{-1}$. PoS is then defined as

$$e_n = \mathbb{E}_\Theta[\beta_n(\Theta)] = \int_{\mathbb{R}^+} \beta_n(\theta) f_\Theta(\theta) d\theta = \int_0^1 y \cdot g(y) dy. \tag{6}$$

A closed-form expression for (6) is not available for this model, but application of Monte Carlo is straightforward to obtain numerical approximation of values of $e_n$, for each given $n$. As shown in the examples of Section 3, exploration of the features of $g(y)$ induced by the choice of $f_\Theta$ is useful to check

DRAFT

whether PoS is a representative summary of $g(y)$. In this case, assuming that $e_n$ is an increasing function of $n$, PoS can be used for sample size determination: we select the minimum sample size such that $e_n$ is above a suitable threshold $\lambda \in (0, 1)$, i.e.

$$n^{\star}_{\text{PoS}} = \min\{n \in \mathbb{N} : e_n > \lambda\}. \tag{7}$$

In the following, for comparison, we refer to the standard criterion for optimal sample size based on the power function, i.e.

$$n^{\star}_{\text{pow}} = \min\{n \in \mathbb{N} : \beta_n(\theta_d) > \lambda\}. \tag{8}$$

## 3. Example

In this section we consider the sample size problem for a reliability experiment on a new technology system, whose components are assumed to have a lifetime cycle well described by a Rayleigh distribution of unknown parameter $\theta$, measured on a given time scale. Suppose that we want to test $H_0 : \theta \geq \theta_0 = 15$ vs $H_1 : \theta < \theta_0 = 15$. First of all, for a given sample size $n = 20$, we study the impact of the design parameters on the density of the power function. In Figure 1, first row, the prior sample size is smaller than the actual sample size (e.g. $n_d = 10$), whereas in the second row the design prior gets more concentrated (e.g. $n_d = 30$). In both cases we consider three different values of the prior mode, $\theta_d = 10, 12, 15$. Table 3. reports the values of PoS computed for the same choices of the design

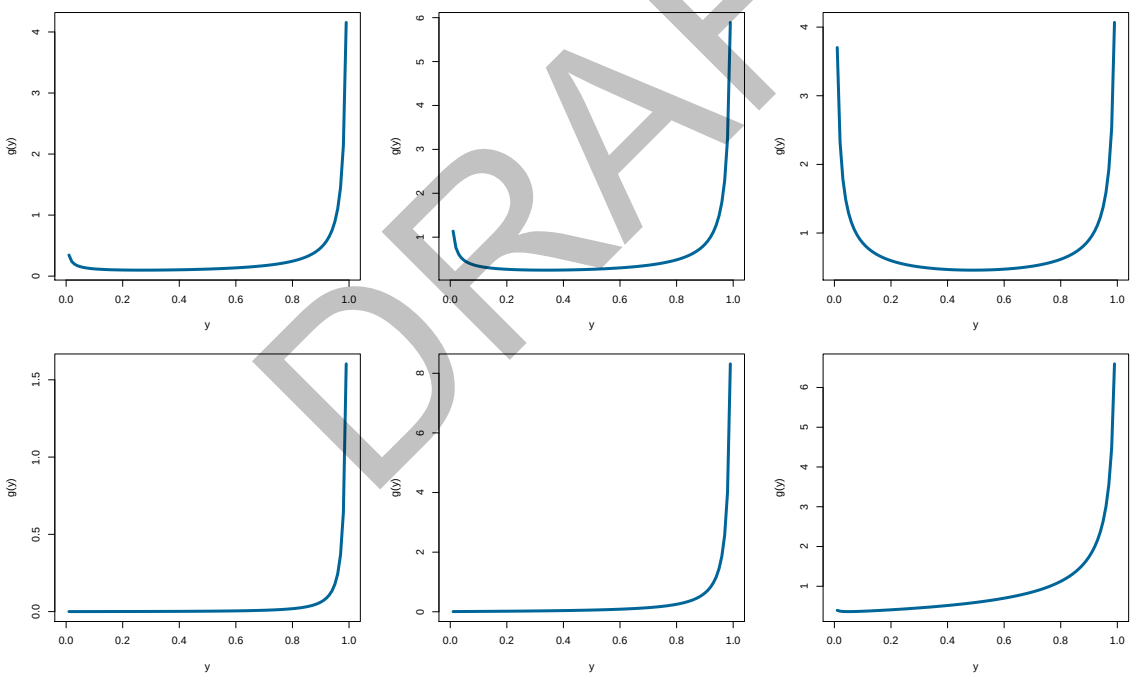



Figure 1: Behavior of $g(y)$ for $n = 20$, $\theta_0 = 15$ and for several choices of $\theta_d = 10, 12, 15$ (from left to right) and of $n_d = 10, 30$ (from top to bottom).

parameters. Note that, when $n_d = 30$, despite the choice of $\theta_d$, the density of the power is always an increasing function of $y$ that assigns large density to values of the power close to 1, as it must be in a well-designed experiment. Conversely, if the design prior sample size is not large enough, $g(y)$ is not well-behaved, i.e. it shows a u-shape that is more and more evident as $\theta_d$ gets closer and closer to the test threshold $\theta_0$. In particular, for $\theta_0 = \theta_d$ and $n_d = 10$, PoS gets as small as 0.511, a value with very low density: as highlighted in (5), this implies that PoS is not a representative summary of $g(y)$. Furthermore $g(y)$ assigns considerably high density to power values very close to 0, which is not desirable in practice.

DRAFT

Similar considerations on the qualitative features of $g(y)$ may be helpful in driving a more careful choice of the design parameters for selecting the sample size. As an example, we consider a square inverse gamma prior with mode at $\theta_d = 12$ and prior sample size $n_d = 30$. Figure 2 compares the frequentist power function evaluated at the design value $\theta_d = 12$ and PoS with the above design prior assumption, as the sample size increases. We notice a well known behavior of these two functions, highlighted for instance in (6): averaging with respect to a design prior fairly concentrated on $H_1$ has the effect of slightly raising PoS for small values of the sample size; whereas, when $n$ increases (in this example, $n > 40$), PoS becomes smaller than the frequentist power. Correspondingly, given a threshold $\lambda = 0.8$, we obtain $n^{\star}_{\text{PoS}} = 22$ and $n^{\star}_{\text{pow}} = 33$; but, if for instance $\lambda = 0.9$ to fulfill the sample size criteria (6) and (8) we obtain $n^{\star}_{\text{PoS}} = 97$ and $n^{\star}_{\text{pow}} = 45$, respectively. As a final remark, note that the standard approach for sample size determination, based on the power function, and the predictive Bayesian approach, based on PoS, coincide when the point-mass prior on the design value is assumed as design prior.

| $n_d$ | $\theta_d$ | | |
|---|---|---|---|
| | 10 | 12 | 15 |
| 10 | 0.923 | 0.797 | 0.511 |
| 30 | 0.997 | 0.967 | 0.717 |

Table 1: Values of PoS for $n = 20$, $\theta_0 = 15$ and for several choices of $\theta_d$ and $n_d$.

DRAFT

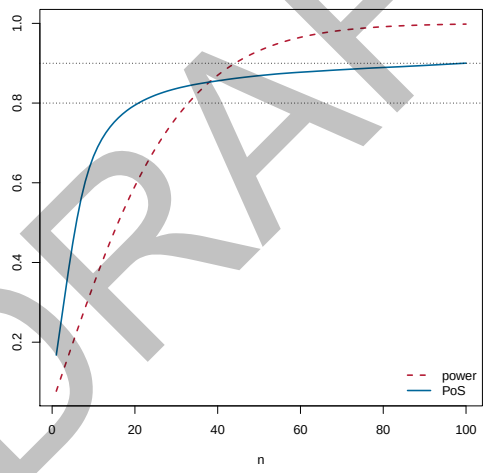



Figure 2: Power function (at design value $\theta_d = 12$) and PoS (with design prior parameters $\theta_d = 12$ and $n_d = 10$) as functions of the sample size.

## 4. Concluding remarks

PoS is a useful hybrid Bayesian-frequentist sample size determination tool, since it is based on the widely used frequentist power function of a test, but it also allows to take into account uncertainty on the design value. However, PoS is not always a suitable summary of the distribution of the power, as in the case of a u-shaped density of the power. In this sense, a deeper investigation on qualitative features of the density of the power in terms of the design parameters may be helpful: if the choice of the design prior is cautious, then bad-shaped densities can be avoided. When the goal of the experiment is to prove inferiority ($H_1 : \theta < \theta_0$), the design prior should be strongly concentrated on values of the parameter in $(0, \theta_0)$. In fact the design prior does not express information on the parameter to be combined with experimental evidence in order to obtain the posterior distribution. It rather models the scenario under which the study is planned. (See (6) for a discussion on the distinction between design and analysis priors). Finally, note that we here consider a conjugate design prior for the sake of analytical tractability. However, alternative choices are possible: for instance, the choice of truncated distributions fully concentrated on $H_1$ could be explored. In this case the proposed method can be easily implemented numerically via Monte Carlo.

## References


[1] Dey, S., Dey, T.: Bayesian estimation and prediction intervals for a Rayleigh distribution under a conjugate prior. Journal of Statistical Computation and Simulation, **82**(11): 1651–1660 (2012)

[2] O'Hagan, A., Stevens, J.W., Campbell, M.J.: Assurance in clinical trial design. Pharmaceutical Statistics **4**(3):187–201 (2005)

[3] Spiegelhalter, D.J., Freedman, L.S., Blackburn, P.R.: Monitoring clinical trials - conditional power or predictive power. Controlled Clinical Trials 7(1):8–17 (1986)

[4] Wang, Y., Fu, H., Kulkarni, P., Kaiser, C.: Evaluating and utilizing prob- ability of study success in clinical development. Clinical Trials 10(3):407–413 (2013)

[5] Rufibach, K., Burger, H.U., Abt, M.: Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. Pharmaceutical Statistics **15** 438–446 (2016)

[6] Brutti, P., De Santis, F., Gubbiotti, S.: Bayesian-frequentist sample size determination: a game of two priors Metron **72**(2), 133–151 (2014)


DRAFT

# Power priors elicitation through Bayes factors

Roberto Macrì Demartino[a], Leonardo Egidi[b], and Nicola Torelli[b]

[a]Department of Statistical Sciences, University of Padova,
roberto.macridemartino@phd.unipd.it
[b]Department of Economics, Business, Mathematics and Statistics, University of Trieste,
legidi@units.it, nicola.torelli@deams.units.it

**Abstract**

In the Bayesian framework, the power priors have been increasingly used in the context of the analysis of clinical trials and similar studies to incorporate external and past information, usually into the prior distribution of some treatment effect. Their use has been shown to be particularly effective in small sample size scenarios and when strong prior information is available. In a fully Bayesian approach, eliciting the initial distribution of the weight parameter controlling the amount of historical information remains a challenge, since it must be carefully chosen to reflect the available prior information accurately and not dominate the posterior inferential conclusions. We propose a novel preliminary method for eliciting the distribution of the weight parameter based on the Bayes factor, which allows the prior distribution will be updated based on the strength of the evidence the data provides.

***Keywords:*** Clinical trial, Historical information, Prior elicitation, Strength of evidence.

DRAFT

## 1. Introduction

In recent years, incorporating historical data into the design and analysis of new clinical trials has gained significant attention, especially for ethical reasons and when patient recruitment is challenging. Particularly, Bayesian methods have gained increasing popularity in this context (Liu, 2018; Nikolakopoulos et al., 2018; Ollier et al., 2020). Furthermore, one of the critical advantages of the Bayesian approach is the ability to elicit informative priors on the model parameters, allowing for the integration of historical data into the analysis. However, informative prior elicitation is widely recognized as a complex undertaking, given the inherent complexity of quantifying and synthesizing prior information into an appropriate prior distribution. Hence, there is a pressing need for developing techniques and methods that can facilitate synthesizing and quantifying prior information more effectively and efficiently (Ibrahim et al., 2015). Specifically, there is a growing concern regarding the adaptive incorporation of historical data, particularly in the presence of data heterogeneity and rapid changes of the initial trial conditions (Ollier et al., 2020).

In this context, one possible approach is using *power priors* (Chen and Ibrahim, 2000), which allows the historical data to influence the prior distribution in a flexible and controlled way. A pivotal role in the power prior methodology is played by a *weight parameter* $\delta$, defined between 0 and 1, that determines the degree to which historical data influences the prior distribution. Multiple strategies exist to specify the parameter $\delta$. One straightforward approach involves fixing $\delta$ on a predetermined value that is considered reasonable based on background knowledge regarding the similarity of the two studies. Particularly,

let consider an historical dataset $D_0$ with corresponding likelihood $L(\boldsymbol{\theta}\,|D_0)$: the basic power prior formulation is

$$\pi\left(\boldsymbol{\theta}|D_0,\delta\right) \propto L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta}), \tag{1}$$

where $0 \leq \delta \leq 1$ is the scalar weight parameter and $\pi_0(\boldsymbol{\theta})$ is the *initial prior* for $\boldsymbol{\theta}$ before observing the historical data $D_0$. The parameter $\delta$ plays a crucial role in determining the shape of the prior distribution for $\boldsymbol{\theta}$, as specified in Equation (1). As $\delta$ decreases, the tails of the distribution become heavier (Ibrahim et al., 2015). When the analyst fixes the weight parameter, sensitivity analysis should be used to identify an appropriate level of borrowing information while accounting for prior-data conflict (Evans and Moshonov, 2006). Several statistical methods have been proposed, including the penalized likelihood-type criterion (PLC), marginal likelihood criterion (MLC), deviance information criterion (DIC), logarithm of the pseudo-marginal likelihood (LPML) criterion, as well as the empirical Bayes (EB) method (see Ibrahim et al., 2015; Gravestock and Held, 2017).

A modification of the power prior, which includes a normalizing factor, allows specifying a hierarchical prior specification by taking $\delta$ as a random quantity (Chen and Ibrahim, 2000; Duan et al., 2006). Hence, the *joint normalized power prior* for $(\delta, \boldsymbol{\theta})$ is

$$\pi\left(\boldsymbol{\theta},\delta|D_0\right) = \frac{L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})\pi_0(\delta)}{\int_{\Theta} L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2}$$

where $\pi_0(\delta)$ is an initial prior distribution for $\delta$. One of the primary advantages of the joint normalized power prior is its adherence to the likelihood principle, which ensures that the posterior distributions accurately reflect the compatibility between the current and historical data. Therefore, it enables adaptive borrowing according to prior-data conflict (Ye et al., 2022).

In the context of a power prior approach, the ability to effectively elicit an appropriate initial prior distribution for the weight parameter $\delta$ is a critical step. However, as far as we know from the current literature, this issue has yet to be explored and fully motivated. To achieve the goal mentioned above, we propose to compare the Bayes factor (Jeffreys, 1998; Kass, 1993; Kass and Raftery, 1995), henceforth BF, of competing prior distributions for $\delta$ to determine which one provides the best fit for the available data. This sort of reverse-Bayes approach (Good, 1950) would allow for a more reliable and informed choice between competing priors, which is crucial in ensuring improvements on the robustness and accuracy of the results. Specifically, we aim to present some preliminary results concerning a well-known example consisting of two clinical trials about the efficacy of an interferon treatment by using the probabilistic programming language Stan (Carpenter et al., 2017) and the `bridgesampling` `R` package (Gronau et al., 2020).

DRAFT

## 2. Elicitation of the $\delta$ initial prior

Determining the optimal value for the weight/discount parameter through sensitivity analysis can only be applied in cases where the parameter $\delta$ is fixed. However, eliciting an appropriate initial prior distribution has proven challenging when the discount parameter is a random variable. Only a few methods have been developed for this aim.

The Bayes factor constitutes a valuable statistical tool in prior elicitation by comparing the marginal likelihoods associated with the joint posterior of competing models that incorporate different initial prior distributions for the weight parameter. This approach is beneficial when prior information is limited and can aid in selecting a prior distribution well-supported by the observed data. Specifically, given the joint power prior distribution (2) and in light of the current data $D$, the joint posterior is

$$\pi\left(\boldsymbol{\theta},\delta|D,D_0\right) = \frac{L\left(\boldsymbol{\theta}|D\right)L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})\pi_0(\delta)}{\int_{\Theta} L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} \times \left[\int_{\Psi} \frac{L\left(\boldsymbol{\theta}|D\right)L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})\pi_0(\delta)}{\int_{\Theta} L\left(\boldsymbol{\theta}|D_0\right)^{\delta}\pi_0(\boldsymbol{\theta})d\boldsymbol{\theta}} d\boldsymbol{\psi}\right]^{-1}, \tag{3}$$

where $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \delta\}$ contains both the model parameters $\boldsymbol{\theta}$ and the weight parameter $\delta$. The second term in Equation (3) represents the inverse marginal likelihood. Moreover, the normalized power prior yields a

marginal likelihood that is not analytically tractable and must be approximated using numerical methods. In recent years, a popular approach for estimating the marginal likelihood is the so-called *bridge sampling* (Meng and Wong, 1996). This method is a Monte Carlo technique that involves generating samples from an auxiliary distribution that bridges the model's prior distribution and posterior distribution. The generated samples are then used to calculate the bridge sampling weights, through which then the bias introduced by the auxiliary distribution is corrected, and an unbiased estimate of the marginal likelihood is obtained.

As far as we know from reviewing the existing literature, no author addressed using BFs to discriminate between alternative power prior choices. Given two competing model specifications, $\mathcal{M}_1$ and $\mathcal{M}_2$, with different initial priors on the weight parameter $\delta$, the corresponding Bayes factor can be expressed as follows

$$BF_{12} = \frac{\displaystyle\int_{\Psi} \frac{L\left(\boldsymbol{\theta}|D\right) L\left(\boldsymbol{\theta}|D_0\right)^{\delta} \pi_0(\boldsymbol{\theta}) \pi_0(\delta|\mathcal{M}_1)}{\int_{\Theta} L\left(\boldsymbol{\theta}|D_0\right)^{\delta} \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} d\boldsymbol{\psi}}{\displaystyle\int_{\Psi} \frac{L\left(\boldsymbol{\theta}|D\right) L\left(\boldsymbol{\theta}|D_0\right)^{\delta} \pi_0(\boldsymbol{\theta}) \pi_0(\delta|\mathcal{M}_2)}{\int_{\Theta} L\left(\boldsymbol{\theta}|D_0\right)^{\delta} \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} d\boldsymbol{\psi}}, \tag{4}$$

where $\pi_0(\delta|\mathcal{M}_1)$ and $\pi_0(\delta|\mathcal{M}_2)$ denote the initial prior distribution for $\delta$ under the two model specifications, respectively. As suggested by Jeffreys (1998), Bayes factors larger than one suggest evidence in favor of model $\mathcal{M}_1$. In contrast, values smaller than one suggest evidence in favor of model $\mathcal{M}_2$: values close to one indicate inconclusive evidence.

As illustrated by a practical application in the next section, the use of BFs proposed in Equation (4) could then serve as a further comparison tool to discriminate between choices for the weight parameter $\delta$. Moreover, this represents to us only a procedural starting point: it would then be of interest to explore some theoretical properties of the power priors BF above, such as the consistency and the information consistency (Kass and Raftery, 1995), and formulate an advanced calibrated protocol for a better elicitation of $\delta$, in the same spirit with the calibration purposes of Garcia-Donato and Chen (2005).

DRAFT

## 3. Application

To illustrate the application of the Bayes factor in the context of prior elicitation for the weight parameter, we consider a scenario in which a current clinical trial is analyzed in conjunction with historical information obtained from a previous study, possibly to corroborate the earlier results about medical treatment. The analysis involves data from two well-known clinical trials on phase III melanoma, a two-arm and a three-arm trial conducted by the Eastern Cooperative Oncology Group (ECOG)—see Ibrahim et al. (2012, 2015) for further details. Specifically, we examine data from the historical trial, E1684, and the current trial, E1690, which involved 284 and 427 patients, respectively. The clinical trials under consideration assigned patients to either receive interferon treatment (IFN) or a placebo. The primary outcome variable was the survival time, which was measured from the time of randomization to death. The analysis also included three covariates: standardized age, sex, and performance status (PS), an indicator of the patient's activity level.

As widely documented by Ibrahim et al. (2012), this is a paradigmatic example of the use of previous practical information in clinical trials: the E1690 was conducted to corroborate the efficacy results from the previous E1684 trial about the IFN treatment; however, the two studies appear to be in conflict one each other. Therefore, the choice/elicitation for $\delta$ is crucial and may lead to controversial clinical findings.

Here, we consider the cure rate model developed by Chen et al. (1999). This statistical model can estimate the proportion of individuals who experience long-term survival and the distribution of survival times for the remaining subjects. Particularly, let $N$ denote the number of carcinogenic cells that remain after initial treatment. We assume that $N$ follows a Poisson distribution with rate parameter $\theta$. Moreover, consider a set of i.i.d random variables $R_k$, with $k = 1, \ldots, N$, characterized by the distribution function $F(t) = 1 - S(t)$. Thus, the variable of interest is represented by the time of relapse and is denoted by

$T = min(R_k)$, with $0 \leq k \leq N$. Let suppose to observe a set of i.i.d durations $\boldsymbol{Y} = (y_1, \ldots, y_n)$ with an associated right censoring vector $\boldsymbol{C} = (c_1, \ldots . c_n)$ and a $n \times p$ design matrix $\boldsymbol{X}$, the likelihood function is then

$$L(\boldsymbol{\beta}, \boldsymbol{\xi} \mid \boldsymbol{Y}, \boldsymbol{C}) = \prod_{i=1}^{n} \left\{ p\left(y_i \mid \boldsymbol{\xi}\right) \theta_i \right\}^{c_i} e^{-F(y_i|\boldsymbol{\xi})\theta_i},$$

where $\boldsymbol{\beta}$ is the regression coefficient vector, $\theta_i = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})$ and $\boldsymbol{\xi} = (\alpha, \lambda)$ is the parameter vector of a Weibull distribution. We adopt a hierarchical prior structure, which is designed as follows

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}\left(\mathbf{0}, 10\boldsymbol{I}_p\right), \\ \lambda &\sim \mathrm{N}(0, 10), \\ \alpha &\sim \mathrm{Gamma}\left(1, 1\right), \end{aligned}$$

where $\boldsymbol{I}_p$ is the identity matrix of rank $p$. Furthermore, let $D_0 = (\boldsymbol{Y_0}, \boldsymbol{C_0}, \boldsymbol{X_0})$ be the historical dataset, then the normalized joint power prior takes the following form:

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\xi}, \delta | D_0\right) = \frac{L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) \pi_0\left(\delta\right)}{\int_{\Phi} L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) d\boldsymbol{\phi}},$$

where $\boldsymbol{\phi} = \{\boldsymbol{\beta}, \boldsymbol{\xi}\}$. Consequently, the joint posterior distribution is

$$\begin{aligned} \pi\left(\boldsymbol{\beta}, \boldsymbol{\xi}, \delta | D, D_0\right) = & \frac{L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D\right) L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) \pi_0(\delta)}{\int_{\Phi} L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) d\boldsymbol{\phi}} \times \\ & \times \left[ \int_{\Psi} \frac{L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D\right) L\left(\boldsymbol{\beta}, \boldsymbol{\xi} \mid D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) \pi_0(\delta)}{\int_{\Phi} L\left(\boldsymbol{\beta}, \boldsymbol{\xi} | D_0\right)^{\delta} \pi_0(\boldsymbol{\beta}, \boldsymbol{\xi}) d\boldsymbol{\phi}} d\boldsymbol{\psi} \right]^{-1}, \end{aligned}$$

where $D = (\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{X})$ denotes the current dataset. We assume as the initial prior for the discount parameter a Beta distribution, $\delta \sim \text{Beta}(\eta, \nu)$, with different choices for the hyper-parameters $(\eta, \nu)$, as described in Table 1.

To establish the most suitable initial prior specification for the weight parameter $\delta$, we calculate the BF in Equation (4) for some competing model combinations. Additionally, we consider optimal fixed values of $\delta$ that have been previously identified by Ibrahim et al. (2012) and Ibrahim et al. (2015) as the options that yield the best model fit according to some Bayesian criteria, such as the Deviance Information Criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML). By doing so, we can evaluate the strength of evidence in favor of each model specification and select the one that best fits the data. Table 1 shows twice the natural logarithm of the Bayes factor for all the assumed competing models. We can notice that the Beta distribution with hyper-parameters $\eta = 1$ and $\nu = 5$ emerges as the most favorable initial prior specification for the weight parameter $\delta$. Specifically, we observe that the BF transformation associated with the Beta$(1, 5)$ distribution largely exceeds 0—as a consequence, the original BF is much greater than 1—indicating that this specification outperforms the other priors and the fixed values choices across all the considered scenarios. Notice that the chosen prior specification implies an average value for $\delta$ equal to 0.17 (with an approximated variance of 0.02), which is less than half of the suggested value of $\hat{\delta} = 0.4$, previously obtained by Ibrahim et al. (2012) using the DIC criterion

We posit that through the above prior elicitation, we imply a more conservative procedure for incorporating historical information into the model. This result makes sense, given the conflict between the two trials' outcomes. Thus, our approach based on the Bayes factor has the potential to minimize the impact of subjective biases and ensure robustness in the inferential results.

DRAFT

| | $\delta \sim \text{Beta}(\eta, \nu)$ | | | | | | | | | Fixed $\delta$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **( 1, 2)** | **(1,5)** | **(2,5)** | **(2,10)** | **(5,5)** | **(2,1)** | **(5,1)** | **(5,2)** | **(10,2)** | $\hat{\delta} = 0.2$ | $\hat{\delta} = 0.4$ | $\hat{\delta} = 0.5$ |
| **(1,1)** | -1.37 | -3.23 | 5.16 | 2.64 | 28.57 | 10.54 | 38.16 | 34.61 | 72.44 | 207.47 | 360.60 | 436.72 |
| **(1,2)** | | -1.87 | 6.53 | 4.00 | 29.94 | 11.91 | 39.53 | 35.97 | 73.80 | 208.84 | 361.96 | 438.09 |
| **(1,5)** | | | 8.39 | 5.87 | 31.80 | 13.77 | 41.39 | 37.84 | 75.67 | 210.71 | 363.83 | 439.95 |
| **(2,5)** | | | | -2.52 | 23.41 | 5.38 | 33.00 | 29.45 | 67.28 | 202.31 | 355.44 | 431.56 |
| **(2,10)** | | | | | 25.93 | 7.90 | 35.52 | 31.97 | 69.80 | 204.84 | 357.96 | 434.08 |
| **(5,5)** | | | | | | -18.03 | 9.59 | 6.04 | 43.87 | 178.90 | 332.03 | 408.15 |
| **(2,1)** | | | | | | | 27.62 | 24.07 | 61.90 | 196.93 | 350.06 | 426.18 |
| **(5,1)** | | | | | | | | -3.55 | 34.28 | 169.31 | 322.44 | 398.556 |
| **(5,2)** | | | | | | | | | 37.83 | 172.87 | 325.99 | 402.11 |
| **(10,2)** | | | | | | | | | | 135.04 | 28.16 | 364.28 |
| $\hat{\delta} = 0.2$ | | | | | | | | | | | 153.12 | 229.25 |
| $\hat{\delta} = 0.4$ | | | | | | | | | | | | 76.12 |

Table 1: Upper triangular matrix of twice the natural logarithm of the Bayes factor, with distinct choices for the Beta hyper-parameters $(\eta, \nu)$ and three optimal fixed values for $\delta$. Models in the BF numerators appear in rows, and models in the BF denominators in columns, respectively.

DRAFT

## 4. Discussion

This paper provides a preliminary and concise assessment of applying the Bayes factor for power prior elicitation in a fully Bayesian approach where a Beta distribution is specified for the discount parameter $\delta$. Our preliminary application suggests that the Bayes factor is a valuable and powerful measure of evidence for discriminating between some competing prior specifications.

However, there are many directions for future research. First, using the BF as a sole criterion is limited due to its asymmetric prior-predictive distribution across models. Therefore, there is a need to explore the potential of a calibrated version of the Bayes factor because, before observing the data, the Bayes factor is a random variable that follows its own sampling distribution (Garcia-Donato and Chen, 2005). Then, we should derive a specific version of the prior-predictive $p$-value that considers and assesses the possibility of a prior-data conflict. Such an approach would permit the selection of competing models based on a more accurate assessment of the available evidence. Finally, consistency checks for the proposed BF should be addressed and studied.

## References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1â32.

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46 – 60.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new Bayesian model for survival data with a

surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919.

Duan, Y., Ye, K., and Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17(1):95–106.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1(4):893 – 914.

Garcia-Donato, G. and Chen, M.-H. (2005). Calibrating Bayes factor under prior predictive distributions. *Statistica Sinica*, 15(2):359–380.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin & Company Limited, London.

Gravestock, I. and Held, L. (2017). Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceutical Statistics*, 16(5):349–360.

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29.

Ibrahim, J. G., Chen, M.-H., and Chu, H. (2012). Bayesian methods in clinical trials: a Bayesian analysis of ecog trials e1684 and e1690. *BMC Medical Research Methodology*, 12(1):1–12.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749.

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42(5):551–560.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Liu, G. F. (2018). A dynamic power prior for borrowing historical data in noninferiority trials with binary endpoint. *Pharmaceutical Statistics*, 17(1):61–73.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Nikolakopoulos, S., van der Tweel, I., and Roes, K. C. B. (2018). Dynamic borrowing through empirical power priors that control type i error. *Biometrics*, 74(3):874–880.

Ollier, A., Morita, S., Ursino, M., and Zohar, S. (2020). An adaptive power prior for sequential clinical trials–Application to bridging studies. *Statistical Methods in Medical Research*, 29(8):2282–2294. PMID: 31729275.

Ye, K., Han, Z., Duan, Y., and Bai, T. (2022). Normalized power prior Bayesian analysis. *Journal of Statistical Planning and Inference*, 216:29–50.

DRAFT

# Predictive Bayes factors

Leonardo Egidi[a] and Ioannis Ntzoufras[b]

[a]Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche *'Bruno de Finetti'*, Università degli Studi di Trieste, legidi@units.it
[b]Department of Statistics, Athens University of Economics and Business, ntzoufras@aueb.gr

**Abstract**

Bayes factors represent one of the most well-known and commonly adopted tools to perform model selection and hypothesis testing according to a Bayesian flavour. Nevertheless, they are often criticized due to some interpretative and computational aspects, including that of not being able to be used with improper priors, or their intrinsic lack of calibration. Another criticism refers to the fact that they measure the model weight of evidence in terms of prior-predictive distributions, but they are rarely used to measure the predictive accuracy arising from competing models. In this paper we tried to fill this gap by proposing a new algorithmic protocol to transform Bayes factors into measures that evaluate the pure and intrinsic predictive capabilities of models in terms of posterior predictive distributions.

***Keywords:*** model selection, calibration, predictive accuracy, algorithmic protocol, posterior predictive distribution.

DRAFT

## 1. Introduction

Nowadays Bayes factors—hereafter, BFs—(7; 8; 9) still represent one of the cornerstones commonly adopted for Bayesian model selection and hypothesis testing. Given some data denoted by $\boldsymbol{y}$ and two competing parametric statistical models $\mathcal{M}_1, \mathcal{M}_2$, the Bayes factor $BF_{12}(\boldsymbol{y})$ is defined as the ratio between the two model marginal likelihoods, and measures the amount of evidence provided by the first model over the second one; ultimately, the BF is used to select which of the two models fits the data better according to their marginal likelihoods.

Although their recognised relevance in Bayesian model selection and hypothesis testing to discriminate between two candidate models or two alternative hypotheses, their use is limited and constrained by some well-known computational and interpretative issues.

First, BFs cannot be applied with improper prior distributions, say $p(\theta) \propto k$, otherwise the ratio would depend on some subjective constants, making then model selection driven by the numeric constant $k$ itself. To fix this issue, BF variants such as the intrinsic Bayes factors (1) and the fractional Bayes factor (11) have been proposed.

Second, the scale of evidence favouring one model against another (7) based solely on the observed value of the BF is usually not appropriate, for such reason BFs could benefit from a calibration via prior-predictive distribution (4) to obtain more robust results.

Third, BFs are usually considered as the Bayesian counterpart of classical $p$-values (6), and as such characterized by a non-negligible amount of subjectivity for choosing between alternative models/hypotheses; on the other hand, the weight of evidence is usually expressed by use of Jeffreys' or similar subjective numerical scales (7).

Fourth, BFs measure the weight of model evidence by use of pairwise model comparisons with respect to some training data, and this comparison could be carried out even *before* the single models are fitted; as brilliantly remarked by (2), Bayes factors measure prior predictive performance rather than posterior predictive performance.

From the aforementioned criticisms, it emerges how BFs represent a valid but somehow controversial tool to perform model selection in the Bayesian framework. However, we feel their use could be extended towards a pure prediction setting (13) to embrace and capture the *'prediction fever'* raised by the modern statistical learning field, given the need for reliable and scientific predictive tools. For such reasons, developing theoretically solid and calibrated Bayes factors with a predictive purpose (12) could be fruitful in many settings and applications, especially for those fields at the boundaries between pure statistical methods and machine learning. To this aim, we propose in this paper a new approach based on the notion of predictive Bayes factors, consisting of a BF evaluation with respect to posterior predictive distribution samples corresponding to distinct test sets. This proposal could be then used as an alternative measure to evaluate predictive accuracy between two competing models, with the desirable property of reporting a measure of the uncertainty arising from the posterior predictive distributions (hereafter, ppd).

The remainder of the paper is organized as follows. Section 2. briefly describes the predictive BF algorithmic protocol. In section 3. we propose an illustrative example of linear regression with $g$-priors, whereas Section 4. concludes.

DRAFT

## 2. A new predictive protocol

Before observing the data $\boldsymbol{y}$, the Bayes factor $BF(\boldsymbol{Y})$ is a random variable following its own distribution (4). We could mimic the same argument not for the data at hand $\boldsymbol{Y}$, but, rather, for some observable but not yet observed test/out-of-sample data, $\boldsymbol{Y}^{pred}$. The final aim could then be that of exploring some theoretical and practical properties of the resulting BF between two candidate models according to a predictive protocol, and select then the best model on the ground of a pure forecasting comparison.

To this aim, we could generate hypothetical test-data replications from the models' posterior predictive distribution for many distinct test/training set configurations, and use these new values for a predictive BF evaluation. In such way, we would have a pure predictive discrimination tool, and the BFs would be evaluated under a posterior predictive perspective; then, the distribution of the predictive BFs, or some credible intervals, could be displayed against the observed value of the BF computed on the test set as a further comparison. We summarise the main steps from our algorithmic protocol in Algorithm 1.

The steps described in Algorithm 1 provide an intrinsic uncertainty coming from the posterior predictive distribution of the generated samples: the only evaluation of the observed BF on the distinct test sets obtained at each algorithmic iteration is not sufficient for acknowledging the amount of variability in the predictions from the ppds, and would not provide a pure probabilistic measure of predictive accuracy. Rather, we need some probabilistic statements about the identification of the best model in forecasting terms, and this can be obtained only by inspection of the posterior predictive distribution.

Note that in our protocol many inputs could be let to the users' preferences, such as the training and test set sample sizes, the model formulation, the prior distributions, or the choice of eventual hyper-prior distributions with hyper-parameters.

## 3. Sketch of application: linear regression with $g$-priors

As an illustrative example, we consider two nested linear regression models $\mathcal{M}_o \subset \mathcal{M}_f$, where $\mathcal{M}_o$ ($\mathcal{M}_f$) has $p_o$ ($p_f$) parameters, with $p_o < p_f$. We assume some $g$-priors (14; 10) for

**Algorithm 1** Predictive Bayes factors

Inputs: dataset $\boldsymbol{y}$; training and test set sample sizes, $n_{train}, n_c$
Models $\mathcal{M}_o, \mathcal{M}_f$
Priors $\pi(\boldsymbol{\beta}_o|\mathcal{M}_o)$, $\pi(\boldsymbol{\beta}_f|\mathcal{M}_f)$
Parameters $\boldsymbol{\beta}_o,\ \boldsymbol{\beta}_f$
Estimates $\hat{\boldsymbol{\beta}}_o,\ \hat{\boldsymbol{\beta}}_f$

For $t = 1, 2, \ldots, T$ repeat:

1. Create a data-split: $\boldsymbol{y} = (\boldsymbol{y}^{O(t)}, \boldsymbol{y}^{C(t)})$, where $\boldsymbol{y}^{O(t)}$ has dimension $n_{train}$, and $\boldsymbol{y}^{C(t)}$ has dimension $n_c$.
2. According to both the models considered, generate two datasets from the posterior predictive distributions: $\boldsymbol{y}_o^{pred} \sim p(\boldsymbol{y}^{pred}|\boldsymbol{y}, \mathcal{M}_o)$, $\boldsymbol{y}_f^{pred} \sim p(\boldsymbol{y}^{pred}|\boldsymbol{y}, \mathcal{M}_f)$, each of size $n_c$.
3. Calculate the two predictive Bayes factors:

$$BF_{of}(\boldsymbol{y}_o^{pred}) = \frac{\int p(\boldsymbol{y}_o^{pred}|\boldsymbol{X}_o, \boldsymbol{\beta}_o, \mathcal{M}_o)\pi(\boldsymbol{\beta}_o|\mathcal{M}_o)d\boldsymbol{\beta}_o}{\int p(\boldsymbol{y}_f^{pred}|\boldsymbol{X}_f, \boldsymbol{\beta}_f, \mathcal{M}_f)\pi(\boldsymbol{\beta}_f|\mathcal{M}_f)d\boldsymbol{\beta}_f}$$

$$BF_{of}(\boldsymbol{y}_f^{pred}) = \frac{\int p(\boldsymbol{y}_o^{pred}|\boldsymbol{X}_o, \boldsymbol{\beta}_o, \mathcal{M}_o)\pi(\boldsymbol{\beta}_o|\mathcal{M}_o)d\boldsymbol{\beta}_o}{\int p(\boldsymbol{y}_f^{pred}|\boldsymbol{X}_f, \boldsymbol{\beta}_f, \mathcal{M}_f)\pi(\boldsymbol{\beta}_f|\mathcal{M}_f)d\boldsymbol{\beta}_f}.$$

4. Produce the ppd for these quantities and derive credible intervals.
5. Compute the observed BF on the test set, $BF(\boldsymbol{y}^C)$, and compare it with the BFs predictive distributions.
6. Assess when $\mathcal{M}_o$ is better than $\mathcal{M}_f$.

the model parameters, denoted by vectors $\boldsymbol{\beta}_o$ and $\boldsymbol{\beta}_f$, with dimensions $p_o$ and $p_f$, respectively: $\boldsymbol{\beta}_o \sim \mathcal{N}(\mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}_o'\boldsymbol{X}_o)^{-1})$, $\boldsymbol{\beta}_f \sim \mathcal{N}(\mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}_f'\boldsymbol{X}_f)^{-1})$, where $\boldsymbol{X}_o$ ($\boldsymbol{X}_f$) is a $n \times p_o$ ($n \times p_f$) predictor matrix of full rank $p_o$ ($p_f$), $g$ is the scaling factor controlling variable/model selection, and $\phi$ is a dispersion parameter. $g$-prior distributions have been widely used in Bayesian variable and model selection settings because of their computational efficiency for marginal likelihoods evaluation and model search and because of their immediate interpretation. As widely remarked by the related literature, the choice of $g$ affects the model selection, with large $g$ favouring parsimonious models with a few large coefficients, and small $g$ tending to concentrate the prior on saturated models with small coefficients—see (10) for a review about some usual choices for $g$, including the *unit information prior* by (9).

One of the main advantages from the Zellner's $g$-priors is that the marginal likelihoods can be written in closed-form expressions; as shown by (10), the resulting Bayes factor in the null-based approach, computed by comparing the posed model $\mathcal{M}_o$ with the null model $\mathcal{M}_n$ having only the intercept, is given by:

$$BF_{on}(\boldsymbol{y}) = (1+g)^{(n-p_o-1)/2}[1+g(1-R_o^2)]^{-(n-1)/2}, \tag{1}$$

where $R_o^2$ represents the usual coefficient of determination under the model $\mathcal{M}_o$. The Bayes factor under the full-based approach, computed by comparing the model $\mathcal{M}_o$ with the full model $\mathcal{M}_f$, is instead given by:

$$BF_{of}(\boldsymbol{y}) = (1+g)^{-(n-p_f-1)/2}\left[1+g\frac{1-R_f^2}{1-R_o^2}\right]^{(n-p_o-1)/2}, \tag{2}$$

where $R_f^2$ represents the coefficient of determination under the model $\mathcal{M}_f$.

DRAFT

To implement the protocol procedure described through the Algorithm 1 in Section 2, we consider a total of $T = 10^3$ data-splits for a sample of $n = 100$ simulated data, where the sample sizes for the training and test set are fixed to $n_{train} = 70$ and $n_c = 30$, respectively; for each data split $t$, we generate $n_{rep} = 10^3$ replicated datasets under both the models from the posterior predictive multivariate student-$t$ distribution as follows:

$$\boldsymbol{y}_o^{pred} \sim t_{n_c}(\omega \boldsymbol{X}_o \hat{\boldsymbol{\beta}}_o, s_o^2(\boldsymbol{I} + \omega \boldsymbol{X}_o(\boldsymbol{X}_o'\boldsymbol{X}_o)^{-1}\boldsymbol{X}_o')),$$
$$\boldsymbol{y}_f^{pred} \sim t_{n_c}(\omega \boldsymbol{X}_f \hat{\boldsymbol{\beta}}_f, s_f^2(\boldsymbol{I} + \omega \boldsymbol{X}_f(\boldsymbol{X}_f'\boldsymbol{X}_f)^{-1}\boldsymbol{X}_f')),$$

where $\omega = g/(g+1)$; $s_o^2 = RSS_o + \frac{1}{g+1}\boldsymbol{\beta}_o'\boldsymbol{X}_o'\boldsymbol{X}_o\boldsymbol{\beta}_o$ and $s_f^2 = RSS_f + \frac{1}{g+1}\boldsymbol{\beta}_f'\boldsymbol{X}_f'\boldsymbol{X}_f\boldsymbol{\beta}_f$ are the sample variance estimates; $\boldsymbol{X}_o, \boldsymbol{X}_f$ denote the predictor matrices for the two models, with dimensions $n \times p_o$ and $n \times p_f$, respectively; and $\hat{\boldsymbol{\beta}}_o = (\boldsymbol{X}_o'\boldsymbol{X}_o)^{-1}\boldsymbol{X}_o'\boldsymbol{y}$, $\hat{\boldsymbol{\beta}}_f = (\boldsymbol{X}_f'\boldsymbol{X}_f)^{-1}\boldsymbol{X}_f'\boldsymbol{y}$ are the maximum likelihood estimates for $\boldsymbol{\beta}_o$, $\boldsymbol{\beta}_f$, respectively. The 95% credible intervals for the Bayes factors $BF_{of}(\boldsymbol{y}_o^{pred})$ and $BF_{of}(\boldsymbol{y}_f^{pred})$ under two distinct simulated scenarios are reported in Figure 1. The top panel displays the distribution of the two Bayes factors under the hypothesis that the original data have been generated from the 'true' model $\mathcal{M}_o$: $y = \beta_0 + \beta_1 x + \epsilon$, with $\beta_0 = 3, \beta_1 = 0.3$, whereas in the bottom panel the data have been generated from the 'true' full model $\mathcal{M}_f$ : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$, with $\beta_0 = 3, \beta_1 = 0.3, \beta_2 = 0.5$ and $\beta_3 = 0.7$. It is worth noting that both the plots display a large evidence for the true model in terms of predictive BFs: in the first scenario $\mathcal{M}_o$ is dramatically favoured with respect to both the predictive datasets $\boldsymbol{y}_o^{pred}$, $\boldsymbol{y}_f^{pred}$, even if the predictive BF is higher for the data generated from the true model $\mathcal{M}_o$, as expected. In the second scenario we have instead an apparent disagreement between the two BFs, since the predictive BF for $\boldsymbol{y}_o^{pred}$ would indicate a clear preference for $\mathcal{M}_o$, whereas $BF_{of}(\boldsymbol{y}_f^{pred})$ definitely suggests the choice of $\mathcal{M}_f$. This apparently controversial finding in the second scenario is not surprising: the predictive BF evaluated for $\boldsymbol{y}_o^{pred}$ is likely to favour the nested and simpler model $\mathcal{M}_o$ over the full and more complex model $\mathcal{M}_f$ due to the well-known *Occam's rasor* issue. In fact, when predictive data $\boldsymbol{y}_o^{pred}$ are generated from the nested model $\mathcal{M}_o$ but the true model is instead the full $\mathcal{M}_f$, $BF_{of}(\boldsymbol{y}_o^{pred})$ will clearly privilege the simpler model, and the posterior predictive distribution for $BF_{of}(\boldsymbol{y}_o^{pred})$ will serve as a sort of upper benchmark of comparison; conversely, $\mathcal{M}_f$ is chosen only when data complexity goes beyond that implied by the simpler model $\mathcal{M}_o$, for instance when $\boldsymbol{y}_f^{pred}$ is generated from the true model $\mathcal{M}_f$, as in the bottom part of the bottom panel.

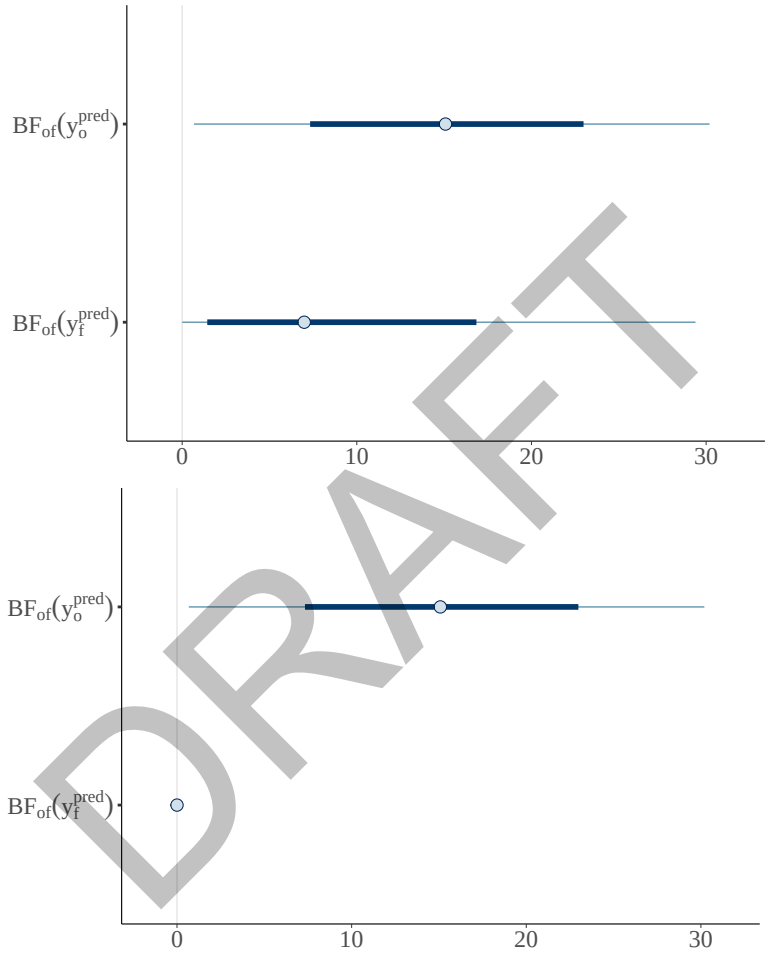
This example helps guide the choice between the two competing nested models: from the whole configuration, it emerges that $\mathcal{M}_o$ is clearly favoured in the first scenario—where both the ppds are far from zero and quite close each other—whereas $\mathcal{M}_f$ is definitely suggested in the second scenario—where the ppd for $BF(\boldsymbol{y}_f^{pred})$ converges at zero—as expected. Moreover, we conclude with a look at the test BF, $BF(\boldsymbol{y}^c)$, over the $T$ splits: in the first scenario the median value is 14.6, whereas in the second scenario is 0. These values show a further confirmatory effect on the choice of the models in the two scenarios.

## 4. Discussion

We proposed a novel predictive protocol to assess model predictive accuracy by evaluating Bayes factors on samples from the posterior predictive distributions of two competing models. In such way, we try to transform BFs into some purely predictive-based tools and make them appealing from a forecasting accuracy perspective.

Many points of future research remain open. We would need to check and prove the consistency of our proposed BFs and the connection with some existing predictive information criteria, such as the BIC, and with leave-one-out cross validation. We should then formalize the final model choice with some *ad-hoc* metrics, possibly based on the Kullback-Leibler divergence.

DRAFT




DRAFT

Figure 1: Linear regression with $g$-priors: predictive Bayes factors $BF_{of}(\boldsymbol{y}_o^{pred})$ and $BF_{of}(\boldsymbol{y}_f^{pred})$ according to two simulated scenarios: $y = 3 + 0.3x + \epsilon$ (top panel); $y = 3 + 0.3x + 0.5x^2 + 0.7x^3 + \epsilon$ (bottom panel).

Moreover, the application of the algorithmic protocol in more complex settings where BFs are not available in closed-forms is required.

We feel our tentative and preliminary procedure is transparently posed and computationally feasible: open-source softwares such as Stan (3), combined with the R package `bridgesampling` (5), may be used for BFs derivation and computation in complex settings.

## References


[1] Berger JO, Pericchi LR (1996) The intrinsic bayes factor for model selection and prediction. Journal of the American Statistical Association 91(433):109–122

[2] Carpenter B (2022) Bayes factors measure prior predictive performance. Statistical Modeling, Causal Inference, and Social Science URL https://statmodeling.stat.columbia.edu/2022/06/25/bayes-factors-measure-prior-predictive-performance/

[3] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. Journal of Statistical Software 76(1)

[4] García-Donato G, Chen MH (2005) Calibrating bayes factor under prior predictive distributions. Statistica Sinica pp 359–380

[5] Gronau QF, Singmann H, Wagenmakers EJ (2020) bridgesampling: An R package for estimating normalizing constants. Journal of Statistical Software 92(10):1–29, DOI 10.18637/jss.v092.i10

[6] Held L, Ott M (2018) On p-values and bayes factors. Annual Review of Statistics and Its Application 5:393–419

[7] Jeffreys H (1998) The theory of probability. OuP Oxford

[8] Kass RE (1993) Bayes factors in practice. Journal of the Royal Statistical Society: Series D (The Statistician) 42(5):551–560

[9] Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90(430):773–795

[10] Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for bayesian variable selection. Journal of the American Statistical Association 103(481):410–423

[11] O'Hagan A (1995) Fractional bayes factors for model comparison. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):99–118

[12] Robert CP (2009) Predictive bayes factors?! Xi'An's Og URL https://xianblog.wordpress.com/2009/09/11/predictive-bayes-factors/

[13] Trotta R (2007) Forecasting the bayes factor of a future observation. Monthly Notices of the Royal Astronomical Society 378(3):819–824

[14] Zellner A (1986) On assessing prior distributions and bayesian regression analysis with g-prior distributions. Bayesian Inference and Decision Techniques


DRAFT

# A Clusterwise Regression Method for Distributional-Valued Data

Antonio Balzanella [a], Rosanna Verde[a], and Francisco de A.T. de Carvalho[b]

[a]Dept. of Mathematics and Physics, University of Campania Luigi Vanvitelli, 81100 Caserta, Italy; antonio.balzanella@unicampania.it, rosanna.verde@unicampania.it
[b]Centro de Informatica, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes s/n - Cidade Universitaria, Recife-PE, Brazil; fatc@cin.ufpe.br

**Abstract**

In this work, we present a clusterwise algorithm based on a new regression method for distributional data. The first contribution of this work is the definition of a new regression model for distributional data, which maps density functions into a Hilbert space through a Logarithmic transformation of the Derivative Quantile functions (LDQ). Then, the proposed clusterwise regression method predicts the response variable by partitioning the set of objects into $K$ clusters, according to the best fit of local regression models. Preliminary results on synthetic data have confirmed the effectiveness of our proposal.

***Keywords:*** Distributional data, Clusterwise regression, clustering

DRAFT

## 1. Introduction

In this work, we present a new regression method and a clusterwise strategy for distributional data. The input consists of distributional variables $Y, X_1, \ldots, X_p$, where $Y$ is the response variable and $X_j$'s are the predictors. Each object is represented by $p+1$ probability functions or empirical ones. Our approach is to predict the response variables by partitioning the set of distributional-valued data into subgroups using a K-means like clustering algorithm. The centroids of the clusters are represented by linear regression models, and objects are assigned to the cluster which minimizes the prediction error. The first contribution of the paper is to define a suitable regression model for distributional data. Previous works, such as (7) and (4), proposed regression models for distributional data based on Non-Linear Least Squared methods and the Wasserstein metric in a linear space, however they imposed the constraint of non-negativity to ensure that the outcome remains a distributional variable. Considering recent developments in distributional data analysis (DDA), we introduce a transformation of the quantile functions into quantile density functions (8), which allows for mapping density functions into a Hilbert space addressing some issues in DDA.

Based on this new regression model, the clusterwise method predicts the response variable by performing a partitioning of the set of objects in $K$ clusters according to the best fit of the local regression models. The clustering process is performed in two alternating steps. Fixed a predefined $K$ number of clusters, the first step involves representing the clusters with regression models; in the second step the elements are assigned to the clusters according to the minimum sum of the squared errors. This process is repeated until convergence to stable clusters.

## 2. Distributional-valued data

Let $E$ be a set of objects. A modal (1; 7) variable $X$ with domain $\mathscr{D}$ on the set $E$ is a mapping $E \to \mathscr{M}$ of all possible measures $\pi$ on $\mathscr{D}$ (completed by a $\sigma$-field):
$e \to X(e) \in \mathscr{M}$, for $e \in E$.

Histograms are a suitable way of representing empirical distributions. In the context of SDA, histogram data are realizations of histogram-valued variables and are considered a special case of modal-valued variables. In this framework, a variable $X$ is considered a histogram-valued variable if each object $e$ is represented by a probability or frequency distribution in the form of a histogram (1).

Formally, let's $x$ be a realization of $X$ with support $S(e) = [a_H, b_H]$, that is partitioned into a set of contiguous intervals (or bins) $I_h = [a_h, b_h]$, such that $I_h \cap I'_h = \emptyset$; $\forall h : I_h \subseteq S(e)$. A non negative weight $\pi_h$ (a probability or a relative frequency) is associated to each $I_h$, so that, a histogram data is defined by a sequence of intervals (bins) with associated the respective weights:

$$x = X(e) = [(I_1, \pi_1), \ldots, (I_h, \pi_h), \ldots, (I_H, \pi_H)] \tag{1}$$

A cumulative distribution function $F(x)$ is associated to each $x$. $F(x)$ is a continuous function and, for empirical distributions, it is a piece-wise function, strictly increasing in the interval $[a_1, b_H]$; $0 \leq F(x) \leq 1$; $F(x) = 0$ $\forall x \leq a_1$ and $F(x) = 1$ $\forall x \geq b_H$; $F(x)$ is differentiable on $[a_1, b_H]$.

The inverse of a distribution function $F^{-1}(x) = Q(t)$ is a quantile function which is a monotone increasing function with support in $[0,1]$ (for histogram variables, it is a piece-wise function). $Q(t)$ is also differentiable on $[0,1]$.

## 2.1 Linear regression model for distributional data

In the framework of distributional-valued data analysis, many statistical analysis methods have been proposed, as clustering methods, regression models, factorial approaches, and many others. A reference book on these themes is (2).

Regression models have been developed to study the dependence relationship between a response variable $Y$ and a set of explanatory variables $X_1, \ldots, X_p$. Both response and independent variables are distributional valued variables. The main regression model approaches (4; 7) are based on a suitable metric to compare distributions: the Wasserstein's squared distance (also known as Mallow's distance). In the unidimensional space, this metric corresponds to the Euclidean distance between the two quantile functions, $Q_i(t)$ and $Q_{i'}(t)$, associated to the two objects $e_i, e_{i'}$ :

$$d_W(e_i, e_{i'}) = \int_0^1 (Q_i(t) - Q_{i'}(t))^2 dt. \tag{2}$$

The squared distance Wasserstein $d_W$ has been assumed as metric between distributional valued data for solving an OLS minimization problem:

$$SSE_{OLS-LDQ} = \left(\tilde{\mathbf{y}}(t) - \tilde{\mathbf{X}}(t)\beta\right)^\top \left(\tilde{\mathbf{y}}(t) - \tilde{\mathbf{X}}(t)\beta\right)$$

Where $\mathbf{y}(t)$ and $\mathbf{X}(t)$ are the matrices of quantile functions associated to the distributional-valued data of $Y$ and $X_1, \ldots, X_p$, while $\beta$ is the vector of the regression coefficients.

The estimation problem is solved differently in the two main regression approaches for distributional-valued data (4; 7), but in both cases, a non-negativity constraint is imposed to the coefficients to ensure that the predicted $\tilde{\mathbf{Y}}(t)$ is a vector of quantile functions.

The main challenge is that probability density functions as well as cumulative distribution functions and their inverse, the quantile functions, are not in a Hilbert space. The solutions introduced in the two previous regression methods only in part address this problem by proposing a symmetrical transformation and a centring transformation of the regression variables.

More recently a suitable transformation has been introduced for mapping probability densities to a Hilbert space of functions through a continuous and invertible map (8). A probability density function is firstly transformed by the derivative of its quantile function:

DRAFT

$$q(t) = \frac{dQ(t)}{dt} = \frac{dF^{-1}(t)}{dt} = \frac{1}{f(Q(t))}$$

which is strictly positive and continuous in its domain $[0,1]$, and then, by the logarithmic transformation of $q(t)$, as follows:

$$l(t) = \ln q(t) = \ln \frac{1}{f(Q(t))} = -\ln f(Q(t))$$

Let $l(t)$ denote the Logarithm Derivative of the Quantile function (LDQ). On these functions it is possible to define the addition and scalar multiplication operations, the inner product and the Euclidean norm. The squared Euclidean distance between two LDQ functions $l_1(t)$ and $l_2(t)$ is given by:

$$d^2_{LDQ}(l_1(t), l_2(t)) = \|l_1(t) - l_2(t)\|_2 = \int_0^1 (l_1(t) - l_2(t))^2 \, dt$$

The main problem with LDQ transformation is that it loses information on the location parameters of the density distribution. In fact, the derivatives of two quantile functions that differ by a constant term are equal.

A linear model on the LDQ transformation functions of $y_i(t)$ and $x_{ij}$(t) was proposed in (10). However, it does not effectively solve the problem concerning the position of the density functions. The new regression model, here proposed, attempts to overcome this issue by dividing the model into two regression models, as follows:

$$y_i^m = \beta_0^m + \sum_{j=1}^{p} \beta_j^m x_{i,j}^m + \varepsilon_i^m$$

$$y_i^l(t) = \beta_0^l(t) + \sum_{j=1}^{p} \beta_j^l x_{i,j}^l(t) + \varepsilon_i^l(t)$$

where:

$y_i^m$ is related to the minimum values $x_{ij}^m$;
$y_i^l(t)$ and $x_{i,j}^l(t)$ are the LQD function of $y_i(t)$ and $x_{i,j}(t)$;
$\varepsilon_i^m$ is the $i$-th residual error of the minimum values;
$\varepsilon_i^l(t)$ is the $i$-th residual error function of the LDQ function.

The parameters of the models are estimated by minimizing the sum of the square errors, given by:

$$SSE_{OLS-LDQ} = \int_0^1 \left[ \left( \tilde{\mathbf{y}}^l(t) - \tilde{\mathbf{X}}^l(t)\beta^l \right)^\top \left( \tilde{\mathbf{y}}^l(t) - \tilde{\mathbf{X}}^l(t)\beta^l \right) \right] dt + \\ + (\mathbf{y}^m - \mathbf{X}^m\beta^m)^\top (\mathbf{y}^m - \mathbf{X}^m\beta^m) \quad (3)$$

The least squares estimators for $\beta^l(t)$ and $\beta^m$ represent the solutions of the two independent systems.

## 3. The CRL-LDQ Clusterwise Regression Method

The CRL-LDQ method combines the dynamic clustering algorithm (6) with the OLS-LDQ regression method for distributional-valued data. For a fixed number $K$ of clusters, it seeks the better partition $P_k = C_1, \ldots, C_K$ and the best fitting models $\hat{y}^k$, for each cluster $C_k$, by minimising the $SSE_{OLS-LDQ}(P_k, \hat{y}^k)$:

$$SSE_{OLS-LDQ}(P_k, \hat{y}^k | \beta^l_{j(k)}, \beta^m_{j(k)}) = \sum_{k=1}^{K} \sum_{e_i \in P_k} \left[ \|\tilde{\varepsilon}^l_{i(k)}(t)\|^2 + (\varepsilon^m_{i(k)})^2 \right]$$

After setting the number K of clusters, the algorithm is performed alternating the following two steps, until the convergence to a stationary value.

DRAFT

- *Step 1 - Representation step* (best fitting):
  The local regression models are estimated by minimizing the objective function $SSE(P_k, \hat{y}^{k*})$ on the parameters $\beta^l_{(k)}(t)$ and $\beta^m_{(k)}$ $(1 \leq k \leq K)$.
  The OLS estimations are provided as solutions of the two independent systems.

- *Step 2 - Assignment step* (partitioning $P_k$):
  The optimal clusters $P_k$ which minimize the criterion $SSE(P^*_k, \hat{y}^k)$, are obtained according to the following assignment rule:

$$P_k = \left\{ e_i \in E : \left[ \|\hat{\varepsilon}^l_{i(k)}(t)\|^2 + \left(\hat{\varepsilon}^m_{i(k)}\right)^2 \right] = \min_{h=1}^{K} \left[ \|\hat{\varepsilon}^l_{i(h)}(t)\|^2 + \left(\hat{\varepsilon}^m_{i(h)}\right)^2 \right] \right\}$$

  Thus, the observation $e_i$ is assigned to cluster $P_k$ if the sum-of-squared errors are minimal for this cluster regression model.

The convergence of the algorithm is guaranteed by the decreasing of the criterion according to the improvement of the cluster regression models fitting.

## 4. Preliminary results on simulated data

The proposed method's efficacy was tested using simulated data. The performance of the regression model was evaluated by measuring its ability to fit the data, which was determined by using values of $K$ ranging from 1 to 4. A value of $K = 1$ represents the use of the model without partitioning the data into clusters. The simulated data consisted of 600 individuals and included one response variable and two predictors, which were generated from Gamma random variables with varying parameters. The response was a combination of the predictors with added Gaussian error. A pseudo $R^2$ index was calculated using a suitable decomposition of the deviance, and the result was 0.59 for $K = 1$. When the clusterwise procedure was introduced, the results improved to 0.61, 0.74, and 0.95 for $K = 3$, which corresponded to the number of clusters in the data.

DRAFT

## References


[1] Bock H., Diday E.: Analysis of Symbolic Data. Springer Berlin, Heidelberg (2012)
[2] Brito P., Dias S.: Analysis of Distributional Data. Chapman Hall (2022)
[3] de Carvalho, F.d.A., Saporta, G., Queiroz, D.N. A: Clusterwise Center and Range Regression Model for Interval-Valued Data. In: Lechevallier, Y., Saporta, G. (eds) Proceedings of COMPSTAT'2010. Physica-Verlag HD. (2010)
[4] Dias, S. and Brito, P.: Linear regression model with histogram-valued variables. Statistical Analy Data Mining, 8: 75-113 (2015) https://doi.org/10.1002/sam.11260
[5] Diday E.: Introduction à l'analyse factorielle typologique, Revue de Statistique Appliquée, XXII(4), 29-38, ( 1974)
[6] Diday E.,Simon J.C.: Clustering analysis. Digital Pattern Recognition, 47-94 Springer (1980)
[7] Irpino, A., Verde, R.: Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. Adv Data Anal Classif 9, 81?106 (2015). https://doi.org/10.1007/s11634-015-0197-7
[8] Petersen A., Müller H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics, Ann. Statist. 44(1), 183-218, (2016)
[9] Spaeth, H.: Clusterwise Linear Regression. Computing 22 (4), 367?373 (1979)
[10] Zhao Q., Wang H., Lu S.: M-LDQ feature embedding and regression modeling for distribution-valued data. Information Sciences, Volume 609 (2022). https://doi.org/10.1016/j.ins.2022.07.064.

# A novel statistical-significance based semi-parametric GLMM for clustering countries standing on their innumeracy levels

Alessandra Ragni[a], Chiara Masci[a], Francesca Ieva[a,b], and Anna Maria Paganoni[a]

[a]MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy; alessandra.ragni@polimi.it, chiara.masci@polimi.it, francesca.ieva@polimi.it, anna.paganoni@polimi.it
[b]CHDS (Center for Health Data Science), Human Technopole, Milan, 20157, Italy.

DRAFT

**Abstract**

We present a semi-parametric linear mixed-effects model for a generalized response which assumes the random effects to follow a discrete distribution. The estimated support points of the discrete distribution are collapsed along the iterations of a tailored EM algorithm - through which the parameters of the model are estimated - inducing a clustering of the higher level of hierarchy. The novelty lies in the collapsing criteria: the two closest estimated support points for which the confidence regions overlap are collapsed. At convergence, an a priori unknown optimal number of statistically different support points is identified. Moreover, the model is applied to the Programme for International Student Assessment (OECD) schools' data for modelling their innumeracy rates considering the school-country nesting.

***Keywords:*** Semi-parametric generalized linear mixed-effects model, Discrete random effects, EM algorithm, Innumeracy rates, Administrative databases

## 1. Introduction

We present a novel method based on statistical significance for clustering groups in Semi-Parametric Generalized Linear Mixed-effects Models (SPGLMMs). Hierarchical data, often encountered in longitudinal studies and repeated measurements, necessitate the use of specialized models to account for both the group and individual levels. Mixed effects models (1) are a common choice, being able to handle both random and fixed effects. Recently, Semi-Parametric Linear Mixed-effects Models (SPLMMs) have been proposed in literature for continuous (2), multinomial (3) and Bernoulli (4) responses. Such a family of models assume the random effects to follow a discrete distribution and make use of a tailored Expectation-Maximization (EM) algorithm (5) for the estimation of the parameters. This approach offers several advantages, such as dimensionality reduction and improved interpretability; it reveals a clustering structure of the hierarchy, a priori latent, being more flexible than the parametric version as it does not require the assumption of normal distribution. However, the main drawback of the state-of-the-art SPLMMs is the collapsing criterion, performed at each iteration of the algorithm, being dependent on the choice of a threshold that determines the merging of discrete masses with lower Euclidean distance.

The presented method involves the computation of confidence regions of a certain level of confidence centred on the two closest support points estimated using Maximum Likelihood Estimators (MLEs) and their asymptotic properties, as described in (6). The overlapping of the confidence regions leads to the merging of the two discrete masses. The advantage of this criterion lies in the identification of the latent structure solely by choosing a level of confidence rather than an arbitrary threshold.

The SPGLMM, formulated for a Poisson response, is applied to data extracted from the Programme for International Student Assessment (PISA) (7) to cluster countries based on their levels of mathematical illiteracy (e.g., innumeracy levels). Specifically, we aim at profiling the percentage of low-achieving students in mathematics within schools standing on the average size of the school and socio-economic index of the students within that school. We are interested in identifying latent clusters of countries. The results show that the proposed method can effectively identify statistically different clusters and provide more interpretable solutions compared to traditional methods.

## 2. The model formulation and the EM algorithm

Given $i = 1, ..., N$ groups, Generalized Linear Mixed-effects Models (GLMMs) (8) are defined such that, conditioned on the random effects $\mathbf{b}_i$ in the $i^{th}$ group, the expectation of the dependent variable $\mathbf{y}_i$ ($n_i$-dimensional vector), being $\mathbf{y}_i$ distributed according to the exponential family, is related to the linear predictor $\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$ via a link function $g(\cdot)$, as follows: $g(\mathbb{E}[\mathbf{y}_i|\mathbf{b}_i]) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$ for $i = 1, ..., N$ where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are, respectively, the $[n_i \times P]$ and $[n_i \times Q]$ matrices of fixed and random covariates in the $i^{th}$ group; $\beta$ is the $P$-dimensional vector of fixed coefficients and $\mathbf{b}_i$ the $Q$-dimensional vector of random coefficients in the $i^{th}$ group, being $\mathbf{b}$ normally distributed within the parametric framework.

In the semi-parametric approach, $\mathbf{b}$ is assumed to follow a discrete distribution $P$, with an a priori unknown $M$ support points $(\mathbf{c}_1, ..., \mathbf{c}_M)$ for $M \leq N$, where each $\mathbf{c}_m \in \mathbb{R}^Q$, $m = 1, ..., M$, corresponds to the $m^{th}$ cluster. Each group $i$ is assigned with a certain probability $\omega_m$ to a cluster $m$ allowing the identification of a latent structure among the groups, in the following fashion:

$$\mathbf{b}_i = \begin{cases} \mathbf{c}_1 & p(\mathbf{b}_i = \mathbf{c}_1) = \omega_1 \\ ... & \\ \mathbf{c}_M & p(\mathbf{b}_i = \mathbf{c}_M) = \omega_M \end{cases} \qquad \sum_{m=1}^{M} \omega_l = 1, \ \omega_m \geq 0 \qquad\qquad \forall i \in N$$

so that the SPGLMM formulation is

$$g(\,\mathbb{E}[y_{ij}|\mathbf{c}_m]\,) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{c}_m \quad \text{for} \quad i = 1, ..., N, \ j = 1, ..., n_i, \ m = 1, ..., M$$

where $\mathbf{x}_{ij}$ is the $[P \times 1]$ vector of fixed covariates and $\mathbf{z}_{ij}$ the $[Q \times 1]$ vector of random covariates. The marginal loglikelihood can be expressed, as proposed in (9), as follows:

$$\ln \mathcal{L}(\boldsymbol{\beta}, \mathbf{c}_1, ..., \mathbf{c}_M|\mathbf{y}) = \sum_{m=1}^{M} \omega_m \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ln\, p(y_{ij}|\boldsymbol{\beta}, \mathbf{c}_m)$$

being $p(y_{ij}|\boldsymbol{\beta}, \mathbf{c}_m)$ the conditional probability mass (or density) function of $y_{ij}$ given random and fixed effects.

The tailored EM algorithm is inspired by the one proposed in (10) and consists of the computation of the expected log-likelihood and its maximization with respect first to random effects and then to fixed effects, in an iterative framework, as described in (4; 6). The pointwise estimates updates of the unknown parameters $\hat{\boldsymbol{\beta}}$, $(\hat{\mathbf{c}}_1, ..., \hat{\mathbf{c}}_M)$ and $(\hat{\omega}_1, ..., \hat{\omega}_M)$ are obtained while proving the increasing likelihood property given a fixed number of clusters $M$. Refer to (6) for a more detailed algorithm formulation. Furthermore, the point estimate $\hat{\mathbf{b}}_i$ of the coefficients $\mathbf{b}_i$ for each group $i = 1, ..., N$, is obtained by maximizing over $m$ the conditional probability $p(\mathbf{b}_i = \hat{\mathbf{c}}_m|\mathbf{y}_i, \hat{\boldsymbol{\beta}})$.

DRAFT

## 3. The statistical significance-based support reduction criterion

As foreseen in Section 1, the state-of-the-art methods (10; 2; 3; 4) reduce support through a check performed at each iteration $k$ of the EM algorithm. This check entails merging two points, $\hat{\mathbf{c}}_l^{(k)}$ and $\hat{\mathbf{c}}_m^{(k)}$, into a single point $\hat{\mathbf{c}}_{l,m}^{(k)} = \frac{\hat{\omega}_l^{(k)}\hat{\mathbf{c}}_l^{(k)} + \hat{\omega}_m^{(k)}\hat{\mathbf{c}}_m^{(k)}}{\hat{\omega}_l^{(k)} + \hat{\omega}_m^{(k)}}$ with weight $\hat{\omega}_{l,m}^{(k)} = \hat{\omega}_l^{(k)} + \hat{\omega}_m^{(k)}$, if their Euclidean distance is lower than a fixed threshold $t$. In real data applications, these methods could result in being computationally expensive because requiring multiple fits of the model with different $t$ for the identification of the best-performing one[1].

In view of untying from the choice of $t$, as explained in (6) we propose to (i) to compute the confidence regions (intervals) of level $1-\alpha$ centered in each of the two closest - in terms of Euclidean distance - estimated support points, exploiting the asymptotic properties of the Maximum Likelihood Estimators (MLEs) and (ii) if the two confidence regions (intervals) overlap, to collapse the two discrete masses to a unique point $\hat{\mathbf{c}}_{l,m}^{(k)}$ with weight $\hat{\omega}_{l,m}^{(k)}$ as in the state-of-the-art methods; if not, the *overlapping condition* is checked for all the other pairs of mass points (ordered by increasing Euclidean distance) until either two confidence regions (intervals) overlap or all pairs have been checked. From a practical point of view, the variance-covariance matrix is computed at each iteration of the algorithm as the inverse of the Information Matrix, retrieved by means of finite differences approximation of the MLE.

Moreover, the *overlapping condition* for $Q = 1$ reduces to a simple inequality which checks the reciprocal position of the extremes of the two confidence intervals, while for $Q \geq 2$ reduces to a unidimensional minimization[2]. As a final check, empty clusters (i.e. the estimated support points to which no groups are associated) or clusters with zero weight are removed; when one or more mass points are deleted, the remaining weights are renormalized in such a way that they sum up to 1.

DRAFT

## 4. The innumeracy levels within schools across countries

We apply the SPGLMM with the support reduction criterion - described in Section 3. - to data concerning mathematical performances PISA (OECD) survay of 2018 (7). The global indicators for the United Nations Sustainable Development Goals identify a minimum Level of Proficiency that all children should acquire by the end of secondary education: students below this level are considered *low-achieving* students. We restrict the analysis to schools with at least 10 students. After removing missing values and aggregating student data at the school level, we are left with 12620 schools across 50 countries. We consider, for each school $j$ within a country $i$, the Poisson distributed response variable $\texttt{Y_MATH}_{ij}$, (i.e., the rounded percentage of low-achieving students) and the two independent variables at school level (i) $\texttt{avg_ESCS_std}_{ij}$ (i.e., the average students' index of economic, social and cultural status, subsequent to standardization to mean 0 and standard deviation 1 within the country of the school, for keeping into account differences between countries) and (ii) $\texttt{SCHSIZE}_{ij}$ (i.e., the sum of students of school $j$), both standardized to mean zero and standard deviation 1. The SPGLMM with Poisson response is

$$\ln(\,\mathbb{E}[\texttt{Y_MATH}_{ij}|c_m]\,) = \mathbf{x}_{ij}'\boldsymbol{\beta} + c_m$$

for $i = 1, ..., 50,\ j = 1, ..., n_i,\ m = 1, ..., M_\alpha$ where $\mathbf{x}_{ij}'$ is the two-dimensional vector of fixed effects covariates at the school level that contains $\texttt{SCHSIZE}_{ij}$ and $\texttt{avg_ESCS_std}_{ij}$; $\boldsymbol{\beta} = [\beta_1, \beta_2]'$ is the vector of fixed effects coefficients; $c_m$ is the random intercept relative to the $m^{\text{th}}$ cluster of countries and $M_\alpha$ is the total number of clusters the model identifies and depends on the level of confidence $\alpha$ chosen. We run the model with different $\alpha$ and we get $\hat{M}_{0.01} = 13$, $\hat{M}_{0.05} = 16$ and $\hat{M}_{0.10} = 18$. As expected, higher values of $\alpha$ correspond to higher values of $M$. Indeed, the higher $\alpha$, the smaller the confidence intervals and the less likely to overlap. By choosing $\alpha = 0.01$, we obtain $\hat{\beta}_1 = -1.361$ and $\hat{\beta}_2 = -0.094$

[1] See S6 of Supplementary Materials in (6) for the description of an *entropy*-based method for the selection of the best $t$.

[2] The *Fast Ellipsoid Intersection Test*, see https://github.com/NickAlger/nalger_helper_functions/blob/master/tutorial_notebooks/ellipsoid_intersection_test_tutorial.ipynb for details concerning the implementation.

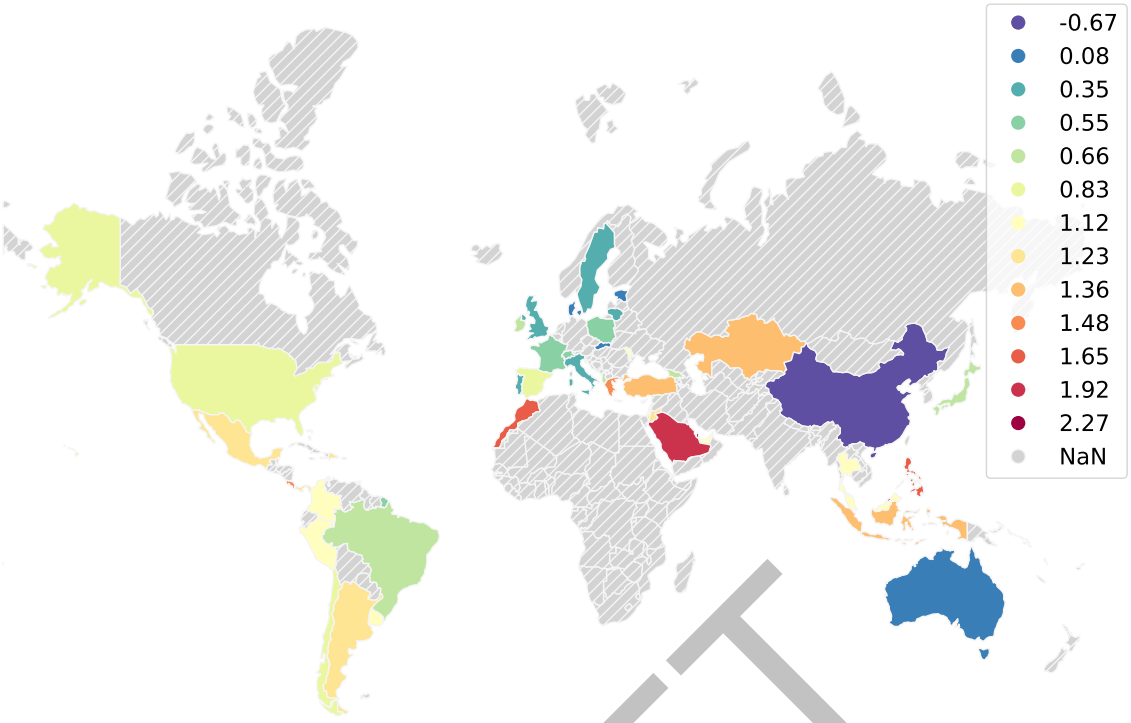



Figure 1: Choropleth map of the clusters of countries identified by the random intercepts for $\alpha$ = 0.01. Countries represented with the same colour belong to the same cluster. The blue-most the colour, the lower the random intercept. Grey-striped countries are missing data.

DRAFT

(having both a statistically significant p-values[3]) and the obtained random intercepts are plotted in Fig. 1. For this specific application, the running time for fitting the SPGLMM with a fixed $\alpha$ is a few hours and the obtained results are stable among different runs. Such an amount of time is mainly due to the iterative nature of the algorithm, whose pseudo-code is reported on page 11 in (6). The running time is much higher than fitting a parametric GLMM, which has the drawback of not being able to identify a clustering structure, but much lower than the SPGLMM based on the choice of the threshold $t$, because it does not require tuning $t$ by fitting multiple models: we can rely on the conventional meaning attributed to the statistical confidence levels.

## 5. Simulation study

In order to evaluate the performance of the proposed model and the reliability of the results in Section 4, we propose a simulation study to test the SPGLMM considering a Poisson-distributed response with a single random intercept and two fixed-effects covariates, to be in line with Section 4. We consider $N = 10$ groups of data, where each group contains $n_{i=1,\dots,N} \sim U(70, 100)$ and we induce the presence of three clusters. The linear predictor $\boldsymbol{\eta}_i = \boldsymbol{\beta}_1 \mathbf{x}_{1i} + \boldsymbol{\beta}_2 \mathbf{x}_{2i} + c_{1i} \mathbb{1}_{n_i}$ is defined by the following Data Generating Process (DGP):

$$\boldsymbol{\eta}_i = \begin{cases} 0.3\mathbf{x}_{1i} + 0.9\mathbf{x}_{2i} + 2.5\ \mathbb{1}_{n_i} & \text{if } i = 1, 2, \\ 0.3\mathbf{x}_{1i} + 0.9\mathbf{x}_{2i} + 1\ \mathbb{1}_{n_i} & \text{if } i = 3, 4, 5, 6, 7, \\ 0.3\mathbf{x}_{1i} + 0.9\mathbf{x}_{2i} - 1\ \mathbb{1}_{n_i} & \text{if } i = 8, 9, 10 \end{cases} \tag{1}$$

[3]Computed through likelihood-ratio test.

Variables $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ follow a distribution $\mathcal{N}(0,1)$. We perform 500 runs of the SPGLMM. Results obtained by using Section 3.'s criterion are shown in Table 1: estimates are reported in terms of mean (sd) on the runs in which the SPGLMM identifies 3 clusters. If the DGP in Eq. (1) is fitted by using the state-of-the-art collapsing criterion (4), the proportion (over 500 iterations) in which 3 clusters are identified (which is maximized for $t = 0.25$) is 0.922, a value that needs to be compared with 0.956 obtained in the best case ($\alpha = 0.01$) and 0.870 obtained in the worst case ($\alpha = 0.10$).

Table 1: Results obtained by SPGLMM for the Poisson response through DGP in Eq. (1).

| $\alpha$ | Proportion (over 500 iterations) in which 3 clusters are identified | $\hat{\omega}$ (sd) | $\hat{c}_1$ (sd) | $\hat{\beta}_1$ (sd) | $\hat{\beta}_2$ (sd) |
|---|---|---|---|---|---|
| $\alpha = 0.01$ | 0.956 | 0.20 (0.00)<br>0.50 (0.00)<br>0.30 (0.00) | 2.50 (0.02)<br>1.00 (0.03)<br>-1.01 (0.08) | 0.30 (0.01) | 0.90 (0.02) |
| $\alpha = 0.05$ | 0.934 | 0.20 (0.00)<br>0.50 (0.00)<br>0.30 (0.00) | 2.50 (0.02)<br>1.00 (0.03)<br>-1.01 (0.08) | 0.30 (0.01) | 0.90 (0.02) |
| $\alpha = 0.10$ | 0.870 | 0.20 (0.21)<br>0.50 (0.31)<br>0.30 (0.07) | 2.49 (0.12)<br>0.99 (0.12)<br>-0.99 (0.09) | 0.30 (0.01) | 0.90 (0.02) |

## 6. Conclusions

We presented a statistical significance-based method for the estimated support points reduction within SPGLMMs, and, more in general, any kind of SPLMMs (2; 3; 4). By choosing a level of confidence $\alpha$, such criterion allows the identification of a latent structure by clustering groups while having no prior knowledge of the "true" number of clusters. Simulation studies proved that the approach outperforms the state-of-the-art approaches, based on the choice of a discretionary threshold, which selection can often be computationally expensive. We applied the criterion to PISA (OECD) for obtaining a clustering of the countries as an alternative to the ranking proposed by a classical parametric GLMM. Yet, the method can be applied in various contexts, such as healthcare or any scenario where clustering groups is a more desirable approach than simply dealing with individual groups.

DRAFT

## References

[1] Pinheiro, J. C., and Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus, 3-56.

[2] Masci, C., Paganoni, A. M. and Ieva, F. (2019). Semiparametric mixed effects models for unsupervised classification of italian schools. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4):1313–1342, 2019.

[3] Masci, C., Ieva, F., and Paganoni, A. M. (2022). Semiparametric multinomial mixed-effects models: A university students profiling tool. The Annals of Applied Statistics, 16(3), 1608-1632.

[4] Ragni, A., Masci, C., Ieva, F., and Paganoni, A. M. (2022). Semi-parametric generalized linear mixed effects models for binary response for the analysis of heart failure hospitalizations. In Proceedings of the 51th Scientific Meeting of the Italian Statistical Society (pp. 2042-2047).

[5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

[6] Ragni, A., Masci, C., Ieva, F., and Paganoni, A.M. (2023). Clustering Hierarchies via a Semi-

Parametric Generalized Linear Mixed Model: a statistical significance-based approach. arXiv preprint arXiv:2302.12103 (2023).
[7] OECD. (2019). PISA 2018 results (Volume I, II, and III): Combined executive summary.
[8] Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American statistical Association, 88(421), 9-25.
[9] Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics, 55(1), 117-128.
[10] Azzimonti, L., Ieva, F., and Maria Paganoni, A. (2013). Nonlinear nonparametric mixed-effects models for unsupervised classification. Computational Statistics, 28, 1549-1570.

DRAFT

# Introducing a novel directional distribution depth function for supervised classification

Edoardo Redivo[a] and Cinzia Viroli[a]

[a]University of Bologna; edoardo.redivo@unibo.it, cinzia.viroli@unibo.it

**Abstract**

Statistical depth functions are introduced as a way to provide a center-outward ordering of the sample points in multidimensional space, which can be used for outlier detection, classification, and other exploratory tools. In this work we propose a novel definition of depth function for multivariate data using random directions, which preserves the Mahalanobis distance of the points in the original space. The proposed method is evaluated through simulated experiments and real data applications, and is shown to be effective in supervised classification problems.

***Keywords:*** depth functions, random projection, supervised classification

DRAFT

## 1. Introduction

In multivariate analysis the identification of order statistics, quantiles and typical or atypical patterns is very challenging due to the lack of an order among observations, which is instead natural in the real line $\mathbb{R}^1$ (3; 7).

To this aim, the most important line of research is rooted in the concept of statistical depth, which leads to a natural center-outward ordering of the sample points in $\mathbb{R}^p$ with $p \geq 2$. More specifically, a depth function is a function that assigns a real number to each point of a multivariate dataset measuring the outlyingness of the point with respect to the barycenter. Thus, it provides a way to quantify how far an observation is from the center of the dataset, and is often used in outlier detection, clustering, classification.

Popular depth functions are the Mahalanobis depth, which is based on the Mahalanobis distance between a point and the center of the dataset and the halfspace depth, which measures the depth of a point as the probability that it lies inside a randomly chosen halfspace that contains the center of the dataset. (4) described different depth functions as valuable exploratory tools in multivariate analysis.

In this paper we propose a novel definition of depth function for multivariate data using random spherical directions. The method is defined to satisfy the essential properties of depth functions (8; 7) while also preserving the Mahalanobis distance of the multivariate points in the original space of dimensionality $p$. The proposed depth function is the expectation of all depths along the potentially infinite random directions, which are, in turn, functions of the point percentiles estimated via parametric or nonparametric models. The proposed strategy is very general thanks to the model choices in the projected spaces, with the flexibility of the *fgld* quantile distribution being a valuable option (6; 1). Additionally, the theoretical definition of the depth function allows prediction out-of-sample. This enables the method to be used as a tool in supervised classification problems, where it can provide a distance measure for points in the test set with respect to the distributions of the different classes. The performance of the proposed method is evaluated through simulated experiments and real data application, and is shown to be very good, especially when compared to other classification methods.

## 2. Directional Distribution Depth

Let $X$ be a multivariate random variable of order $p$ with a probability distribution $F$. A data depth measures how deep (or central) a given value $x$ of $X$ is with respect to the data cloud or a given distribution function. A simple example is the Mahalanobis depth (5), which is inversely proportional to the Mahalanobis distance:

$$MD(x) = \left[1 + (x-\mu)\Sigma^{-1}(x-\mu)\right]^{-1},$$

where $\mu$ and $\Sigma$ are the mean vector and dispersion matrix of the theoretical distribution of $X$, that can be estimated by the data obtaining the sample version of $MD(x)$. (8) introduced a general definition of depth function and defined four essential properties a depth function should have. More precisely, a depth function is a nonnegative and bounded function, which is: (i) invariant to the coordinate system or to the scale of the underlying measurements (affine invariance); (ii) maximum at its center; (iii) monotonically decreasing when a point moves away from the deepest central point and (iv) it should approach zero as a point approaches infinity. In addition, a depth function should be consistent, as the sample size increases, and computationally efficient, *i.e.*, it should be possible to compute the depth values of data points efficiently even for large $p$.

Here we introduce a novel notion of depth function that satisfies the previous properties under the constraint of sphered data and, in addition, it has two additional desirable features. It is general enough to comprise in its definition arbitrary probabilistic formulations and it preserves the Mahalanobis distances of the values to the data barycenter (as the $MD(x)$ naturally does).

Let $S$ be a random vector of length $p$ with a uniform distribution on the sphere. In other terms, a random value of $S$ is a unit-norm direction $s \in \mathbb{S}^{p-1}$ with $\|s\| = 1$. At each random projection, the depth can be defined an inverse function of the percentile of the univariate value $s^\top x$. More specifically, for a whitened random variable the projection depth is the mapping $\mathbb{R}^p \times \mathcal{F} \to [0,1]$ defined as

$$D(x,F) = E_S\left[1 - 2|F_{S^\top X}(S^\top x) - 0.5|\right], \tag{1}$$

where $E_S$ is the expectation with respect to the random vector $S$, $F$ is the probability distribution of the multivariate data and $F_{S^\top X}$ is the marginal probability distribution of the transformation $S^\top X$ evaluated at $S^\top x$. $F_{S^\top X}$ can be any (probabilistic or nonparametric) univariate distribution function differently parameterized along each direction. In this work we will focus and compare the depth based on the Gaussian distribution, on the *fgld* quantile function due to its large flexibility (6; 1) and the nonparametric kernel density estimation. Figure 1 shows the depth contourns obtained by applying the proposed depth with three distributions on data generated from a standard Gaussian and from a Chi-squared distribution with 3 degrees of freedom.

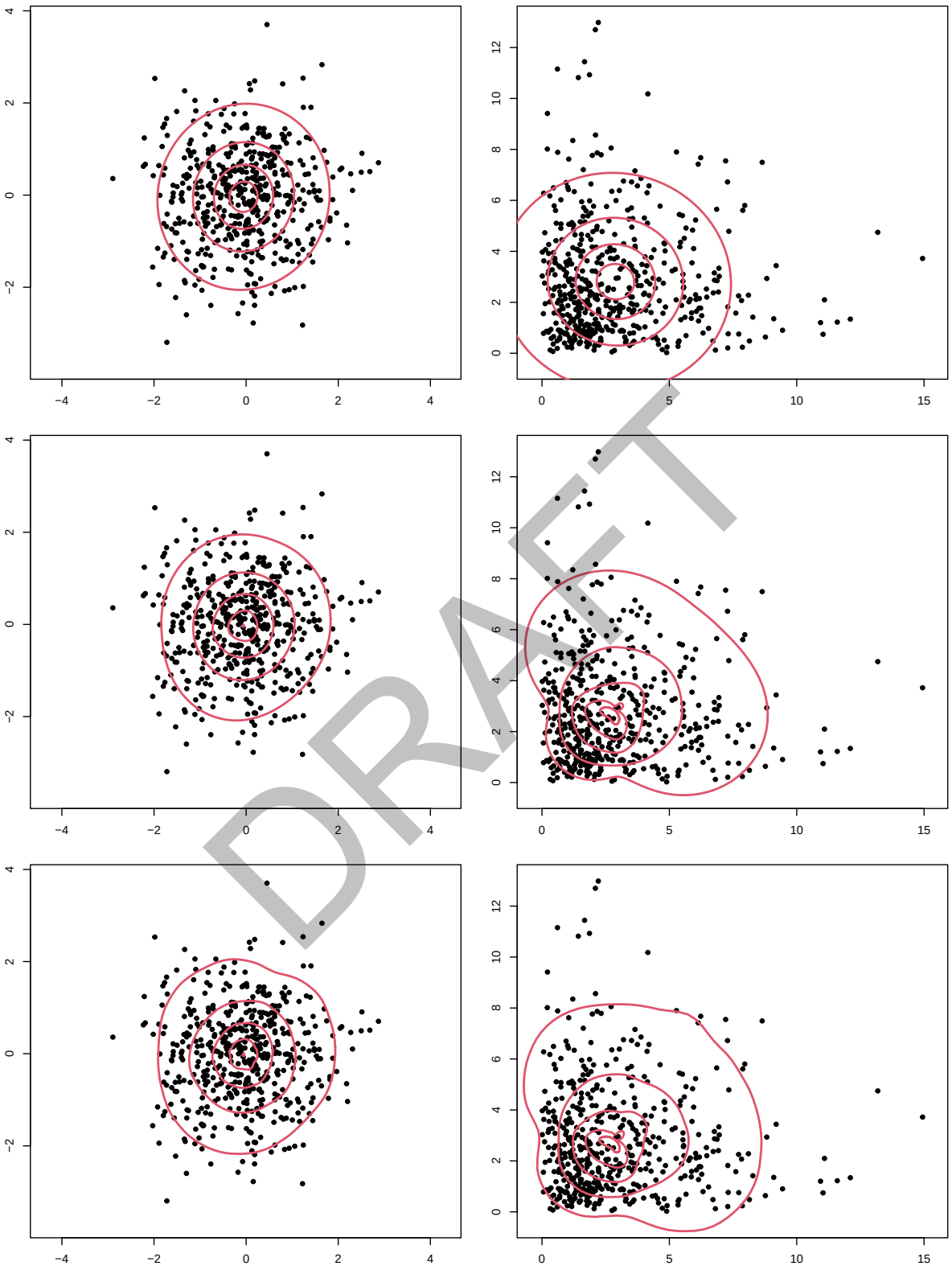
Given whatever model choice of $F_S$, the depth defined in (1) is a proper depth function in the sense of the definition given by (8). Another interesting property of the proposed depth function is that the expected value of the distances between univariate projections on sphered data is proportional to the Mahalanobis distance in the original multivariate data. This is very important from the empirical point of view because it ensures coherence to the depths. Take, for example, the point cloud represented in the first panel of plot 2. The red point with coordinates {2,1} is clearly further away from the point cloud than the green point with coordinates {3.5,4.5}. Their Mahalanobis distances are 6.4 and 2.6, respectively. Table 1 shows the depths obtained on sphered and non-sphered data.

Table 1: Computed depths for the two points on raw data and on sphered data.

| | $\mathbf{x}_a = \{2,1\}$ | $\mathbf{x}_b = \{3.5,4.5\}$ |
|---|---|---|
| without sphering | 0.193 | 0.085 |
| with spering | 0.082 | 0.206 |

From this example, it is clear that on raw data the green point $\mathbf{x}_a$ is projected far from the barycenter of the data in a limited number of directions, while $\mathbf{x}_a$ is projected far from the barycenter for most

DRAFT




DRAFT

Figure 1: Normal, fgld and nonparametric contourns for p=0.25, 0.5, 0.75 and 0.95 on data drawn from a standard Gaussian and a Chi-squared distribution with 3 degrees of freedom.

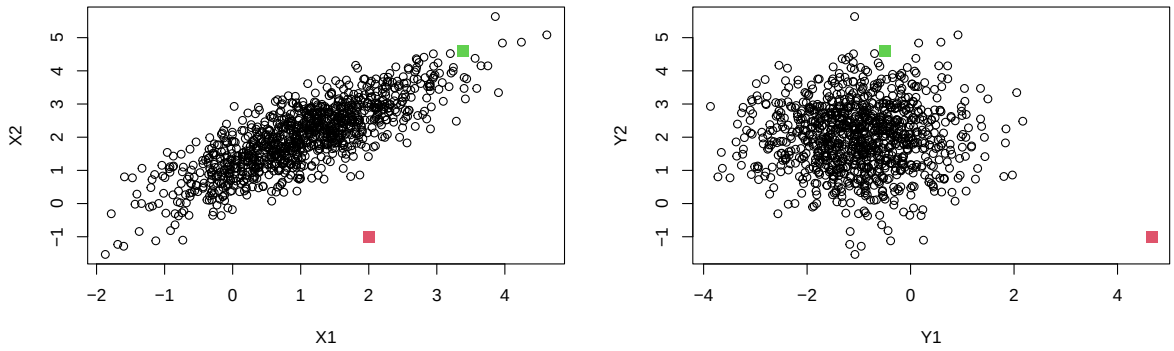



Figure 2: First panel: data drawn from a multivariate Gaussian. Second panel: sphered data.

directions. But sphering makes angles and distances constant along the different projections, and on sphered data the depth indicated that $\mathbf{x}_b$ is deeper than $\mathbf{x}_a$. This is also evident from the second panel of Figure 2.

## 2.1 Sample version

Let $X_n$ be a sample of size $n$ from $X$, without loss of generality we assume it to sphered. Let $\mathbf{s}$ be a set of $B$ spherical directions. Then the sample version of the directional distribution depth for a generic point $x_i$ is

$$D_n(x_i, F) = \frac{\sum_{b=1}^{B} \left[1 - 2|\hat{F}_{s_b^\top X_n}(s_b^\top x_i) - 0.5|\right]}{B}, \tag{2}$$

Please note that unit norm projections of sphered data produce marginal distributions with the same variability. In fact, the marginal distribution along each direction has unit variance. This result is advantageous from a computational point of view, for (at least) two reasons. It simplifies the estimation problem in cases where a parameter represents data dispersion and makes the distributions comparable along different directions.
In the following result we prove the strong consistency of $D_n(x, F)$.

**Theorem 1.** *As $n \to \infty$ and $B \to \infty$*

$$D_n(x, F) \xrightarrow{a.s.} D(x, F) \tag{3}$$

*Proof.* By the strong law of large numbers we first observe that

$$D_n(x, F) \xrightarrow{a.s.} E_S[1 - 2|\hat{F}_{S^\top X_n}(S^\top x) - 0.5|]$$

as $B \to \infty$. Now, let $\Psi_n(x, F) = 1 - 2|\hat{F}_{S^\top X_n}(S^\top x) - 0.5|$ and $\Psi(x, F) = 1 - 2|F_{S^\top X}(S^\top x) - 0.5|$. Notice that $\Psi_n(x, F) = O_p(1)$. Thus by the theorem of dominate convergence

$$E[\Psi_n(x, F)] \to E[\Psi(x, F)]$$

as $n \to \infty$.

**Remark 1.** *Direction importance.*

DRAFT

One might wonder if the directions $s$ could contribute differently to determining the depth of a point. Therefore, the problem arises as to whether it makes sense to weigh the directions by transforming formula (1) into a weighted average using coefficients $w_b$ to be determined according to some criterion. Intuitively, a direction is good when it is able to concentrate the units more around the barycenter. However, as previously observed, the variability along each direction is constant. Furthermore, for each direction, the sum of the depths of all points is also constant. In particular, given $d_{ib} = 1 - 2|\hat{F}_{s_b^\top X_n}(s_b^\top x_i) - 0.5|$, we have

$$\sum_{i=1}^{n} d_{ib} = \frac{n}{2},$$

for every $b$. Therefore, surprisingly, for the purpose of determining the depth, the directions have the same importance.

**Remark 2.** *Prediction.*

The proposed depth is based on a model (parametric or non-parametric) given by $F_S$ and thus it allows for a population version. This is a great advantage in terms of prediction. In other words, unlike other depth functions are defined only for empirical distributions, it is possible to estimate the depth for new out-of-sample values once the distributions along each direction have been estimated. The ability to make prediction makes this depth function a candidate tool for supervised classification purposes, as will be shown in the next sections.

## 3. Application to Supervised Classification

We apply to proposed depth function to supervised classification by allocating a new observation to the class with the maximum depth among the $K$ populations (2). Thus for a statistical unit $x$ the predicted class is

$$\arg\max_{k \in K} D(x, F_k).$$

The performance of the proposed classifier is evaluated through a large simulation study and on real data sets. We compare it with other maximum depth classifiers, based on Mahalanobis, projection and halfspace depths, with linear and quadratic discriminant analysis, and for the real data sets also with K-nearest neighbours and SVM.

DRAFT

## References


[1] Chakrabarty, T. K. and Sharma, D. (2021). A Generalization of the Quantile-Based Flattened Logistic Distribution. *Annals of Data Science*, 8(3):603–627.

[2] Ghosh, A. K. and Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350.

[3] Kong, L. and Mizera, I. (2012). Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica*, 22(4):1589–1610.

[4] Liu, R., Parelius, J., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27(3):783–858.

[5] Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, 2:49–55.

[6] Redivo, E., Viroli, C., and Farcomeni, A. (2023). Quantile-based distribution functions and their use for classification, with application to naïve bayes classifiers. *Statistics and Computing*, (forthcoming).

[7] Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56(2):214–232.

[8] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461 – 482.

# Clustering alternatives in the preference-approval context

Alessandro Albano[a], José Luis García-Lapresta[b], Mariangela Sciandra[a], and Antonella Plaia[a]

[a]Department of Economics, Business and Statistics - University of Palermo;
alessandro.albano@unipa.it, mariangela.sciandra@unipa.it
antonella.plaia@unipa.it
[b]Departamento de Economía Aplicada, Universidad de Valladolid ; lapresta@uva.es

DRAFT

**Abstract**

Preference-approval structures combine preference rankings and approval voting to express preferences over a set of alternatives. This paper proposes a new method for clustering and visualizing alternatives in the context of preference-approvals. Firstly, we present a new family of pseudometrics defined on a set of alternatives, evaluated via preference-approvals. The distances among alternatives are used as input in the Ranked k-medoids algorithm to find clusters. Finally, clusters are visualized in a two-dimensional space using non-metric multidimensional scaling. We show, through an application to real data, that our approach allows for reducing the complexity of the preference-approval space and facilitates its interpretation.

***Keywords:*** Preference rankings, Approval voting, Preference-approvals, Kemeny distance

## 1. Introduction

Preference-approval structures involve collecting data about people's opinions on a set of alternatives to discover which they prefer and which they consider unacceptable. In preference-approval structures, voters are asked to classify the alternatives as acceptable or unacceptable and rank them in order of preference. This information can then be used to make decisions or to understand how people feel about the different options.

Preference-approval structures have been the subject of recent research from various perspectives, as shown in several papers (3; 4; 2; 1). However, there has been little focus on creating clustering algorithms specifically for preference-approvals.

This paper introduces a method for clustering and visualizing alternatives in the preference-approval context. Firstly, we define a new family of pseudometrics on the set of alternatives in the preference-approval context. Then, the Ranked k-medoids algorithm (see (5)) is employed to create clusters of the alternatives based on the similarities calculated using these pseudometrics. Finally, the clusters are represented in a two-dimensional space using non-metric multidimensional scaling. This method allows for visualizing data in reduced dimensions while preserving the relationships between points. We demonstrate how our approach can be applied to create meaningful clusters of the alternatives and visualize them.

The paper is organized as follows. Section 2 introduces basic notation and concepts we use throughout the article. Section 3 contains our proposal for clustering alternatives. Section 4 includes a case study. Finally, Section 5 concludes the paper with some remarks.

## 2. Notation

Suppose a set of voters $V = \{v_1, \ldots, v_m\}$, with $m \geq 2$, are asked to order $n$ different alternatives, the ranking $\pi$ is a mapping function from the set of alternatives $X = \{x_1, \ldots, x_n\}$ to the set of ranks $\pi = \{P_\pi(x_1), \ldots, P_\pi(x_i), \ldots, P_\pi(x_n)\}$, where $P_\pi : X \longrightarrow \{1, \ldots, n\}$ assigns the rank of each alternative. If the $n$ alternatives are ranked in $n$ different ranks, a complete (full) ranking or linear order is achieved. In certain cases, some alternatives could receive the same rank, and then a tied ranking or a weak order is obtained.
In the framework of preference-approval modelling, each preference ranking, $\pi$, is paired with an approval vector, $A$. For any given set $X$ of alternatives, we define approvals by partitioning $X$ into $G$, the set of good alternatives, and $U = X \backslash G$ the set of unacceptable alternatives, where $G$ and $U$ can be empty sets.
We represent a voter's preference-approval profile by a top-down order of alternatives with a horizontal bar: alternatives above the bar are approved, and those below are rejected.

$$\begin{array}{c} x_2 \\ x_1 \\ \hline x_3 \\ \hline x_4 \end{array}$$

The previous representation indicates that the voter's three top-ranked alternatives ordered as: $x_2 \succ x_1 \succ x_3$ are approved and and the voter's bottom-ranked alternative $x_4$ is disapproved. The preference-approval profile is codified as follows:

$$\pi_1 = (2, 1, 3, 4) \quad A_1 = (1, 1, 1, 0).$$

## 3. The proposal

Given a profile $\big((\pi_1, A_1), \ldots, (\pi_n, A_n)\big) \in \mathcal{R}(X)^n$, where $\mathcal{R}(X)$ denotes the set of preference-approvals on $X$, and two alternatives $x_i, x_j \in X$, we now present two indices that quantify the distance between these alternatives in terms of preference and approvals, respectively, for each voter $v_k \in V$.
The *preference-discordance* between $x_i$ and $x_j$ for the voter $v_k \in V$ is defined as

$$p_{ij}^k = \frac{1}{n-1} \cdot |P_{\pi_k}(x_i) - P_{\pi_k}(x_j)|, \tag{1}$$

where $p_{ij}^k \in [0, 1]$. The *approval-discordance* between $x_i$ and $x_j$ for the voter $v_k \in V$ is defined as

$$a_{ij}^k = |I_{A_k}(x_i) - I_{A_k}(x_j)|, \tag{2}$$

where $a_{ij}^k \in \{0, 1\}$. In order to define an overall measure of discordance between each pair of alternatives, we consider the family of *weighted means*, $h : [0, 1] \times [0, 1] \longrightarrow [0, 1]$, where $\lambda$ is the weighting parameter, defined as

$$h(x, y) = \lambda \cdot x + (1 - \lambda) \cdot y, \tag{3}$$

where $\lambda \in [0, 1]$. Taking into account the preference and approval discordances introduced in Eqs. (1) and (2), respectively, and the family of weighted means defined in Eq. (3), we now introduce a global measure of discordance between pairs of alternatives.

**Definition 1.** *Given a profile* $\big((\pi_1, A_1), \ldots, (\pi_n, A_n)\big) \in \mathcal{R}(X)^n$ *and* $\lambda \in [0, 1]$, *the mapping* $\delta_\lambda : X \times X \longrightarrow [0, 1]$ *is defined as*

$$\delta_\lambda(x_i, x_j) = \frac{1}{m} \cdot \sum_{k=1}^{m} \Big(\lambda \cdot p_{ij}^k + (1 - \lambda) \cdot a_{ij}^k\Big). \tag{4}$$

The $\delta_\lambda$ measure is used in the clustering procedure.

DRAFT

## 3.1 Clustering procedure and visualization

In this paper, we use the *Ranked $k$-medoids* (RKM) algorithm (see Zadegan et al. (5)) to find clusters. The Ranked $k$-medoids (RKM) algorithm is a distance-based clustering technique used to find clusters in a dataset. It works by assigning a rank to each pair of alternatives in the dataset based on how similar they are to each other. The rank matrix, $K$, is a measure of the hostility relationship between the alternatives in the dataset, with lower ranks indicating greater similarity. The $i$th hostility value, $hv_i$, is computed by summing all of the $i$th object's ranks with respect to a set of alternatives. The RKM algorithm first selects the medoids randomly, then selects the group of most similar objects to each medoid, calculates the hostility values of each object in the group, and selects the object with the highest hostility value as the new medoid. This process is repeated until the objects are assigned to the most similar medoid. The procedure needs the number of clusters to be specified in advance, but the best configuration can be found by using methods such as the Silhouette Coefficient. The resulting clusters can be visualized in a 2-dimensional space using multidimensional scaling (MDS), specifically Non-metric Multidimensional Scaling in this case. MDS attempts to represent a distance matrix as a simple geometrical model or map, such that the perceived distance between two alternatives is reflected in the distance between the points representing them in the model. Non-metric MDS preserves the rank order of the proximities, and the coordinates are found by minimizing the Stress function, which is based on the squared differences between the observed proximities and the derived disparities.

# 4. Case study

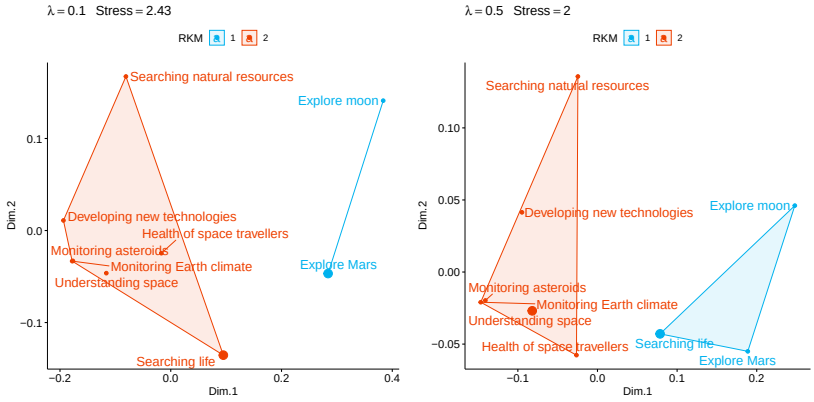
Data used in this section comes from the survey "American Trends Panel Wave 33" conducted by the Pew Research Center. This research organization conducts data-driven social science research, including public opinion surveys and demographic studies. In this particular analysis, the survey is used to gather opinions from United States citizens on the priorities of the space agency NASA. The survey was conducted from March 27 to April 9, 2018, and a total of $m = 2541$ respondents were asked to assess the priority that NASA should give to a list of $n = 9$ lines of action. The respondents used a qualitative scale to express their opinions, where $l_1$ indicates a top priority option, $l_2$ an important but lower priority, $l_3$ not too important, $l_4$ should not be done, and $l_5$ indicates no answer.
This data is used to assess the priorities of United States citizens for NASA.

Table 1: The $n = 9$ alternatives (NASA lines of action).

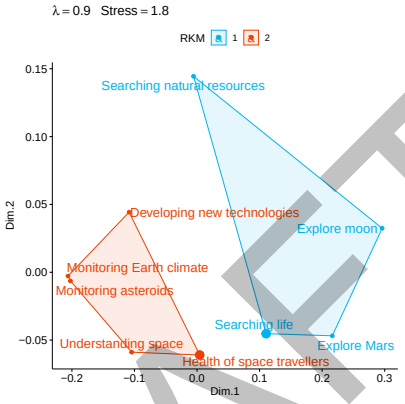
| Alternatives | Names |
|---|---|
| $x_1$ | Searching for life and planets that could support life |
| $x_2$ | Searching for raw materials and natural resources that could be used on Earth |
| $x_3$ | Conducting basic scientific research to increase knowledge and understanding of space |
| $x_4$ | Developing technologies that could be adapted for uses other than space exploration |
| $x_5$ | Monitoring asteroids and other objects that could potentially hit the Earth |
| $x_6$ | Monitoring key parts of the Earth's climate system |
| $x_7$ | Sending human astronauts to explore the moon |
| $x_8$ | Sending human astronauts to explore Mars |
| $x_9$ | Conducting scientific research on how space travel affects human health |

To remove neutral answers from the survey data, the respondents who provided at least a "No answer" response were excluded from the analysis. This represented about 3% of the total sample size $m$. The remaining responses were then arranged into a preference-approval, with the linguistic terms "$l_1$" and "$l_2$" indicating an acceptable alternative. This allows for a more focused analysis of the respondents' preferences and approval ratings for the different lines of action.
The clusters estimated by the RKM algorithm, using three different values of $\lambda = 0.1, 0.5, 0.9$, are shown in Fig 1. For each scenario, the clusters, medoids and Stress values reached by the MDS are illustrated for graphical representation.

DRAFT




(a) $\lambda = 0.1$, Stress=2.43 (b) $\lambda = 0.5$, Stress=2




(c) $\lambda = 0.9$, Stress=1.8

Figure 1: Graphical representation of RKM clusters.

DRAFT

The Stress coefficient, which measures the model's goodness of fit, varies between 1.8 and 2.43 in this analysis, indicating an excellent fit that tends to improve as $\lambda$ increases. The optimal number of clusters, as determined by the Silhouette criterion, is 2, regardless of the value of $\lambda$.
The parameter $\lambda$ has a visible effect on the formation of the clusters. When $\lambda = 0.1$, Cluster 1 includes only the alternatives related to space exploration (Moon and Mars), the respondents tend to give similar approval ratings to these two alternatives. Increasing the value of $\lambda$ to 0.5 causes Cluster 1 to include the alternative "Searching life". When $\lambda = 0.9$, giving more weight to the rankings, Cluster 1 also includes the alternative "Searching natural resources". As a result, Cluster 1 includes the four alternatives that are most frequently placed at the bottom of the respondents' preference-approvals.

## 5. Conclusions

This paper presents a new method for clustering alternatives in preference-approval structures, which are increasingly important in social choice because they allow decision-makers to describe their preferences using more flexible and intuitive ordinal information. The method involves introducing a family of pseudometrics, $\delta_\lambda$, which can quantify the distance between alternatives based on the "preference-discordance" and "approval-discordance", and a parameter $\lambda$ that regulates the weight given to each component. The Ranked $k$-medoids partitioning algorithm is used to find clusters based on the similarities between pairs of alternatives calculated using the pseudometrics. The resulting clusters are visualized in a 2-dimensional space using Non-Metric Multidimensional Scaling.

## References


[1] Albano, A., García-Lapresta, J. L., Plaia, A., and Sciandra, M. (2022). A family of distances between preference-approvals. *Annals of Operation Research*, pages 1–29.

[2] Barokas, G. and Sprumont, Y. (2022). The broken Borda rule and other refinements of approval ranking. *Social Choice and Welfare*, 58(1):187–199.

[3] Erdamar, B., García-Lapresta, J. L., Pérez-Román, D., and Sanver, M. R. (2014). Measuring consensus in a preference-approval context. *Information Fusion*, 17:14–21.

[4] Kruger, J. and Sanver, M. R. (2021). An Arrovian impossibility in combining ranking and evaluation. *Social Choice and Welfare*, 57(3):535–555.

[5] Zadegan, S. M. R., Mirzaie, M., and Sadoughi, F. (2013). Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems*, 39:133–143.


DRAFT

# Computational assessment of k-means clustering on a Structural Equation Model based index

Mariaelena Bottazzi Schenone[a], Elena Grimaccia[b], Maurizio Vichi[a]

[a] Department of Statistical Sciences, Sapienza University, Rome; mariaelena.bottazzischenone@uniroma1.it,maurizio.vichi@uniroma1.it

[b] ISTAT - Italian National Institute of Statistics, Rome; elgrimac@istat.it

**Abstract**

This paper proposes an alternative method for the choice of the number of centroids in a cluster analysis, when the groups' order is relevant. Differently from commonly used approaches, aimed at finding the minimum number of clusters, the illustrated method aims at finding the maximum one. Given that the clusters are ordered, this allows to define a granular ranking among them. The *k*-means partitioning algorithm is applied to an index resulting from a Structural Equation Model. The procedure is implemented on a measure of air pollution in urban areas: a clustering of main Italian cities, according to the optimal number of air pollution levels, is the final result. The analysis' interpretation provides useful information to design policies aimed at reducing air pollution.

**Key words:** Cluster analysis, Computational assessment, Structural Equation Models, Air Pollution, Latent variable models, Environmental statistical models.

DRAFT

## 1. Introduction

Most of the commonly used methods to identify the optimal number of clusters in a cluster analysis aim at finding the minimum *k* (Gap statistics [16]; Silhouette method [13]; Elbow method [14]). In this paper, the idea behind the choice of *k* is the opposite: the goal is to find the maximum number of "*distinguished*" clusters: what are the "well-distinguished" clusters is defined by means of bootstrap confidence intervals for clusters' centroids. Given a sample of units and a number of clusters *k*, these intervals can be computed by bootstrapping those units a number B of times and computing each time the *k*-means. The results are B vectors of *k* centroids. The final estimates of the *k* clusters' centroids as well as their empirical distributions can be obtained computing the mean and plotting the histograms of the bootstrap replicates, respectively. Given the *k* centroids' point estimates with the corresponding $\alpha/2$ and $(1-\alpha/2)$ percentile estimates, it is possible to build *k* percentile confidence intervals of the desired confidence level $\alpha$. If some of these confidence intervals do overlap by more than a fixed small constant $\varepsilon$, then the clusters are not "well-distinguished". The optimal number of clusters $k^*$ will be the maximum *k* such that none of the *k* intervals do overlap by more than $\varepsilon$.

A new air quality index has been estimated by means of the application of a Structural Equation Model (SEM) to pollutants variables, as it takes into account the multivariate relationships among contaminants ([2]). The analysis of the new index distribution among the Italian metropolitan areas can be useful to draw policy conclusions devoted to reducing air pollution levels.

The clustering analysis homogeneously groups Italian cities with respect to different air pollution levels. Assigning a rank to the cities within the same cluster, it is possible to classify them from the most to the less polluted.

The paper is structured as follows: Section 2 presents data used for the empirical study, Section 3 introduces SEM's specifications and modelling and the cluster analysis technique employed. In Section 4, an application on air quality for Italian cities is provided. In Section 5 concluding remarks are drawn.

## 2. Data

The dataset on which the study is conducted comes from the European Environmental Agency (EEA) and refers to 6 pollutants' emissions of 106 Italian provinces in 2022. For each city, the pollutant emission is computed as the average over the 365 days of daily median emissions. These 6 gases are the ones that the Environmental Protection Agency (EPA) individuated as major air pollutants and they are regulated by the Clean Air Act. They are: Ground-level ozone (O3), Particle pollution (also known as Particulate Matter (PM), including PM2.5 and PM10), Carbon monoxide (CO), Sulfur dioxide (SO2) and Nitrogen dioxide (NO2).

## 3. Methods: Structural Equation Model and Cluster analysis

In this paper, an index is built employing a Structural Equation Model. This model combines Confirmatory Factor Analysis and Multiple Regression Analysis into a comprehensive modelling framework that involves both endogenous and exogenous variables. In the so called "measurement part" of the model, the relations between the observed (manifest) variables (MVs) and the latent factors (LVs) can be studied. Moreover, the "structural part" of the model studies the causal relationships of the latent constructs among themselves. All these relations are estimated simultaneously ([8]; [3]; [15]; [6]).

In order to rank units with respect to the SEM-based index, the centroid-based model of *k*-means ([17]) is employed. The *k*-means method assumes that each observation is equal to one of the *k* centroids. All the observations assigned to each centroid, perturbed by error in measuring the features, forms a cluster. The clustering goal is to partition the units in a disjoint set of *k* clusters to maximise the dissimilarity between centroids of the clusters.

In this analysis, a centroid of a cluster is the mean value of the index. For this univariate clustering problem, the optimal number of clusters $k^*$ is chosen to be the highest possible, as long as centroids are statistically different according to a desired confidence level α. The idea is to rank clusters by their centroids, so that units belonging to a cluster have the rank of that cluster. Hence, $k^*$ is the maximum value for which all $k^*$ centroids' confidence intervals at significance level (1-α) do not overlap by more than a fixed constant ε.

However, because of its deterministic nature, *k*-means does not yield confidence information about centroids' distribution and estimated cluster memberships, although this could be useful for inferential purposes. It is possible to obtain such information by means of a non-parametric bootstrap procedure. This procedure provides centroids' distributions ([9]) which can be used to derive probabilistic membership information on each object from all bootstrap samples. It also yields confidence information about the centroids in the form of confidence intervals ([7]).

Given a sample of size n, for a given *k*, the partitioning algorithm is run. The corresponding *k* centroids' estimates and corresponding confidence intervals are built applying bootstrap to that sample of n units.

For each of the B bootstrap iterations, the *k* centroids are ordered from the smallest to the largest. This allows to easily match the centroids over different bootstrap samples. Once the matching is done, the final estimate of each of the *k* centroids is obtained as the mean of the corresponding B centroids' values. The α/2% and (1- α/2)% centroids' percentiles can be estimated in a similar way and the (1-α)% percentile bootstrap confidence intervals can be computed as follows: $\left[percentile_{\frac{\alpha}{2}}\,;\, percentile_{\frac{1-\alpha}{2}}\right]$.

The partitioning algorithm chosen is a centroid-based 1-dimensional *k*-means and units are classified according to the SEM-based index. In this particular, unidimensional case, an optimal dynamic programming *k*-means algorithm has been developed by Froese et al. ([5]). The algorithm is implemented in the Ckmeans.1d.dp R package ([18]).

The clustering algorithm is run for different values of *k*, starting from $k = 2$. At each iteration, if the *k* bootstrap confidence intervals do not overlap by more than ε, the *k* clusters can be considered well separated. Then, *k* is increased by one and the partitioning algorithm is run again. The procedure is

DRAFT

iterated until two overlapping confidence intervals are found. A crucial point is the need of ordering the clusters with respect to their centroid value, from the smallest to the largest. This allows to find the consecutive clusters' confidence intervals to be compared.

## 4. Application to urban air pollution for Italian metropolitan areas

In this study, a multidimensional index to measure air pollution is built by means of a hierarchical SEM ([4]). This flexible model has the advantages of taking simultaneously into account a number of levels in the hierarchy and to exploit the information available in meaningful explanatory variables.

Based on the results of a preliminary Explanatory Factor Analysis on the six main pollutants, a hierarchical, two latent factors model is estimated using the "sem" function from the R "lavaan" package ([12]). This function automatically standardizes the six pollutants, assigning negative weights to the variables that the factor reconstructs in the opposite direction from the others.

The resulting index, called Model Based-Air Pollution Index (MB-API), is normalized in 0-1. This index allows to rank cities with respect to their air pollution level.

The estimated model can be written as follows.

$$\begin{cases} f1 = 0.64PM2.5 + 0.62PM10 - 0.37O3 + 0.41SO2 \\ f2 = CO - 0.65NO2 \\ MBAPI = f1 + 0.79f2 \end{cases} \quad (1)$$

Fig.1 shows the different levels of air pollution, according to the MB-API, for the 106 Italian cities. The most polluted areas are in Piedmont, Lombardy, Veneto and Emilia-Romagna, while the less polluted cities are in the islands and in Calabria.

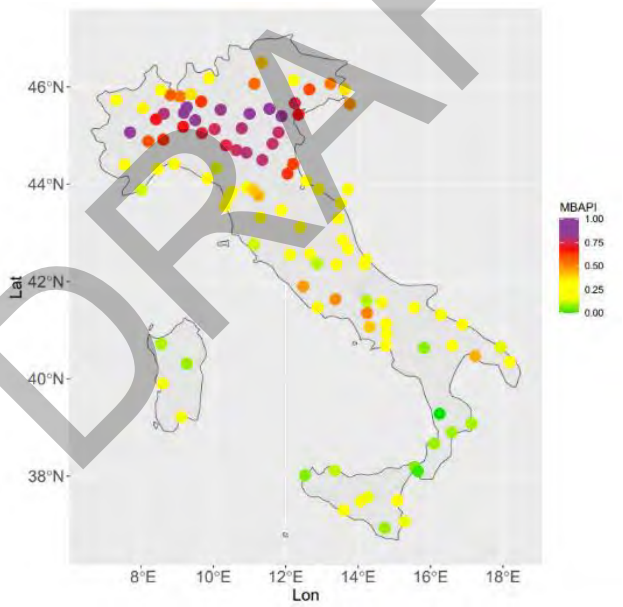



Figure 1: Air pollution index values in Italian metropolitan areas.

Cluster analysis is then applied to find groups of cities homogeneous with respect to the air pollution level. The Italian cities are grouped into clusters, each represented by a centroid that corresponds to an index value. It is important to note that to allow cities' ranking, cluster must be ordered with respect to the corresponding centroids. The clustering algorithm is run for $k = 2$ up to 10. The bootstrap procedure (with a number of bootstrap replicates equal to 10000) is used to compute the corresponding centroids' confidence intervals at 90% shown in Fig. 2.

Two clusters are considered well separated if the difference between the upper bound of a cluster and the lower bound of the consecutive one is smaller than a constant $\varepsilon = 0.01$ (1% of the index range, equal to 1). It is possible to note that for $k = 6$ the clusters do overlap by more than $\varepsilon$ and therefore the optimal $k$ should be 5. However, for $k = 7$ the clusters are well distinguished and therefore the optimal maximum number of non-overlapping clusters $k^*$ is 7.

DRAFT

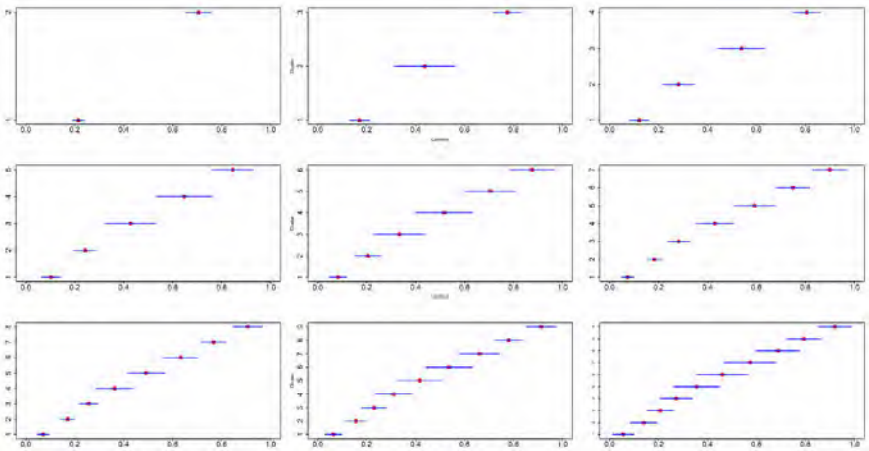


Figure 2: 90% bootstrap confidence intervals for *k*-means centroids. *k* ranges in 2-10.

Of course, the optimal number $k^*$ depends on the α and ε values chosen *a priori*. It is worthy to note that this method of choosing *k* is very different with respect to the classical Elbow, Silhouette and Gap Statistics methods, whose aim is to find the minimum (and not the maximum) *k* such that the clusters are well defined. In all 3 cases, in fact, $k^*=2$ (as shown in the plots of Fig.3).

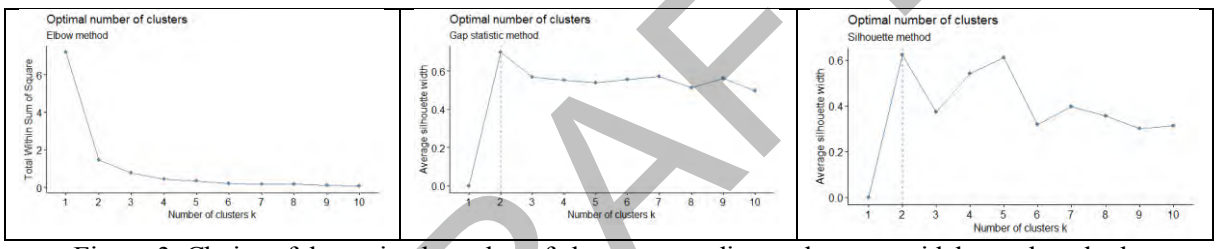



Figure 3: Choice of the optimal number of clusters according to the most widely used methods.

The maximum optimal number of clusters can be found also considering the maximum number of clusters whose centroids are simultaneously significantly different, according to nonparametric Wilcoxon tests ([10]), with confidence level chosen following the Bonferroni's correction ([1]). In this case, $k^*=10$ and therefore the more parsimonious solution $k^* = 7$ seems even more reasonable.

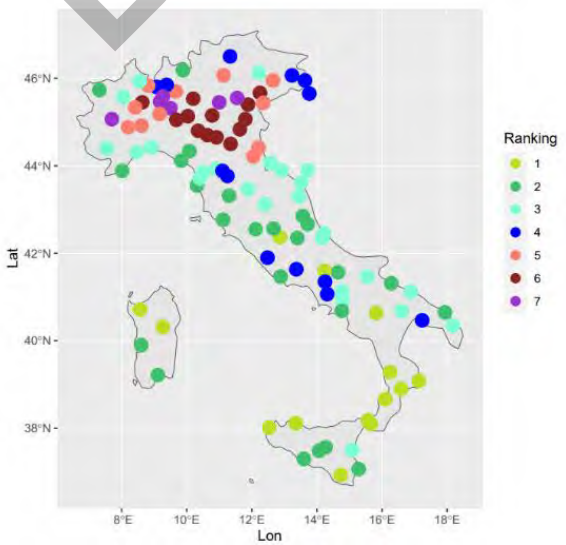



Figure 4: Choice of the optimal number of clusters according to the most widely used methods.

DRAFT

Basing on the previous results, groups are ranked from 1 to 7 considering the centroids' values from the highest to the lowest: rank 1 corresponds to the lowest centroid and therefore to the group of less air polluted cities.

The map in Fig. 4 shows air pollution distribution in Italy in 2022, highlighting groups of cities with a similar situation in terms of air pollution levels. It is possible to note that close points tend to have the same colour: cities in the same region often have a similar air pollution level.

## 5. Concluding remarks

The novelty of our paper resides in the fact that the *k*-means cluster analysis is employed in order to estimate the maximum number of significantly different centroids. The significance is assessed by means of percentile confidence intervals, built with a bootstrap procedure.

The performance of the proposed method is shown by an air pollution analysis: a measure of air quality in metropolitan areas has been developed based on a structural equation model. The procedure obtain a ranking of Italian cities, according to the optimal number of clusters of air pollution.

As a possible extensions of the study, the air pollution index can be improved considering also exogenous explanatory variables in the Structural Equation Model, such as the number of cars in the city or the percentage of green spaces. The model immediately integrates the new covariates. In case of adding more covariates, a different multidimensional clustering technique must be used and the obtained results can be compared with the unidimensional case.

Furthermore, a sensitivity and robustness analysis of the ranking can be conducted in a simulation framework, for instance computing average absolute shift in ranking index.

Thanks to this granular ranking, policy makers can easily identify the most polluted metropolitan areas and employ the estimated index as a unique measure of air pollution.

DRAFT

### References


[1] Armstrong R. : When to use the Bonferroni correction. Ophthalmic Physiol, 34(5):502-8 (2014).
[2] Boaz R. M., Lawson A. B., Pearce J. L. : Multivariate air pollution prediction modelling with partial missingness. Environmetrics, 30(7): e2592 (2019).
[3] Bollen K.A. : Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models. MIS Quarterly, 35(2), 359-372 (2011).
[4] Cavicchia C., Vichi M. : Second-order disjoint factor analysis. Psychometrika, 87 (1), 289–309 (2022).
[5] Froese R., Klassen J. W., Leung C. K. and Loewen T. S. : The Border K-Means Clustering Algorithm for One Dimensional Data. IEEE International Conference on Big Data and Smart Computing, pp. 35-42 (2022).
[6] Hair J. F., Sarstedt M. : Explanation plus prediction – The logical focus of project management research. Project Management Journal, 52(4), 319–322 (2021).
[7] Hofmans J. : On the Added Value of Bootstrap Analysis for K-Means Clustering (2015).
[8] Landis R. S., Beal D. J., Tesluk P.E. : Comparison of Approaches to Forming Composite Measures in Structural Equation Models. Organizational Research Methods 2000 3: 186 (2000).
[9] Martella F., Vichi M. : Clustering microarray data using model-based double K-means (2012).
[10] Rey D., Neuhäuser M. : Wilcoxon-Signed-Rank Test. International Encyclopedia of Statistical Science. Springer (2011).
[11] Rizzo M. : Statistical Computing with R. Computer Science and Data Analysis Series. Chapman & Hall/CRC The R Series. p. 198 (2008).
[12] Rosseel Y. : lavaan: An R Package for Structural Equation Modeling, Journal of Statistical Software, 48 (2) (2012).
[13] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
[14] Shi C., Wei B., Wei S. Wang W., Liu H., Liu J. : A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 1-16 (2021).
[15] Tarka P. : An overview of structural equation modeling: its beginnings, historical development,

usefulness and controversies in the social sciences. Quality & Quantity, 52, 313–354 (2018).
[16] Tibshirani R., Walther, G., Hastie, T. : Estimating the Number of Clusters in a Data Set via the Gap Statistic.Journal of the Royal Statistical Society. Series B (Statistical Methodology), 63(2), 411– 423 (2001).
[17] Vichi M., Cavicchia C., Groenen P. J. F. : Hierarchical Means Clustering, Journal of Classification. https://doi.org/10.1007/s00357-022-09419-7 (2022).
[18] Wang H. and Song M. : Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. The R Journal Vol. 3/2 (2018).

DRAFT

# Handling missing data in complex phenomena: an ultrametric model-based approach for clustering

Francesca Greselin[a] and Giorgia Zaccaria[a]

[a]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy;
francesca.greselin@unimib.it, giorgia.zaccaria@unimib.it

**Abstract**

In the model-based clustering literature, we find several methodologies to study latent structures underlying the data, among which mixtures of factor analyzers. However, none of them can detect hierarchical relationships among latent variables. The Ultrametric Gaussian Mixture Model (UGMM) is intended to reach this goal by identifying a hierarchy of variables, starting by partitioning the variables into a reduced number of groups per mixture component. Nonetheless, up to now, it requires complete observations, which is often not the case in real data collections. In this paper, we propose the extension of UGMM in the missing data framework. The proposal is applied to a real data set for inspecting the relationships among features of songs of different genres.

***Keywords:*** ultrametricity, Gaussian mixture models, missing information, hierarchy of latent concepts, heterogeneous populations, features of songs

DRAFT

## 1. Introduction

Model-based clustering is one of the most well-known approaches to model heterogeneous multivariate data by specifying a probability distribution (2). The central statistical methodology for clustering assumes the data coming from a heterogeneous population defined by a finite collection of $G$ sub-populations, and is based on finite mixture models. Specifically, the dominant model is the Gaussian mixture model in which the distribution of the mixture components is Gaussian with mean vector and covariance matrix representing its parameters (9). In the specialized literature, different parameterizations of the component-covariance matrix have been proposed with a twofold goal: (i) specifying more parsimonious models by reducing the number of parameters that mostly derive from the component-covariance matrices, i.e., $Gp(p+1)/2$ parameters, and (ii) introducing models able to inspect latent structures underlying the data, e.g., factorial structure. The parameterization introduced by (1), and extended by (4), fits the first aim and is based upon the eigen-decomposition of the covariance matrix. Indeed, its number of parameters can vary between 1 and $Gp(p+1)/2$, by constraining them to be equal within and/or across components. The models proposed by (6), and further developed by (10), are instead able to reach both goals previously described by parameterizing the component-covariance matrices via a factorial structure with $q$ factors, whose number of parameters varies between $(pq - q(q-1)/2)$ and $G(pq - q(q-1)/2) + Gp$ depending on constraints. Nonetheless, none of these models is able to detect hierarchical relationships among variables within components. To fill this gap, (3) introduced a Gaussian mixture model with an extended ultrametric covariance structure which is able to pinpoint the hierarchical structure of variables. This model assumes that, within each component, the variables are partitioned into $q$ groups characterized by three features: the group variance, the covariance within the groups and

the covariance between the groups identifying their aggregation in pairs (under the ultrametricity constraint). The ultrametric parameterization leads to the estimation of $2q + p - 1$ parameters for each component of the mixture.

Although useful in several real applications, the aforementioned methodologies can only handle the occurrence of missing information (i.e., a data matrix in which some entries are not observed) by using ad hoc methods. Case deletion or imputation are used to force the incomplete data set into a rectangular complete-data format beforehand the model estimation. To overcome this issue, assumptions on the missing data probability have to be taken into account (11). The missing mechanism can be of three types: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR), where the probability of a missing datum in an observation does not depend on the other values, is independent of the missing value or depends on the missing value itself, respectively. Under the MCAR and the less restrictive MAR assumptions the missing mechanism is considered to be ignorable, i.e., the parameters of the missing data distribution are distinct from the parameters of the observed data distribution (8; 12). The extension of Gaussian mixture models in the missing data framework was introduced by (7), whereas (15) proposed mixtures of factor analyzers models with missing information, both under the MAR mechanism. Only recently, Gaussian mixture models were studied under the MNAR mechanism (14).

In this paper, we introduce the Ultrametric Gaussian Mixture Model (UGMM) that handles incomplete information when the missing mechanism is assumed to be MAR. The proposal is intended to identify hierarchical structures of variables characterizing a multidimensional phenomenon within heterogeneous populations, when the data are affected by missingness, as often occurs in real data applications. The proposed methodology is applied herein for studying the hierarchical relationships among features of songs pertaining to different genres.

The outline of the paper is as follows. In Section 2, the ultrametric Gaussian mixture model with missing data handling is described in details together with its estimation. A real data application is provided in Section 3. A final discussion concludes the paper in Section 4.

DRAFT

## 2. The model

Let $\boldsymbol{x}$ be a random sample composed of $n$ observations of dimension $p$, that is $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$. The GMM density of $\boldsymbol{x}_i$ is defined by a mixture of $G$ multivariate Gaussian distributions as follows

$$f(\boldsymbol{x}_i \mid \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \, \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where $\pi_g$ is the probability that an observation has been generated by the $g$th component of the mixture, so that $\pi_g > 0, g = 1, \ldots, G$, and $\sum_{g=1}^{G} \pi_g = 1$, and $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the mean vector and covariance matrix of the $g$th normal component, respectively. In UGMM, the latter has a particular form defined by

$$\boldsymbol{\Sigma}_g = \boldsymbol{V}_g(\boldsymbol{\Sigma}_{\mathrm{W}_g} + \boldsymbol{\Sigma}_{\mathrm{B}_g})\boldsymbol{V}_g' - \mathrm{diag}(\boldsymbol{V}_g\boldsymbol{\Sigma}_{\mathrm{W}_g}\boldsymbol{V}_g') + \mathrm{diag}(\boldsymbol{V}_g\boldsymbol{\Sigma}_{\mathrm{V}_g}\boldsymbol{V}_g'), \tag{2}$$

that is extended ultrametric (Definition 2 in 3) if, for each component $g$ of the mixture, $\boldsymbol{V}_g$ is binary and row-stochastic, $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ is symmetric with zero diagonal entries and off-diagonal values whose triplets comply with the ultrametric condition and whose maximum is lower than or equal to the minimum of $\boldsymbol{\Sigma}_{\mathrm{W}_g}$, and such that $\boldsymbol{\Sigma}_g$ is positive definite. Thus, the parameters in $\boldsymbol{\Psi}$ are $\boldsymbol{\pi} = [\pi_g]_{g=1}^{G}, \boldsymbol{\mu} = [\boldsymbol{\mu}_g]_{g=1}^{G}, \boldsymbol{V} = [\boldsymbol{V}_g]_{g=1}^{G}, \boldsymbol{\Sigma}_{\mathrm{V}} = [\boldsymbol{\Sigma}_{\mathrm{V}_g}]_{g=1}^{G}, \boldsymbol{\Sigma}_{\mathrm{W}} = [\boldsymbol{\Sigma}_{\mathrm{W}_g}]_{g=1}^{G}, \boldsymbol{\Sigma}_{\mathrm{B}} = [\boldsymbol{\Sigma}_{\mathrm{B}_g}]_{g=1}^{G}$, where $\boldsymbol{V}_g$ is the variable-group membership matrix that specifies a partition of the variable space into a reduced number $q$ of groups in each component, $\boldsymbol{\Sigma}_{\mathrm{V}_g}$ is the diagonal group-variance matrix identifying the variance of each of the $q$ groups per component, $\boldsymbol{\Sigma}_{\mathrm{W}_g}$ is the diagonal within-group covariance matrix representing the covariance within the variable groups for each mixture component, and $\boldsymbol{\Sigma}_{\mathrm{B}_g}$ is the component between-group covariance matrix.

Commonly, Gaussian – and more generally finite – mixture models are estimated via the Expectation-Maximization algorithm (EM, 5) that works in an incomplete data framework, where the source of miss-

Table 1: Song features and their description

| Song feature | Description |
|---|---|
| Danceability | Sustainability for dancing based on a combination of musical elements |
| Energy | Perceptual measure of intensity and activity |
| Loudness | Overall loudness in decibel |
| Speechiness | Presence of spoken words |
| Acousticness | Measure of acousticness |
| Instrumentalness | Non-vocal presence |
| Liveness | Presence of an audience in the recording |
| Valence | Musical positiveness |
| Tempo | Overall estimated tempo in beats per minute |

ingness originates from the unit-component membership variable $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)'$, with $w_{ig} = 1$ if the $i$th observation belongs to the $g$th component, $w_{ig} = 0$ otherwise. When missing information occurs in the data as missing values at random (i.e., MAR), a second source of missingness has to be considered. In this case, each observation can be split in two parts: the *observed* part of dimension $p_i$ and the *missing* part of dimension $(p - p_i)$. The complete data log-likelihood function of UGMM, i.e., subject to the aforementioned constraints on the component-covariance matrix expressed in (2), is

$$\ell_c(\boldsymbol{\Psi}; \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{w}) = \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} \ln\Big(\pi_g \phi(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{x}_i^{\text{mis}}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\Big). \tag{3}$$

Under the MAR mechanism, $\phi(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{x}_i^{\text{mis}}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \phi(\boldsymbol{x}_i^{\text{obs}}; \boldsymbol{\mu}_g^{\text{obs}}, \boldsymbol{\Sigma}_g^{\text{obs}}) \times \phi(\boldsymbol{x}_i^{\text{mis}} \mid \boldsymbol{x}_i^{\text{obs}}; \boldsymbol{\mu}_g^{\text{mis}}, \boldsymbol{\Sigma}_g^{\text{mis}})$ thanks to the separability assumption and the properties of the Gaussian distribution. It has to be noticed that the pattern of missingness depends on the observation, even if this specifying is removed herein for simplicity reasons.

The EM algorithm can be thus implemented to obtain the estimates of the model parameters in $\boldsymbol{\Psi}$. Particularly, after setting initial values $\boldsymbol{\pi}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{V}^{(0)}$, $\boldsymbol{\Sigma}_{\text{V}}^{(0)}$, $\boldsymbol{\Sigma}_{\text{W}}^{(0)}$, $\boldsymbol{\Sigma}_{\text{B}}^{(0)}$, the two following steps are repeated until convergence.

*E-step.* At iteration $r$, the expectation of the complete-data log-likelihood in Eq. (3) has to be computed. It corresponds to the calculation of the expected values of $W_{ig}$, $W_{ig}\boldsymbol{X}_i^{\text{mis}}$ and $W_{ig}\boldsymbol{X}_i^{\text{mis}}\boldsymbol{X}_i^{\text{mis}\prime}$ given the observed data $\boldsymbol{x}_i^{\text{obs}}$ and the current estimates of the unknown parameter, i.e., $\widehat{\boldsymbol{\Psi}}^{(r-1)}$.

*M-step.* At iteration $r$, the estimates of the prior probabilities $\pi_g$ and the mean vectors $\boldsymbol{\mu}_g$ are derived (see 9, among others), as well as those of the component-covariance matrices $\boldsymbol{\Sigma}_g$ (see 3, for the estimation of $\boldsymbol{V}_g$, $\boldsymbol{\Sigma}_{\text{V}_g}$, $\boldsymbol{\Sigma}_{\text{W}_g}$ and $\boldsymbol{\Sigma}_{\text{B}_g}$).

## 3. Clustering songs with respect to their features in the presence of missing information

The proposed methodology is applied herein for classifying songs with respect to their genre (*Rhythm and Blues (RnB)*, *hardstyle* and *techno*) and identifying relationships among song features within different genres. Indeed, musical elements can play different roles in the definition of the pleasantness of a song, depending on the song genre.

The data set analyzed in this paper is composed of 165 songs – 52 RnB, 39 hardstyle and 74 techno – characterized by the nine features described in Table 1. The missing rate over the total number of values is 7%. UGMM is estimated by fixing $G = 3$, i.e., the number of genres in the data set, and selecting $q$ into the set $\{1, \ldots, 5\}$ according to the Bayesian Information Criterion (13). The best $q$ turns out to be $q = 3$.

DRAFT

Table 2: Clustering structure of songs in genres

| Theoretical | Estimated | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| RnB | 50 | 0 | 2 |
| hardstyle | 5 | 0 | 34 |
| techno | 0 | 69 | 5 |

The clustering structure is summarized in Table 2, where the theoretical and the estimated partitions in genres are compared. The latter is obtained by considering the membership of each song to the component of the mixture on which it has the highest posterior probability (i.e., maximum a posteriori). As a result, the theoretical partition in genres turns out to be recovered except for twelve songs, corresponding to a misclassification rate of 7%. Specifically, the estimated components can be labeled as follows: component 1 represents the RnB songs, component 2 the techno songs and component 3 the hardstyle songs of the sample.

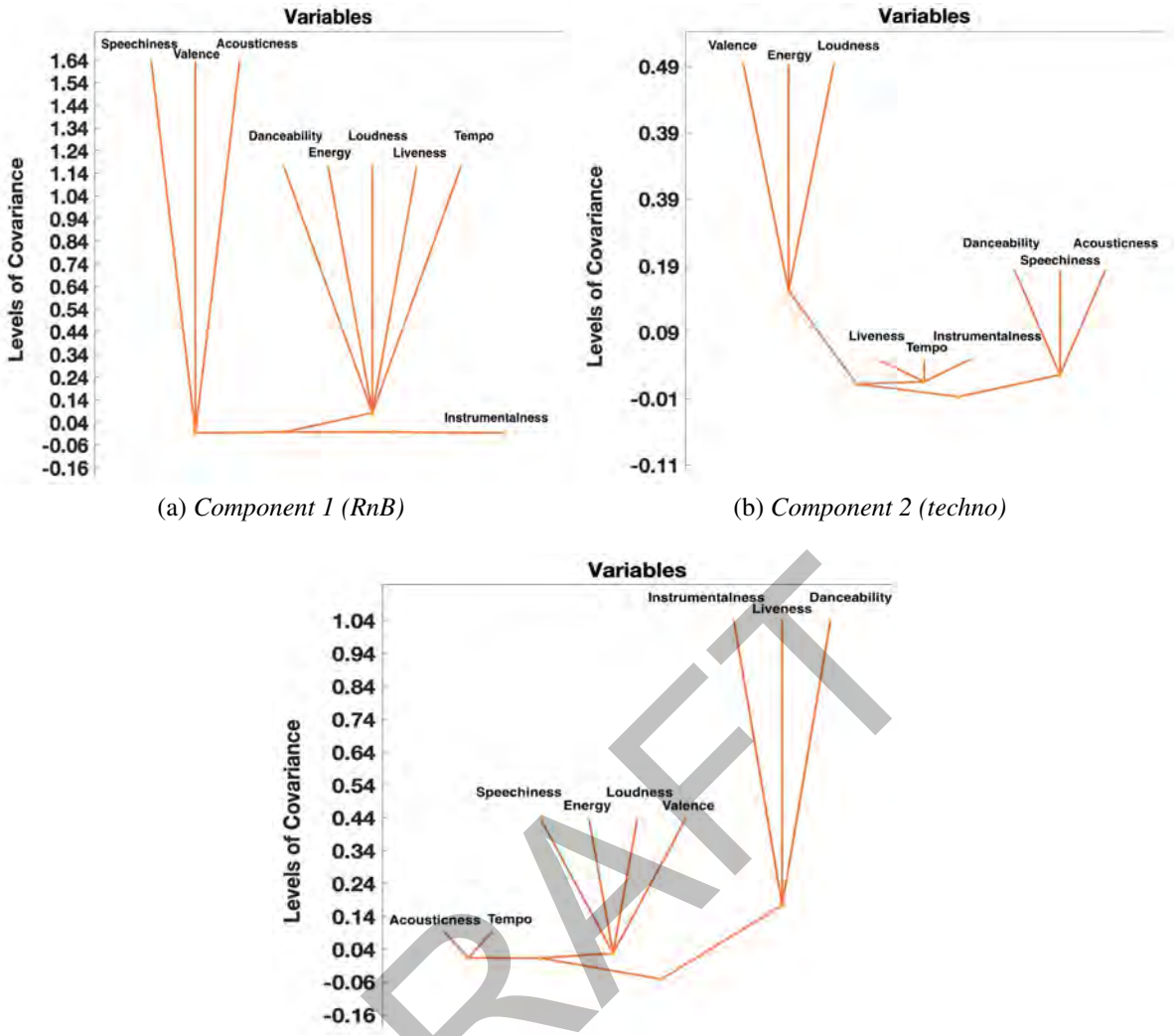
The hierarchical structure of variables per component is depicted in Figure 1. By analyzing the song features' partition in groups, we can notice that the main features characterizing each genre are component-specific. For instance, in the RnB component, the first aggregation in group occurs among *Danceability*, *Energy*, *Loudness*, *Liveness* and *Tempo* with a positive relationship of among these song features, that reasonably seem to be the most important ones. For this genre, it can be noticed that the variables *Speechiness*, *Valence* and *Acousticness* are lumped together in a group, at the same level at which they are merged with the first group previously described and the singleton composed of *Instrumentalness*. Moreover, their aggregation (covariance) level is around 0, revealing that each of these features defines a unique part of the song pleasantness. In the component representing techno songs, the group with the highest internal consistency is defined by *Valence*, *Energy* and *Loudness*. The aggregation of the three groups occurs at a low level of covariance by highlighting, also in this case, the evidence of uncorrelated groups that separately contribute to the definition of the pleasantness of techno songs. This characteristic is even shared by the hardstyle songs, whose highly consistent group is defined by *Instrumentalness*, *Liveness* and *Danceability*. Furthermore, the hierarchical structure results in Figure 1 show "recurring relationships" between song features across components that can find a theoretical explanation. For instance, *Energy* and *Loudness* are in the same variable group for each genre; indeed, energetic songs are usually loud. The differences among genres and the definition of song pleasantness within them prove the need of introducing a new methodology in the literature able to detect component-specific hierarchical structure of variables also in the presence of missing values.

DRAFT

## 4. Conclusions

In this paper we introduce a model-based clustering approach for detecting hierarchical latent structures of variables when missing information occurs in the data. To reach this goal, the component-covariance matrices of a Gaussian mixture model are parameterized by an extended ultrametric covariance structure able to pinpoint a variable partition into groups and their hierarchical relationships within each component (3).

The proposal is applied herein on a data set with 7% of missing values, referring to three genres of songs that represent the clustering structure. Within each genre, the hierarchy of song features is inspected and component-specific characteristics defining the song pleasantness are highlighted. It has to be noticed that the three variable groups identified for each mixture component – that differ across components – result to be uncorrelated. Furthermore, this happens also within groups, e.g, for the Rnb songs where a higher number of groups would probably provide a better fit ($q = 3$ for each genre). For this reason, a further development of the proposed methodology can address this problem by allowing the number of variable groups $q$ to vary across components.




(a) *Component 1 (RnB)*

(b) *Component 2 (techno)*

(c) *Component 3 (hardstyle)*

Figure 1: Path diagram representing the hierarchical relationships among variables within components

DRAFT

**Acknowledgments** The Authors' work was supported by Milano-Bicocca University Fund for Scientific Research, 2021-ATE-0707.

## References

[1] Banfield, J.D. and Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)

[2] Bouveyron, C., Celeux, G., Murphy, T.B., Raftery A.E.: Model-based clustering and classification for data science: With applications in R. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge (2019)

[3] Cavicchia, C., Vichi, M., Zaccaria, G.: Gaussian Mixture Model with an extended ultrametric covariance structure. Adv. Data Anal. Classif. **16**, 399–427 (2022)

[4] Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**, 781–793 (1995)

[5] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B Stat. Methodol. **39**, 1–38 (1977)

[6] Ghahramani, Z., Hinton, G.H.: The EM algorithm for factor analyzers. Technical report CRG-TR-96-1, University of Toronto, Toronto (1997)
[7] Ghahramani, Z., Jordan, M.I.: Learning from incomplete data. Technical report, AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab (1994) Available at https://mlg.eng.cam.ac.uk/zoubin/papers/foo.pdf
[8] Little, R., Rubin, D.B.: Statistical analysis with missing data (3rd edition). Wiley, New York (2019)
[9] McLachlan, G.J., Peel, D.: Finite mixture models. Wiley, New York (2000)
[10] McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)
[11] Rubin, D. B.: Inference and missing data. Biometrika **63**, 581–592 (1976)
[12] Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC, New York (1997)
[13] Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
[14] Sportisse, A., Marbac, M., Biernacki, C., Boyer, C., Celeux, G., et al.: Model-based clustering with Missing Not At Random data. hal-03494674v2 (2021) Available at https://arxiv.org/abs/2112.10425
[15] Wang, W.L., Lin, T.I.: Automated learning of mixtures of factor analysis models with missing information. TEST **29**, 1098–1124 (2020)

DRAFT

# A multivariate ranking analysis on the employability of young adults

Rosa Arboretti[a], Elena Barzizza[b], Nicolò Biasetton[b], Riccardo Ceccato[b], Monica Fedeli[c], and Concetta Tino[c]

[a]Department of Civile, Environmental and Architectural Engineering, University of Padova, Via Marzolo, 9, Padova, 35131, Italy
[b]Department of Management Engineering, University of Padova, Stradella San Nicola, 3, Vicenza, 36100, Italy riccardo.ceccato.1@unipd.it
[c]Department of Philosophy, Sociology, Pedagogy and Applied Psychology, University of Padova, Via Beato Pellegrino, 28, Padova, 35137, Italy

DRAFT

**Abstract**

Unemployment among young adults has always been a relevant topic. To be employable, it is fundamental that young people learn how to plan and manage their careers and they need to be aware of their their skills and competences. This study focuses on the identification of the skills that students in the late stages of their university journey perceive as important and which of them they were able to develop. To this purpose data were collected through a survey and analyzed using the NonParametric Combination (NPC) methodology and a multivariate ranking procedure.

***Keywords:*** NPC, ranking, employability

## 1. Introduction

Unemployment among young adults has always been a relevant topic for the European Commission (2009) and the Organization for Economic Cooperation and Development (OECD, 2012). A misalignment does exist between young people's skills and the skills required by the labor market and in this the lack of dialogue between higher education institutions (HEIs) and businesses plays a central role.

To be employable, it is fundamental that young people learn how to plan and manage their careers (3), but also their personal and professional development (2). To this aim, they need to be aware of their self-perceived employability (6), their skills and competences, and knowledge of labor market demands and opportunities.

This study focuses on investigating the perception of higher education students on their skills and the existing gaps between the skills they feel they have developed and the skills they consider to be important in order to be employable. A survey was administered to students of the faculty of Engineering at the University of Padua in Italy, in order to collect data about students' labor market perceptions and their career planning and control. To analyze these data, we take advantage of the NonParametric Combination (NPC) methodology (5) and the multivariate ranking procedure by Arboretti et al. (1). These techniques are adopted in order to rank a number of skills according to their perceived importance, to their degree of development and according to both these aspects. This allows us also to identify eventual discrepancies between the skills the students were able to develop during their academic studies and the most important skills according to their students.

Section 2 is devoted to the introduction of the case study, the definition of the adopted methods and the presentation and discussion of the achieved results. In Section 3 conclusions are drawn.

## 2. Case study

A survey was administered to students of the faculty of Engineering at the University of Padua, selected people from the final year of undergraduate courses, from both years of Master's degree courses, and from the last three years of single-cycle courses, given that students at these stages of their university journey are more aware of their career goals and the challenges of university-to-work transition than the other undergraduates. The final sample size was equal to 2129.

Part of the questionnaire concerned the identification of the skills being perceived relevant for the labor market. The considered skills were: analytical thinking and innovation; active learning and strategies; resolving complex problems; critical thinking; creativity, originality and initiative; leadership; use of technologies; technological design and programming; resilience, stress tolerance, and flexibility; reasoning and ideation; emotional intelligence; persuasion and negotiation. Given this set of skills, the respondents were asked to which ones he considered relevant and should be promoted by university, so that a skill-specific binary variable (Yes: if this skill is perceived relevant - No: if this skill is not perceived relevant) was achieved for each skill.

For each skill, the respondent was also asked to choose the skills that he felt to have developed. Skill-specific binary variables were again collected.

The analysis of collected data had two main objectives:

- identify the main skills in terms of perceived relevance and development,
- identify eventual discrepancies between the skills the students were able to develop during their academic studies and the most important skills according to their students.

To this purpose, we adopted appropriate methods to rank a number of skills according to their perceived importance, to their degree of development and according to both these aspects.

DRAFT

### 2.1 Methods

The NonParametric Combination (NPC) (5) is a permutation-based methodology, widely adopted in multivariate scenarios thanks to the fact that it allows us to implicitly take into account any forms of dependency.

NPC does not make any strict assumption about data distribution, given that the key assumption of permutation tests is in relation to their exchangeability with respect to groups. Provided that the exchangeability condition is satisfied, they are conditional on a set of sufficient statistics (i.e. the data set itself) and independent of the likelihood model related to the unknown data distribution (5; 8; 4). Given these characteristics, this methodology can be applied for the analysis of multiple data types, such as continuous, ordinal and even mixed data. This makes it particularly suitable for the case study of interest, given that the variables of interest do not follow a normal distribution and are not even continuous, but they can assume only two different levels: "Yes" and "No".

NPC can be applied to both two-sample and $C$-sample problems even when the samples are dependent. In fact, by accurately decomposing the system of hypotheses, by defining the permutation test statistic and the permutation approach, it is even possible to compare multiple paired samples.

According to the NonParametric Combination methodology, the global system of hypotheses is initially decomposed into several sub-systems. Considering a generic multivariate equality in distribution problem:

$$\begin{cases} H_0 : \mathbf{F}^X = \mathbf{F}^Y \\ H_1 : \mathbf{F}^X \neq \mathbf{F}^Y. \end{cases}$$

where $\mathbf{F}^X$ and $\mathbf{F}^Y$ are the multivariate distribution functions of $\mathbf{X}_{n\times V}$ and $\mathbf{Y}_{n\times V}$, $n$ is the sample size and $V$ is the number of variables. According to Pesarin and Salmaso (5) it is possible to re-write this system of hypotheses as:

$$\begin{cases} H_0 : \bigcap_{v=1}^{V} H_{v0} = \bigcap_{v=1}^{V} [F_v^X = F_v^Y] \\ H_1 : \bigcup_{v=1}^{V} H_{v1} = \bigcup_{v=1}^{V} [F_v^X \neq F_v^Y]. \end{cases}$$

Each sub-system can then be addressed individually using appropriate permutation tests.

Let us now suppose that $\mathbf{X}$ and $\mathbf{Y}$ are paired samples and $\mathbf{Z}_{n\times V} = \mathbf{X} - \mathbf{Y}$, the algorithm continues as follows:

1. Compute the $V$-dimensional vector of test statistics $\mathbf{T} : \mathbf{T}^o = \mathbf{T}(\mathbf{Z})$, applying the chosen test statistic to each component of the multivariate $\mathbf{Z}$.
2. Randomly generate $s_i^* \sim \text{Bin}(1, 0.5), \forall i = 1, \ldots, n$, compute $\mathbf{Z}^* = \mathbf{Z} \cdot s^*$ and $\mathbf{T}^* = \mathbf{T}(\mathbf{Z}^*)$. Perform $B$ independent repetitions of this step (Conditional Monte Carlo - CMC). The CMC results $\{\mathbf{T}_b^*, b = 1, \ldots, B\}$ represent a random sampling from the permutation multivariate distribution of $\mathbf{T}$.
3. Use the CMC results to compute a consistent estimate of the marginal $p$-value $\lambda_v = \Pr\{T_v^* \geq T_v^o | \mathcal{Z}_{/\mathbf{Z}}\}$ for each test statistic $T_v$ as $\hat{\lambda}_v = \hat{L}_v(T_v^o | \mathcal{Z}_{/\mathbf{Z}}) = \frac{\sum_b I(T_{vb}^* \geq T_v^o)}{(B+1)}$, $v = 1, \ldots, V$.
4. Simulate the distribution of $\hat{\lambda}_v$ using $\hat{\lambda}_{vr}^* = \hat{L}_v(T_{vr}^* | \mathcal{Z}_{/\mathbf{Z}^*}) = \frac{\sum_b I(T_{vb}^* \geq T_{vr}^*)}{(B+1)}$, $v = 1, \ldots, V$.

After obtaining marginal p-values, the combination steps takes place:

5. By applying an appropriate combining function $\theta(\cdot)$, combine the $V$ observed $p$-values estimated by $\hat{\lambda}_v = \hat{L}_v(T_v^o | \mathcal{Z}_{/\mathbf{Z}})$ (see Table **??**) using $T''^o = \theta(\hat{\lambda}_1, \ldots, \hat{\lambda}_V)$ and the $V$-dimensional vectors $\{\hat{\lambda}_{vb}^* = \hat{L}_v(T_{vb}^* | \mathcal{Z}_{/\mathbf{Z}^*}), v = 1, \ldots, V\}$ using $T_b''^* = \theta(\hat{\lambda}_{1b}^*, \ldots, \hat{\lambda}_{Vb}^*)$.
6. Estimate the $p$-value of the combined test $T''$ as $\hat{\lambda}_\theta'' = \sum_b I(T_b''^* \geq T''^o)/B$.

The choice of combining function is a fundamental step in this procedure. A suitable non-degenerate combining function $\theta : [0, 1]^V \to \mathbb{R}^1$ is non-increasing in each argument, attains its supremum value (possibly not finite) when one or more of its arguments attain 0, and provides a finite critical value for a certain significance level $\alpha$ which is strictly smaller than the supremum value. In this study we considered Fisher's $\theta_F(\lambda) = -2 \cdot \sum_v \log(\lambda_v)$.

Another fundamental step is the choice of partial test statistics according to the nature of sub-hypotheses defined in the initial phase. In this study we have to deal with binary data, so we choose to simply use $T(Z_v) = \sum_{i=1}^n |Z_{iv}|$ for two-sided alternative hypotheses and $T(Z_v) = \sum_{i=1}^n Z_{iv}$ for one-sided ones.

## 2.2 Ranking of multivariate populations

When $C > 2$ groups need to be compared, multivariate ranking procedures can be extremely useful. The procedure proposed by Arboretti et al. (1) takes advantage of the aforementioned NPC methodology and allows us to rank $C$ populations $\mathbf{X}_c$, $c = 1, \ldots, C$, using the p-values achieved by conducting all the possible pairwise comparisons between groups:

$$\begin{cases} H_0(s,t) : \mathbf{F}^s = \mathbf{F}^t \\ H_1(s,t) : \mathbf{F}^s < \mathbf{F}^t. \end{cases}$$

Let $\mathbf{\Lambda}$ denote the $C \times C$ matrix containing the p-values $\hat{\boldsymbol{\lambda}}^{s,t}$, $\forall s, t = 1, \ldots, C$ and $s \neq t$ related to the one-sided comparison between the samples $\mathbf{X}_s$ and $\mathbf{X}_t$. A multiplicity correction is firstly applied to $\mathbf{\Lambda}$ by using one of the many suitable methods proposed in the literature, such as the Bonferroni-Holm-Shaffer correction (7). A matrix of adjusted p-values $\mathbf{\Lambda}_{adj}$ is therefore achieved. A significance matrix $\mathbf{S}$ is then created to keep track of the significant comparisons, where:

$$\begin{cases} S_{s,t} = 1, & \text{if } \hat{\boldsymbol{\lambda}}_{adj}^{s,t} \leq \alpha/2 \\ S_{s,t} = 0, & \text{otherwise} \end{cases}$$

and $\alpha$ is the desired significance level.

For $s = 1, \ldots, C$, we compute the upward rank estimate as $\left\{r_u^s = 1 + \#\left[(C - \sum_{t=1}^C S_{s,t}) > (C - \sum_{t=1}^C S_{s',t})\right], s' = 1, \ldots, C, s' \neq s\right\}$, where $\#$ means number of times. Then, for $t = 1, \ldots, C$

DRAFT

we calculate the downward rank estimate as $\left\{r_d^t = 1 + \sum_{s=1}^{C} S_{s,t}\right\}$. The vector $\mathbf{r}$ of ranking estimates is finally achieved as $\left\{r^c = 1 + \#\left[\frac{(r_u^c + r_d^c)}{2} > \frac{(r_u^t + r_d^t)}{2}\right], t = 1, \ldots, C, c \neq t\right\}, c = 1, \ldots, C$.

This procedure can be applied on both the partial and global p-values obtained using the aforementioned $NPC$ methodology, so that we can decide to focus on a single component or multiple components of interest.

## 2.3 Results and discussion

A descriptive analysis was initially conducted. looking at Figure 1, a few interesting insights were achieved.

The five skills which are believed to be most relevant are use of technologies, analytical thinking and innovation, creativity, originality and initiative, critical thinking, and active learning and strategies, all with a percentage of Yes among the respondents greater than 40%.

The five skills which most of the respondents believe they posses are critical thinking, resilience, stress tolerance, and flexibility, active learning and strategies, use of technologies, and reasoning and ideation. With the only exception of analytical thinking and innovation and resolving complex problems, all the remaining skills were chosen by less than the 25% of the students.

For some of the observed skills, there is a substantial misalignment between their perceived importance and their development:

- creativity, originality and initiative: 43.4% of the respondents consider it relevant, but only 20.9% of the students stated that they have developed it;
- critical thinking: 42.4% of the students answered it is relevant and 59.2% of the respondents feel that they have developed it;
- resilience, stress tolerance, and flexibility: 24.1% of the respondents consider it relevant and 59.2% of the students stated that they have developed it;
- technological design and programming: 35% of the students answered it is relevant and 24.9% of the respondents believe that they have developed it;
- persuasion and negotiation: 14.6% of the respondents consider it relevant and 24.9% of the students stated that they have developed it;
- analytical thinking and innovation: about 44% of respondents consider this skill important and about 37% of them feel they have developed this skill.

For the ranking analysis we then set the significance level equal to 0.05 and the number of permutations $B$ equal to 2000. The achieved rankings are reported in Table 2.3.

Use of technologies and analytical thinking and innovation are the skills which are perceived as most relevant and stochastically dominate the others. On the other hand, in terms of development, the percentage of Yes for critical thinking is significantly larger than the percentage of Yes for the others. We can see that persuasion and negotiation is at the same time the least relevant and the least developed skill.

The achieve rankings highlight the existence of discrepancies between relevant and developed capabilities, as already shown by the descriptive analysis. For example, creativity, originality and initiative is among the most important skills (with a ranking position equal to 3), but at the same time it is among the least developed ones (with a ranking position equal to 10).

Considering the global ranking, we can finally see that use of technologies and critical thinking are the skills that are at the same time most important and most developed among the considered ones, while the opposite goes for persuasion and negotiation, emotional intelligence, and leadership.

# 3. Conclusions

This study focuses on the analysis of data related to a particularly relevant topic: employability of young adults. A survey was conducted in order to collect information on how students in the late stages of their university journey perceive the labor market and their career planning and control. In particular,

DRAFT

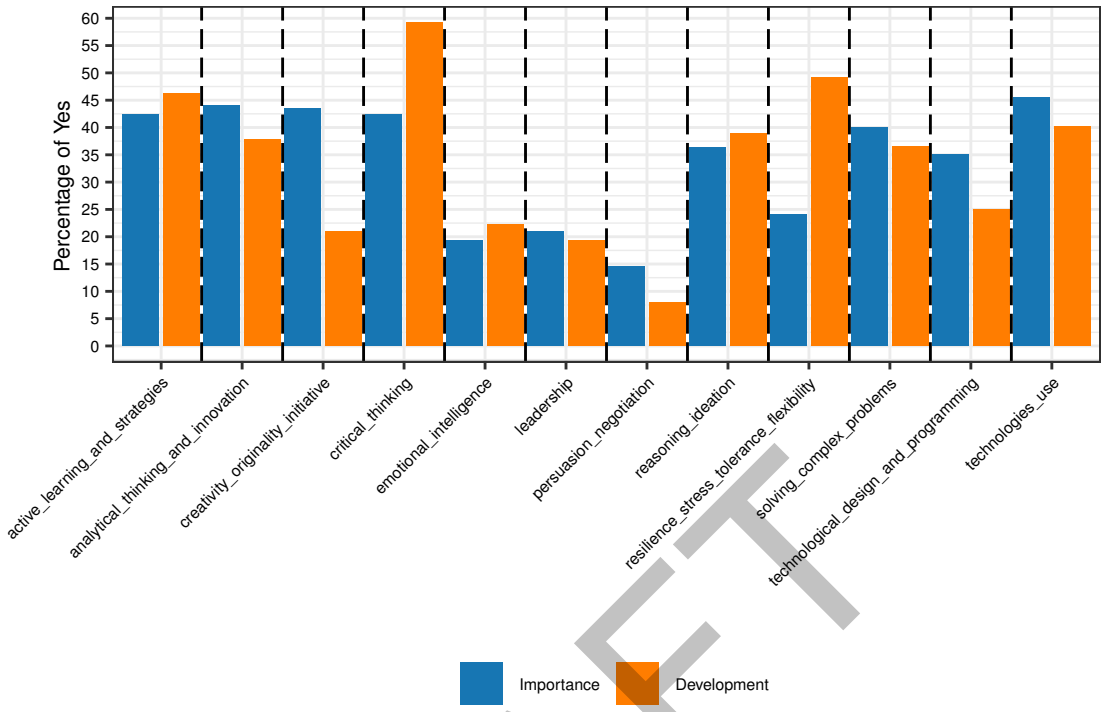



Figure 1: Descriptive analysis results

DRAFT

Table 1: Partial and global rankings

| Skill | Importance | Development | Global |
|---|---|---|---|
| active_learning_and_strategies | 3 | 3 | 3 |
| analytical_thinking_and_innovation | 1 | 6 | 4 |
| creativity_originality_initiative | 3 | 10 | 9 |
| critical_thinking | 3 | 1 | 1 |
| emotional_intelligence | 11 | 10 | 11 |
| leadership | 10 | 9 | 9 |
| persuasion_negotiation | 12 | 12 | 12 |
| reasoning_ideation | 7 | 4 | 8 |
| resilience_stress_tolerance_flexibility | 9 | 2 | 6 |
| solving_complex_problems | 6 | 6 | 4 |
| technological_design_and_programming | 8 | 8 | 6 |
| technologies_use | 1 | 4 | 1 |

we wanted to identify the skills that are perceived as most important for employability and the ones that the students believe they have developed the most, but also identify eventual misalignment between this aspects.

The adoption of the NonParametric Combination (NPC) methodology and the multivariate ranking procedure by Arboretti et al. (1), allowed us to see that the use of technologies and critical thinking are the skills that are at the same time most important and most developed among the considered ones. It also showed that creativity, originality and initiative is a skill that is perceived as really important, but it rarely developed in students at the end of their university journey.

## References


[1] Arboretti, R., Bonnini, S., Corain, L., and Salmaso, L. (2014). A permutation approach for ranking of multivariate populations. *Journal of Multivariate Analysis*, 132:39–57.

[2] Buckley, P. & Casson, M. A theory of cooperation in international business. *The Multinational Enterprise Revisited*. pp. 41-67 (2010)

[3] Fugate, M. & Kinicki, A. A dispositional approach to employability: Development of a measure and test of implications for employee reactions to organizational change. *Journal Of Occupational And Organizational Psychology*. **81**, 503-527 (2008)

[4] Good, P. (2000). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, New York, NY.

[5] Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. Wiley.

[6] Rothwell, A., Jewell, S. & Hardie, M. Self-perceived employability: Investigating the responses of post-graduate students. *Journal Of Vocational Behavior*. **75**, 152-161 (2009)

[7] Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831.

[8] Sierra, V., Solanas, A., and Vicenç, Q. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education*, 73:140–160.


DRAFT

# Analysis of the Gender Pay Gap in the Italian Labour Market

Giulia Cappelletti[a], Daniele Toninelli[a]

[a] University of Bergamo, via dei Caniana 2, 24127, Bergamo (Italy);
giuliacappelletti@outlook.it, daniele.toninelli@unibg.it

**Abstract**

Gender inequality is a complex and multidimensional phenomenon. In recent decades, its measurement has become a focus of interest not only for academic researchers, but also for political institutions and policy makers. However, due to the epidemic from Covid-19, the advancement of gender equality has faced multiple setbacks, especially in the labour market. The objective of this study is to support the implementation of policy makers' strategies and to contribute to the existing literature by shedding light on the importance of quantifying the impact of certain key characteristics related to the workplace and to the individual on the gender pay gap. In particular, we focus on the Italian job market, exploring the phenomenon using updated official statistics data coming from Istat by means of a regression analysis. This study shows how workers are treated differently on the basis of the demographic group to which they belong, providing also an in-depth knowledge of key determinants behind women discrimination in the labour market.

**Key words:** gender pay gap, gender equality, discrimination, labour market, Covid-19, linear regression models

DRAFT

## 1. Introduction

The gender pay gap is commonly defined as the percentage difference in average gross hourly earnings between women and men (EIGE, 2019). At the European level, the principle of equal pay for men and women was, first, enforced in the Treaty of Rome (1957). In the following years, the European Union commitment has been steadily renewed through a series of strategies and action plans, such as: i) the Roadmap for Equality between Women and Men (2006-2010); ii) the Gender Equality Strategy (2010-2015); iii) the European Pact for Gender Equality (2011-2020); iv) the Strategic Commitment to Gender Equality (2016-2019); v) the new Gender Equality Strategy (2020-2025). This increasing interest and the need of measuring the targets' achievement level show how important became developing a reliable index to assess this phenomenon. Over the past decades, academic researchers and international organizations have gradually developed a variety of indicators to measure gender inequality. Indeed, the gender pay gap measurement is not a straightforward task, as the results heavily depend upon peculiar background choices. The main options in the gender pay gap measurement relate to its estimate according to an unadjusted or to an adjusted form (Klammer *et al.,* 2018). Unlike the unadjusted gender pay gap – which is simply estimated as the difference between the average gross hourly earnings of men and women – the adjusted gender pay gap takes into account all possible factors that can play a role in explaining the gap, such as differences in terms of education level, employment sector, job title, working hours, etc. (Klammer *et al*., 2018). Thus, the adjusted gender pay gap makes it possible to measure the part of the gap related to differences in the characteristics of male and female workers, distinguishing this from the part that, instead, can be attributed to discrimination (Klammer *et al*., 2018).

To date, the most commonly used index is the Global Gender Gap Index. This was introduced by the World Economic Forum (2006) in order to benchmark progress towards gender parity and to compare different countries' gender gap. The index covers four dimensions: Economic Participation and Opportunities, Educational Attainment, Health and Survival, and Political Empowerment (World Economic Forum, 2022). Each of these dimensions is ranked on a 0 to 100 scale, which can be interpreted as the distance covered towards parity (i.e., the percentage of the gender gap that has been closed). In 2022, the overall Global Gender Gap Index was equal to 68.1%. This means that the remaining gap to close stands at 31.9%. This level represents an improvement of 0.4 percentage points, if compared to the previous index estimate (World Economic Forum, 2022). Such a slight progress is not even sufficient to compensate for the generational loss which occurred between 2020 and 2021.

In particular, the gender gap referred to the dimension of Economic Participation and Opportunities is the second largest of the four key gaps tracked by the index. This subindex includes three concepts: the participation gap, the remuneration gap and the advancement gap. The participation gap originates from the difference between women and men in labour-force participation rates. The remuneration gap evaluates the female-to-male difference in earned income. The advancement gap is estimated as the ratio of women to men among legislators, senior officials and managers, and as the ratio of women to men among technical and professional workers (World Economic Forum, 2022). In 2022 only 60.3% of this gap has been closed. Moreover, estimates say that it will take another 151 years to close the gap (World Economic Forum, 2022). Among all considered regions, South Asia is the farthest from achieving gender equality (only 35.7% of the gap has been close so far), whereas North America leads all regions, with 76.9% of gender gap closed (World Economic Forum, 2022). In this scenario, Italy is one of the least performing countries: in the general ranking, it is placed at the 63$^{rd}$ position (on 146), but when it comes to the gender gap in labour market, Italy slips to the 110$^{th}$ place (World Economic Forum, 2022).

The 2022 Report highlights an overall stalemate of gender parity, partly due to the impact of the Covid-19 pandemic. This is especially true for the female labour market. Despite the most recent studies (e.g., McKinsey Global Institute, 2020) showed that both men and women were hit hard during and after the pandemic, the most negative impact was suffered by women. This leaded to a setback in the advancement of gender equality, in terms of employment and labour force participation, work structure and intensification of childcare and domestic education activities (International Labour Organization, 2021).

Despite the epidemic from Covid-19 negative impact, signs of good omen for the future are certainly not lacking, nowadays, in Europe: in fact, during the pandemic years, many important decisions were taken by women, such as the President of the European Commission, Ursula Von Der Leyen, and the President of the European Central Bank, Christine Lagarde. In Italy, within the National Plan for Recovery and Resilience (PNRR), approved in July 2021, the need to close the gender gap in the labour market has been strongly marked. Focused on inclusion and social cohesion, Mission V enforces targeted policies and action plans to fight gender discrimination and towards women's empowerment in the five-year period 2021-2026. For this purpose, with the aim of sustaining female entrepreneurship, the Italian Ministry of Economic Development has established the *Women's Enterprise Fund*. The Ministry also introduced the *Gender Equality Certification*, issued starting from January 1$^{st}$, 2022, to all businesses having adopted targeted policies and measures in order to reduce the gender gap in terms of employment, equal pay and equal opportunities, according to 33 specific key performance indicators (KPIs).

To date, most of the studies have focused primarily on the quantification and isolation of the discriminatory component of the gender pay gap. This study aims at showing not only how workers are treated differently on the basis of the demographic group to which they belong, but provides also an in-depth knowledge of key determinants behind women discrimination in the labour market.

## 2. Literature review

Since the late 1950s, several sociological and economic theories have been advanced to explain the underlying causes of the presence of gender discrimination in payment. One of the first contributions, by Becker and Schulz (1960), highlights the relationship between human capital, individual productivity and economic growth. Human capital can be defined as the set of knowledge, skills and abilities possessed by the workforce and offered to the market in exchange for remuneration (Becker, 1964). In this

DRAFT

sense, human capital can be considered as a resource subjected to accumulation processes through investment in education, training and experience. The level of education, in particular, is a key factor in assessing the wage differential. This is crucial to access more lucrative career paths (Becker, 1964), as well as to affect employment levels. Thus, Becker's theory of human capital links the gender pay gap to differences in terms of investment in human capital.

Contrarily, the segregation theory justifies the gender pay gap starting from the employment (Geiler and Renneboog, 2015). In particular, two forms of occupational segregation are distinguished in the economic literature: horizontal (or sectoral) segregation and vertical segregation. The first refers to the concentration of female employment in certain productive sectors and professions; the second is linked to the gender uneven distribution in the different hierarchical levels of a same occupation (Geiler and Renneboog, 2015). This phenomenon is the result of socio-cultural stereotypes that, together with the organisational rigidities of enterprises, lead to more or less explicit forms of discrimination (or sometimes exclusion) against female labour supply. All this perpetuated the gender characterisation of certain occupations. This concept is also closely related to the phenomena of glass ceilings and sticky floors (Booth *et al*., 2003; Haslam *et al*., 2009).

The comparable worth theory was first proposed by Treiman and Hartmann (1981). The authors studied the issue of the devaluation of women's work. According to their work, the gender pay gap stems from the fact that society, due to socio-cultural mechanisms and stereotypes, ascribes less value to female-dominated industries (Treiman and Hartmann, 1981). In other words, occupations with a higher proportion of women receive lower wages, as women positions are considered socially less valuable than men ones and, therefore, less deserving of financial recognition.

The crowding theory (Bergmann, 1974) analyses the effects of women's limited access to certain occupations and professions on their pay level. In fact, the high concentration of workers in a narrow segment of the market leads to the emergence of an artificially high labour supply. This, accordingly to the market principles, results in lower wages than the ones expected in case of a more equal division of labour among the different segments (Bergmann, 1974). In other words, due to discriminatory mechanisms, women cannot move freely in the labour market. Hence, female labour supply is higher in certain occupations rather than others, resulting in a reduction in the equilibrium wage (Bergmann, 1974).

Another cause of the gender pay gap is wage bargaining. It is the main tool for regulating industrial relations and issues concerning pay aspects and working conditions. Wage bargaining takes place at two levels: as individual and as collective bargaining. The economic literature showed that the degree of centralisation of the collective bargaining system has a significant impact on the gender pay gap (Calmfors and Driffill, 1988).

In conclusion, the literature confirms the existence of a gender gap (in labour force and in income) that may vary in size and nature depending upon the different determinants considered. Consequently, it becomes clear how important is to reliably assess the level of women discrimination in order to identify its key determinants and in order to set and implement policies helpful in reducing this phenomenon and to identify tools capable to assess policies efficacy and the linked progress.

DRAFT

## 3. Empirical analysis

This study contributes to the existing literature by shedding light on the importance of certain characteristics related to the job position and/or to the individual worker in causing the gender pay gap. We focus on the Italian context and, by means of identifying key factors, we aim at providing support to policy makers in implementing targeted strategies and policies in order to reduce gender pay differentials and to promote gender equality at all levels.

### 3.1 Data and methodology

The analysis is based on the latest release of Istat data on "Labour and Wages" (2014-2019) (link: https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati/statbase). In particular, we analyse the gross hourly earnings for men and women with relation to the following variables: i) educational level, ii) type of contract, iii) job title, iv) company size, v) industry. First, we derive the gender wage differential. This is computed as the ratio of the percentage difference between the men gross hourly earnings and the women ones on the average hourly earnings of men. Then, the data were split into two separate

datasets. The first one was used to conduct exploratory analyses, focusing on the wage differential. It was also used to estimate a benchmark model, at the national level, and other sub-models related to each of the Italian macro-area set by Istat (North-West, North-East, Centre, South, Islands). The second dataset was the base to estimate OLS linear models aimed at investigating the linkage between the gender wage differential, the worker position and individual characteristics. Alphanumeric variables were recoded into dummies as follows:

(a) Educational level: 1 = the worker has at least a degree, 0 = otherwise.
(b) Contract type: 1 = the worker has a permanent contract, 0 = otherwise.
(c) Job title: 1 = the worker is qualified as manager or employee, 0 = otherwise.
(d) Company size: 1 = large enterprise, 0 = otherwise.
(e) Industry: 1 = female-dominated sector, 0 = otherwise.

Table 1 shows the main works developed on the gender pay gap focusing on the Italian context: they guided us in the selection of data sources and of the variables used in our regression analysis.

Table 1: List of empirical studies on gender pay gap carried out in the Italian literature

| **Authors** | **Data** | **Methodology** |
|---|---|---|
| Addis, Waldmann (1986) | SHIW[1] | Pooled OLS |
| Flabbi (2001) | SHIW | Oaxaca-Blinder |
| Beblo et al. (2003) | ECHP[2] | Oaxaca-Blinder |
| Pissarides et al. (2005) | ECHP | Oaxaca-Blinder |
| Olivetti, Petrongolo (2005) | ECHP | Heckman correction |
| Rustichelli (2005) | INPS[3] | Random Effects Model |
| Mundo, Rustichelli (2005) | INPS-ISFOL | Pooled OLS, Oaxaca-Blinder |
| Addabbo, Favaro (2007) | ECHP | Oaxaca-Blinder |
| Centra, Cutillo (2009) | ISFOL[4] | Oaxaca-Blinder |
| Zizza (2013) | SHIW | Pooled OLS |

[1] Survey on Household Income and Wealth.
[2] European Community Household Panel.
[3] Istituto Nazionale Previdenza Sociale.
[4] Istituto per lo Sviluppo della Formazione professionale dei Lavoratori.

DRAFT

### 3.2 Results

Table 2 summarizes the main descriptive statistics at the national level about the gender pay gap by some key variables (educational level, contract type, job title, company size and industry). The average pay differential in earnings between men and women increases with an higher level of education (9.35% among non-graduates, 19.86% for graduates), with more stable job contract (3.66% among fixed-term contracts, 9.68% for permanent contracts), with a higher qualification level (10.5% among unskilled workers, 23.36% for employees and managers), according to the company size (9.14% among SME, 16.86% for large enterprises) and when considering male-dominated industries (Financial and insurance fields 20.19%; Professional, scientific, and technical fields 13.74%), if compared to female-dominated industries (Education 9.78%; Health

Table 2: Gender pay gap (percentages) by key variables (national level)

| **Educational level** | **N. obs.** | **Mean** | **Std. dev.** | **Min.** | **Max.** |
|---|---|---|---|---|---|
| High school | 30 | 9,35 | 2,59 | 5,65 | 13,41 |
| Degree | 30 | 19,86 | 3,99 | 13,51 | 26,63 |
| **Contract type** | | | | | |
| Fixed-term | 30 | 3,66 | 1,60 | 1,30 | 6,77 |
| Permanent | 30 | 9,68 | 2,49 | 5,40 | 13,04 |
| **Job title** | | | | | |
| Worker | 30 | 10,50 | 1,29 | 8,37 | 13,30 |
| Employee, Executive | 30 | 23,36 | 4,29 | 16,00 | 29,63 |
| **Company size** | | | | | |
| Small-medium enterprise | 30 | 9,14 | 1,34 | 6,75 | 11,82 |
| Large enterprise | 30 | 16,86 | 1,88 | 13,60 | 20,41 |

| **Industry** | | | | | |
|---|---|---|---|---|---|
| Financial and insurance fields | 30 | 20,19 | 1,57 | 17,12 | 23,43 |
| Scientific and technical fields | 30 | 13,74 | 6,67 | 5,91 | 25,96 |
| Education | 30 | 9,78 | 5,88 | 0,53 | 23,84 |
| Health and social assistance | 30 | 4,59 | 1,81 | 1,71 | 8,88 |

and social assistance 4.59%). Table 3 shows simple linear regression models estimates obtained at the national level to study the relationship between the gender wage differential (dependent variable) and, as regressors, the key variables related to the worker position and to the individual previously listed, recoded as dummies. As the gross hourly earnings of men and women provided by Istat is specifically computed each time in relation to the variable considered, it was not possible to estimate one unique multiple regression model. Thus, we estimate five simple regression models.

Table 3: Simple linear regression models for gender pay gap (%) vs key variables (national level).

| | **Iterc./Regr. (key var.)** | **Coeff.** | **Std. Err.** | **t-value** | **Pr > \|t\|** | **R²** |
|---|---|---|---|---|---|---|
| MOD. 1 | Intercept | 9.35 | 0.61 | 15.21 | <.0001* | 0.72 |
| | Educational level | 10.51 | 0.87 | 12.08 | <.0001* | |
| MOD. 2 | Intercept | 3.66 | 0.38 | 9.56 | <.0001* | 0.68 |
| | Contract type | 6.02 | 0.54 | 11.13 | <.0001* | |
| MOD. 3 | Intercept | 10.5 | 0.58 | 18.15 | <.0001* | 0.81 |
| | Job title | 12.86 | 0.82 | 15.73 | <.0001* | |
| MOD. 4 | Intercept | 9.14 | 0.3 | 30.61 | <.0001* | 0.85 |
| | Company size | 7.72 | 0.42 | 18.27 | <.0001* | |
| MOD. 5 | Intercept | 16.97 | 0.7 | 24.17 | <.0001* | 0.45 |
| | Industry | -9.78 | 0.99 | -9.85 | <.0001* | |

All models obtained in Table 3 show a statistically significant relationship between the gender pay differential and the key dummy variables related to the workplace and the individual (for $\alpha = 0.01$). In fact, in each model the *p*-value of the independent variable (<,0001) is lower than alpha. Moreover, the estimated models, despite including one regressor only, show a good predictive capacity: the $R^2$ ranges between 0.45 (considering, as regressor, *Industry*) to over 0.8, including *Job title* (0.81), and *Company size* (0.85) as regressors. The empirical findings confirm the existence of a statistically significant relationship between the gender pay differential and all the considered key variables. As few studies have been carried out in the literature to this regard, the results obtained in this work might represent a starting point for future studies to deepen the impact of these variables on the gender pay gap, eventually estimating multiple regression models, when possible. Moreover, since the analysis was carried out on data covering a relatively short time period of six years (2014-2019), it would be interesting to repeat the analysis by integrating the data that will be available for more recent years (2020, 2021 and 2022), in order to deepen the impact of the Covid-19 pandemic on the labour market and to assess how the labour market itself has faced the recovery process.

## 4. Conclusions

The results obtained in this study are useful for policy makers, shading light on a phenomenon still affecting the job market in Italy. In particular, the level of education is a key factor in the assessment of the wage differential, as it affects not only employment levels but it is also crucial to access more lucrative career paths (Becker, 1964). Moreover, it is observed that in Italy the average gender pay differential increases as the level of contractual framing increases, resulting in phenomena such as glass-ceiling and sticky floor that cause women to be under-framed by being stuck in unprofitable career paths (Geiler and Renneboog, 2015). The empirical results obtained in this study also confirm crowding theory (Bergmann, 1974) and comparable worth theory (Treiman and Hartmann, 1981), which consider the employment sector as one of the primary causes of the gender pay gap. At the same time, however, the results obtained highlight the current need to: i) increase women labour force in all industries and, in particular, in those occupations that traditionally have been a male prerogative; ii) encourage female entrepreneurship and career advancement in managerial roles and top positions; iii) level out differences in human

DRAFT

capital, especially improving the education level, and encouraging women to become more involved in STEM fields; iv) promote the continuous participation of women in the labour market through stable career paths and contracts; v) develop binding national control systems to monitor the gender gap in employment, pay and equal opportunities in each company on a regular basis.

However, it is necessary to highlight how acting on the key variables considered in this study may not be sufficient to fill the gender pay gap. In fact, results obtained in this study highlight a part of gender pay gap that remains unexplained and that can be presumably attributed to a discrimination effect. Overcoming gender discrimination in the labour market, indeed, requires a deep cultural change, based on the deconstruction of all stereotypes of cultural, social and psychological nature that for centuries have dictated a categorization of male and female roles in our society. Hence, social dialogue can be a real keystone, in reducing gender inequalities and in defining a future of gender equality. The full achievement of gender equality should not be an exclusively female issue, but it should be a priority and a moral imperative for every truly modern, inclusive and egalitarian society as a whole.

## References


[1] Addabbo, T., Favaro, D.: Differenziali salariali per sesso in Italia. Problemi di stima ed evidenze empiriche, in Rustichelli E. (a cura di), "Esiste un differenziale retributivo di genere in Italia", pp. 199-237, ISFOL (2007).

[2] Addis, E., Waldmann, R: Struttura salariale e differenziale per sesso in Italia, Economia e Lavoro, 1, pp. 87-103 (1996).

[3] Azcue, X., Krishnan, M., Madgavkar, A., Mahajan, D., White, O.: COVID-19 and gender equality: Countering the regressive effects. McKinsey Global Institute (2020).

[4] Beblo, M., Beninger, D., Heinze, A., Laisney, F.: Measuring selectivity corrected gender wage gaps in the EU", ZEW Discussion paper N. 03-74 (2003).

[5] Becker, G.: Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. Columbia University Press, New York (1964).

[6] Bergmann, B. R.: Occupational segregation, wages and profits when employers discriminate by race or sex. Eastern economic journal, 1(2), pp. 103-110 (1974).

[7] Calmfors, L., Driffil, J.: Bargaining structure, corporatism and macroeconomic performance. Economic policy, 3(6), pp. 13-61 (1988).

[8] Centra, M., Cutillo, A.: Differenziale salariale di genere e lavori tipicamente femminili, Studi Isfol, 2 (2009).

[9] Klammer, U., Klenner, C., Lillemeier, S.: Comparable Worth: Arbeitsbewertungen als blinder Fleck in der Ursachenanalyse des Gender Pay Gaps? WSI Study, 14 (2018).

[10] International Labour Organization, ILO Monitor: COVID-19 and the world of work. Eighth edition. Available on: https://www.ilo.org/rome/pubblicazioni/WCMS_824092/lang--en/index.htm (Last consultation: 26/07/2022).

[11] Mundo, A., Rustichelli, E.: Differenziali retributivi di genere: evidenze dai dati di fonte amministrativa, in Rustichelli E. (a cura di), "Esiste un differenziale retributivo di genere in Italia", ISFOL (2007).

[12] Olivetti, C., Petrongolo, B.: Unequal pay or unequal employment? A cross country analysis of gender gaps", CEP Discussion paper, 711 (2005).

[13] Pissarides, C., Garibaldi, P., Olivetti, C., Petrongolo, B., Wasmer, E.: Wage gaps, in Boeri, T., Del Boca, D., Pissarides, C. (eds) "Women at work. An economic perspective", Oxford University Press (2005).

[14] Rustichelli, E.: I differenziali retributivi di genere", in Battistoni L. (a cura di) "I numeri delle donne", Italia Lavoro (2005).

[15] Treiman, D. J., Hartmann, H. I.: Women, work and Wages: Equal pay for jobs of Equal value. Washington DC: National Accademy Press (1981).

[16] Zizza, R.: The gender wage gap in Italy, Questioni di economia e finanza, 172 (2013).

[17] World Economic Forum, The Future of Jobs Report 2020. Available on: https://www.weforum.org/reports/the-future-of-jobs-report-2020/ (Last consultation: 26/07/2022).

[18] World Economic Forum, The Global Gender Gap Report 2022. Available on: https://www.weforum.org/reports/global-gender-gap-report-2022/ (Last consultation: 26/07/2022).


DRAFT

# Evaluating the effect of home-based working employing causal Bayesian networks and potential outcomes

Lorenzo Giammei[a]

[a]University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1 Milan; lorenzo.giammei@unimb.it

**Abstract**

Covid-19 generated an unprecedented shock on the Italian economy, which severely affected firm performance. This work focuses on estimating the causal effect of implementing home-based working (HBW) after the pandemic outbreak on firms' expected revenues. The analysis uses a unique firm-level dataset, which captures a rich set of features before and after the spread of the virus. Causal effect estimation is performed implementing an integrated approach that merges Causal Graphs and Potential Outcomes frameworks. The results are consistent with the fact that HBW equips firms with greater flexibility and helps contain productivity decreases in Covid times.

***Keywords:*** Causal Bayesian Networks, Potential Outcomes, Home-based Working

DRAFT

## 1. Introduction

The outbreak of Covid-19 in March 2020 had unprecedented consequences on the Italian economy. As the virus spread, consumer spending dropped, and lockdown policies forced many firms to temporarily cease their activity, thus generating both a demand and a supply shock. As soon as the economic consequences of the covid outbreak became clear, firms tried to do everything possible to minimize losses.

This work focuses on the implementation of home-based working (HBW), one of the key firms' countermeasures to the pandemic. The implications of switching to HBW have been thoroughly studied over the past years and its related literature has spiked in covid times. The benefits of home working on employees performance have been studied in (4). On the other hand, evidence from workers who switched to home working in covid times suggests that being far from the workplace for a prolonged period can negatively affect mental health (5). (3) use firm-level surveys to investigate the spread of home-based working during the pandemic. The authors find out that industries with more educated workers are associated with a higher rate of remote working and perceive a lower productivity loss associated with this kind of work. In addition, about 40% of interviewed firms declare that at least 40% of their workers that switched to homeworking will continue doing so even after the crisis, and this represents a strong indicator of the persistence of the phenomenon. This work contributes to the fast-growing literature of home-based working by evaluating the effect of implementing home working in covid times on future expected firm revenues. A rigorous causal evaluation of this kind seems to be missing in the literature and could provide a quantifiable measure of the impact of enabling employees to work from home.

## 2. Data

The analysis uses a unique firm-level dataset provided by MET, a research centre based in Rome, which conducts one of the most comprehensive surveys on the Italian manufacturing and production service sectors. The dataset originates from merging two different MET surveys over the same panel of firms. The same data source has already been employed to study the economic effects of the Covid-19 shock in (2).

The first survey is the 2019 wave of the MET survey on the italian industrial system. The questionnaire covers a vast group of firm features such as structure, performance and strategies. Almost 24000 firms were interviewed according to their size, sector, and area to obtain a representative sample of the Italian manufacturing and production services population. Reaction to Covid-19 was then measured with another questionnaire between March 24 and April 7, 2020, administered to the 24000 respondents of the 2019 MET survey. The exceptional timing thus produces two snapshots of the same group of firms, just before and after the spread of the pandemic. The answers to both surveys have been merged to obtain a final dataset of 7800 respondents.

The treatment variable originates from the post-covid questionnaire and is a binary variable defining if a firm has implemented home-based working for a portion of their employees right after the lockdown policies introduction. The outcome variable describes post-covid expectations towards future variation in revenues with respect to past revenues. The variable derives from the post-covid survey and can take four different modalities, which identify increase, stability, decrease or strong decrease in expected revenues.

## 3. Methodological background

Causal graphs, also known as causal Bayesian networks, are models providing a clear representation of causal problems and a set of tools to derive causal estimates (8). A graph $G = (\mathbf{X}, \mathbf{E})$ is a collection of nodes $\mathbf{X}$ and edges $\mathbf{E}$. When an edge goes out from a node into another is called a directed edge, if there is no such orientation the edge is undirected. A graph that contains only directed edges is a directed graph. Given a graph $G$ with ensemble of nodes $\mathbf{X}$ and two nodes $X_i$ and $X_j$ belonging to $\mathbf{X}$, any sequence of edges which connects $X_i$ and $X_j$, regardless of their direction, is called a path. If every edge of the path is directed and has the same orientation along the path, then it is called a directed path. A directed path which begins and ends with the same node is a cycle. If a directed graph does not contain cycles then it is a directed acyclic graph (DAG). In the context of this work, the nodes of the DAG represent random variables and edges describe the causal relations between them. For a given DAG $G$, the structure of the graph allows factorizing the joint probability distribution of its nodes $(X_1, ..., X_n)$ as follows

$$P(X_1, ..., X_n) = \prod_i P(X_i|pa_i) \tag{1}$$

where $pa_i$ is the set of nodes that have an outgoing edge pointing to $X_i$, according to the graph.

To answer causal queries, we are interested in studying how the model would react to an $intervention$ on one or more variables. (9) introduces the notation $do(X_i = x_i)$ to denote that a variable $X_i$ is set to the value $x_i$ through an intervention. The quantity $P(X_j|do(X_i = x_i))$ represents then the distribution of $X_j$ given that $X_i$ is forced to take value $x_i$, while $P(X_j|X_i = x_i)$ describes the distribution of $X_j$ given that we observe $X_i$ take value $x_i$. Interventional quantities can be used to estimate causal effects, that can be expressed as comparison of interventional distribution summary statistics. A common way to represent causal effects is the *average treatment effect* (ATE) (7).

In observational studies, the interventional distribution $P(X_j|do(X_i = x_i))$ is not directly measured. In order to express this distribution through observational quantities, (8) introduces a graphical test called the back-door criterion. In particular, given a graph $G$ with ensemble of nodes $\mathbf{X}$, a treatment $T$ and an outcome $Y$ belonging to $\mathbf{X}$, if a set $\mathbf{S} \subset \mathbf{X}$ satisfies the back-door criterion, then interventional distributions can be expressed in observational terms:

$$P(Y|do(T = t)) = \sum_{\boldsymbol{S}} P(Y|\mathbf{S} = \mathbf{s}; T = t)P(\mathbf{S} = \mathbf{s}) \tag{2}$$

DRAFT

A set **S** that satisfies the assumptions is called a *sufficient adjustment set*. As shown in (2), given a known causal graph, an adjustment set obtained through the back-door criterion allows the estimation of unbiased interventional distributions. Note that the proposed methods strongly rely on the structure of the graph, which is often partially or entirely unknown when dealing with real problems. However, in this case, the causal graph can be recovered from a dataset containing the variables of interest through structural learning algorithms (12).

Potential outcomes (11) are an alternative framework to deal with causality. This approach is widely used in economics and allows estimating causal effects from experiments and some specific observational contexts. Let us consider an outcome variable $Y$ and a treatment $T$. We denote $T = 0$ and $T = 1$, respectively, the treated and the not treated condition. Then we can define $Y_i(T = 1)$ the potential outcome we would have observed if unit $i$ had received the treatment and as $Y_i(T = 0)$ the potential outcome we would have observed if the same unit had not received the treatment. The methods that belong to the framework usually require the *unconfoundedness* assumption, implying that the treatment assignment mechanism is conditionally independent of the potential outcomes given the covariates.

Here *full matching* (6) to estimate the treatment effect will be used. This technique groups all the units into a series of matched subclasses, containing at least one treated and one control unit. Similar units are gathered in the same subclass and its size depends on the number of comparable units: the more the available similar units, the larger the generated subclass and vice-versa. The similarity between units $i$ and $j$ is described by discrepancy measure $\delta_{ij}$, which is usually calculated as a difference of distance measures, such as a propensity score. Full matching has been chosen because it allows the estimation of the ATE, and thus the results of the matching procedure are coherent with the interventional do-notation defined in the context of causal graphs.

## 4. Analysis and results

The first stage of the analysis will consist in estimating a causal graph on the dataset to study the interactions between the considered variables. In a second step, given the structure of the obtained graph, a sufficient adjustment set will be selected via the back-door criterion. Lastly, the selected set will be used to implement a full matching procedure and estimate causal effects.

A score-based algorithm called *Tabu Search* (12) has been implemented to learn the causal graph from data. Tabu Search has been selected because it is faster and more accurate than most algorithms for both small and large sample sizes (14). This part of the analysis has been carried out employing the *Bnlearn* package (13) in R Statistical Software (10). Most structural learning algorithms, including Tabu Search, allow the inclusion of prior knowledge in the learning procedure by introducing constraints on the graph's structure. A known causal relationship between variables or the absence of it is encoded in the graph by imposing or forbidding a directed edge between two nodes. Prior knowledge of the subject matter has been synthesized in Table 1. The variables have been divided into four logical groups according to their type. The first group is *Precovid demographics* and contains primary firms' features such as their size, the geographical area where they operate and their business sector. The second group contains additional firms' traits that characterized them before the outbreak. The variation in their revenues and number of employees relative to the last three years, their past strategical activities and exposure to credit rationing are, among others, included in this category. The third group contains the variable which describes the firms' future expectations concerning revenues, measured prior to the pandemic. Post-covid features, such as the number of confirmed covid infections in their province and if they have been targeted by lockdown measures or not, belong to the fourth and last group.

The idea behind this logical categorization is that we assume that variables belonging to a specific group cannot affect the preceding groups described in Table 1. For example, a variable of the last group cannot cause variables of the other three groups, the third group cannot affect the first two and so forth. The four categories are then translated into constraints in the graph structure and used in the structural learning procedure.

In addition to the mentioned constraints, some additional assumptions have been included in the model. Firstly, the treatment and the outcome variable are not allowed to cause any of the pre-treatment

DRAFT

Table 1: Logical variable groups

| Precovid demographics | Other precovid features | Precovid expectations | Postcovid features |
|---|---|---|---|
| Size (n. of employees) | Innovation, R&D | Pre-covid delta expcted revenues | Confirmed covid infections |
| Geographical area | Credit rationing | | Essential business sector |
| Business sector | Export | | |
| Manager education | Delta number of employees | | |
| | Digital litteracy | | |
| | Past delta revenues | | |

variables. This constraint originates from the fact that pre-treatment variable cannot be affected by the treatment or the outcome since they are measured before the treatment is applied. It is also assumed that *geographical area* cannot be affected by the other variables in the first group. Lastly, we assumed that the variable *Essential business sector* could only be affected by the firms' business sector by definition.

The graph learnt by the algorithm is shown in Figure 1. Dashed arrows denote edges that have been forced to be present, based on assumptions regarding the existence of causal relations between variables. This set of assumptions, based on prior knowledge, complements the forbidden arcs assumptions deriving from the logical categories in Table 1.

The causal graph unveils the complex network of causal relations between the variables. The graph is dense, and the emerging structure highlights how all the different considered dimensions contribute to shaping the outcome variable *Post-covid delta expected revenues*. The obtained causal graph is then used to select a sufficient adjustment set of covariates to estimate the effect of implementing home-based working on expected revenues. The R Statistical Software (10) package *Dagitty* (17) has been employed. Given the graph in Figure 1, the minimal set which satisfies the back-door criterion for the effect of $T$ on $Y$ is

$$S_{adj} = \{\textit{Essential business sector}; \textit{Pre-covid delta expected revenues}\}.$$

Full matching is employed for causal effect estimation. The methodology has been implemented in R Statistical Software (10), using the package *MatchIt* (16). Covariate balance improves considerably after matching. Once propensity score is estimated and balance is achieved, the average causal effect of $T$ on $Y$ is calculated by regressing the outcome on the treatment and the adjustment set in a weighted regression model, also referred to as the *outcome model*. The fitted model is then used to predict the distribution of the outcome if all units were controls and if all units were treated. This kind of procedure, also called *g-computation* (15), is required when we include additional covariates in the outcome model and we are interested in estimating a marginal effect. The obtained distribution of the outcome under treatment administration $Y_1$ and control administration $Y_0$ are then averaged and used to compute the causal risk ratio

$$ATE = \frac{E[Y_1]}{E[Y_0]}. \tag{3}$$

Standard errors are computed through block bootstrap (1). ATE point estimates, and bootstrapped 95% confidence intervals (C.I.) are shown in Table 2. The different width of confidence intervals is primarily due to an uneven frequency distribution in the outcome variable levels.

Table 2: ATE estimates of HBW implementation on post-covid $\Delta$ expected revenues

| Post-covid delta expected revenues | Point Estimate | 95% C.I. |
|---|---|---|
| Increase (>+5%) | 1.80 | (0.96, 3.13) |
| Stable (between -5% and +5%) | 2.11 | (1.67, 3.03) |
| Decrease (between -15% and -5%) | 0.97 | (0.61, 1.43) |
| Strong decrease (<-15%) | 0.73 | (0.56, 1.07) |

DRAFT

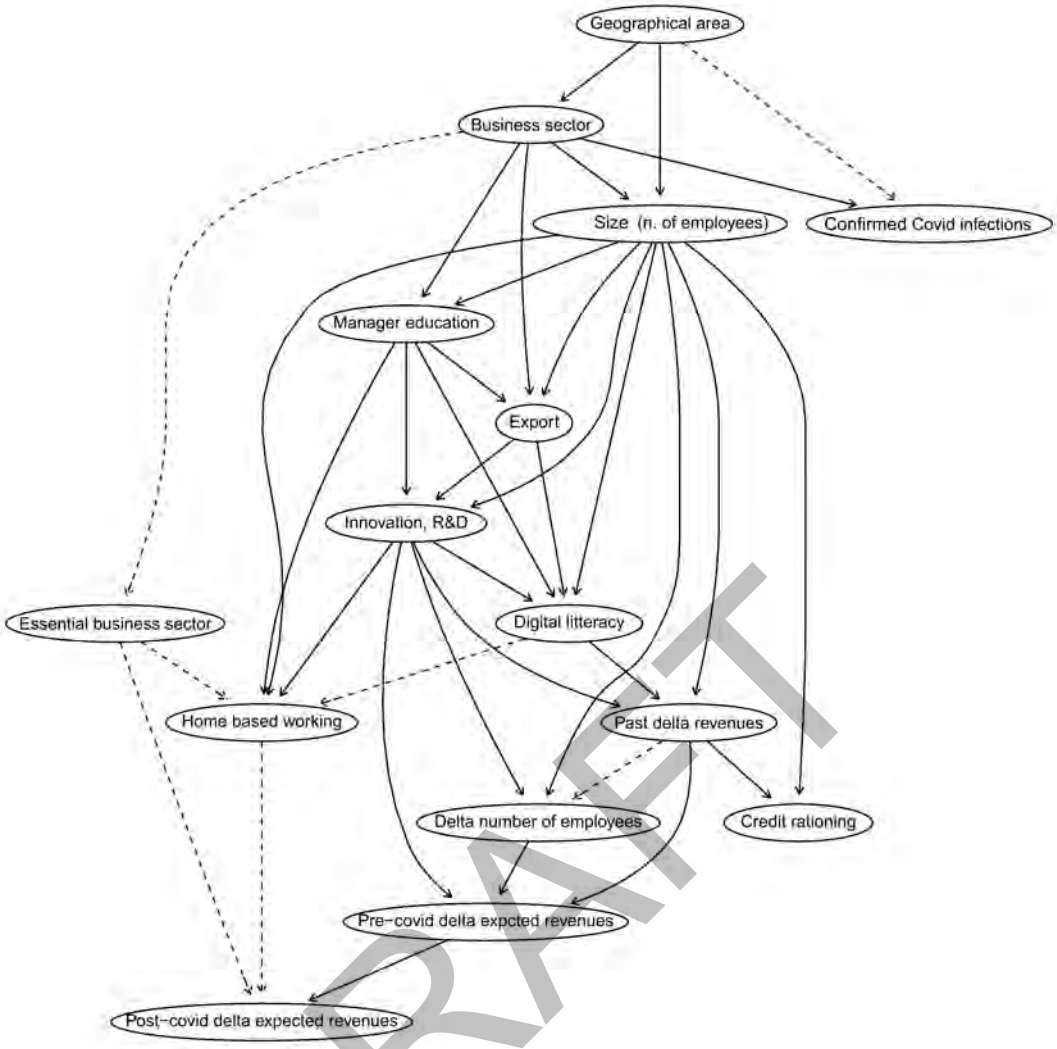



Figure 1: Causal graph learnt with the Tabu Search algorithm

DRAFT

## 5. Discussion

Causal graphs are yet to be the most used approach in economic causal inference literature, but we believe they constitute an irreplaceable resource. Learning a causal graph from data is a process that translates causal information into a more transparent medium. The graph constitutes itself a set of assumptions concerning the causal relationships between variables on which causal estimates are based. Encoding information into a causal diagram thus improves the analysis's clarity and understandability of how results are derived. In addition, the graph learning step allows the introduction of prior knowledge into the model, imposing theory or evidence-based relations between variables. Estimation of ATE using the adjustment set selected from the graph via back-door criterion can be then performed with a method of choice, such as simple regression or matching, depending on the assumptions being made. Regardless of the chosen method, using the graph-selected adjustment set ensures unbiased ATE estimation if the implied assumptions are satisfied. The estimated ATE shows that implementing HBW from the beginning of the pandemic helps mitigate the harmful effects of Covid-19. In particular, treated firms have a higher probability of expecting stable or increasing future revenues and a lower probability of a strong decrease. In other words, the outcome variation generated by the treatment always goes in the opposite direction of the observed change induced by the pandemic outbreak. This finding is coherent with HBW literature and, in particular, with its impact on flexibility and productivity. However, this mitigating effect was yet to be quantified in a comprehensive causal framework.

## References


[1] Abadie, A., Spiess, J.: Robust Post-Matching Inference. Journal of the American Statistical Association **0**(0), 1–13 (2020). DOI 10.1080/01621459.2020.1840383

[2] Balduzzi, P., Brancati, E., Brianti, M., Schiantarelli, F.: The Economic Effects of COVID-19 and Credit Constraints: Evidence from Italian Firms' Expectations and Plans. SSRN Scholarly Paper ID 3682943, Social Science Research Network, Rochester, NY (2020)

[3] Bartik, A.W., Cullen, Z.B., Glaeser, E.L., Luca, M., Stanton, C.T.: What Jobs are Being Done at Home During the Covid-19 Crisis? Evidence from Firm-Level Surveys. Working Paper 27422, National Bureau of Economic Research (2020). DOI 10.3386/w27422

[4] Bloom, N., Liang, J., Roberts, J., Ying, Z.J.: Does Working from Home Work? Evidence from a Chinese Experiment *. The Quarterly Journal of Economics **130**(1), 165–218 (2015). DOI 10.1093/qje/qju032

[5] Felstead, A., Reuschke, D.: Homeworking in the UK: Before and during the 2020 lockdown. https://wiserd.ac.uk/publications/homeworking-uk-and-during-2020-lockdown (2020)

[6] Hansen, B.B.: Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association **99**(467), 609–618 (2004)

[7] Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press (2015)

[8] Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–688 (1995)

[9] Pearl, J.: Causal inference in statistics: An overview. Statistics Surveys **3**(none) (2009). DOI 10.1214/09-SS057

[10] R. Core Team: R: A language and environment for statistical computing (2013)

[11] Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005)

[12] Russell, S., Norvig, P.: Artificial intelligence: A modern approach (2002)

[13] Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software **35**, 1–22 (2010). DOI 10.18637/jss.v035.i03

[14] Scutari, M., Graafland, C.E., Gutiérrez, J.M.: Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms. arXiv:1805.11908 [stat] (2019)

[15] Snowden, J.M., Rose, S., Mortimer, K.M.: Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. American journal of epidemiology **173**(7), 731–738 (2011)

[16] Stuart, E.A., King, G., Imai, K., Ho, D.: MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of statistical software (2011)

[17] Textor, J., van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M., Ellison, G.T.: Robust causal inference using directed acyclic graphs: The R package 'dagitty'. International journal of epidemiology **45**(6), 1887–1894 (2016)


DRAFT

# Patterns of flexible employment careers. Does measurement error matter?

Dimitris Pavlopoulos[a], Mauricio Garnier-Villarreal[a], and Roberta Varriale[b]

[a]Vrije Universiteit Amsterdam; d.pavlopoulos@vu.nl, m.garniervillarreal@vu.nl
[b]Sapienza University of Rome; roberta.varriale@uniroma1.it

**Abstract**

In recent years, the debate on flexible employment has been at the center of political and scientific discussion in Europe. Findings on mobility from flexible to permanent employment can be severely biased due to measurement error, usually present in the data used for analysis. The aim of this paper is to use a mixed hidden Markov model (MHMM) to study the effect of measurement error on the role of flexible employment in the life course. Specifically, we employ a MHMM with two indicators for the employment contract, coming from linked data from the Labour Force Survey and the Employment Register of the Netherlands for the period 2007-2015.

***Keywords:*** mixed hidden Markov model, latent variable model, multisource data, employment careers

DRAFT

## 1. Introduction

In recent years, the debate on flexible employment has been at the center of political and scientific discussion in Europe. In Eurozone countries, in 2021, 11.4% of all employees were employed on a temporary contract (14), while the probability of getting a job on a temporary contract increased by 36 percent between 2013 and 2019 (11). The Netherlands also experienced a sharp increase in the incidence of temporary employment, that increased from 13.2% in 2010 to 20.5% in 2021. The role of temporary contracts can be evaluated from a dynamic or life course perspective: in particular, one of the main questions the research seeks to answer is when temporary work is a stepping stone to permanent employment and when it becomes a trap of precarious jobs (12).

Research has shown that findings on mobility from flexible to permanent employment can be severely biased due to measurement error, usually present in the data used for analysis. In fact, even a small amount of error in the measurement of the type of the employment contract may considerably inflate transitions e.g. from a temporary to a permanent contract, and lead thus to wrongly informed policy decisions (16; 15). In survey data, this measurement error is the result of issues related to cognitive processes, social desirability, design and implementation problems. In register data, measurement error is the result of administrative delays, wrong registration, or erroneous administrative procedures.

A possible approach to deal with measurement errors when multiple data sources are available is based on the use of latent variable models. More specifically, when all the sources contain information

*This paper is part of the project DYNANSE that has received funding from the European Research Council (ERC) under the European Unionâs Horizon 2020 research and innovation programme (grant agreement No 864471)

closely related to the target variable, but none can be assumed to be a corrected measure of the target variable, the latent variable models can be used to predict the true target value given the observed measurements in the data sources (7). In this context, Latent Class Analysis (LCA) is considered as a method to identify a latent categorical construct of interest using observed variables that can be used to evaluate measurement errors (18; 5): hidden Markov Models (HMMs) represent a potential extension when longitudinal data are available. Their particular usefulness is that they do not require the use of an auxiliary dataset that has to be considered as error-free ('gold standard'). These models have been used to correct for measurement error in mobility between employment states (3) and employment contracts (16; 15). Moreover, they have been used to estimate employment status in the Italian employment register (6; 7).

The aim of this paper is to use a mixed hidden Markov model (MHMM) to extend our knowledge on the effect of measurement error on the role of flexible employment in the life course. Specifically, we employ a MHMM with two indicators for the employment contract, coming from linked data from the Labour Force Survey (LFS) and the Employment Register of the Netherlands for the period 2007-2015.

The rest of the paper is organized as follows. In section 2, we present the data. Section 3 and 4 present the MHMM that is used togther with some results. In Section 4 some conclusions are briefly described.

## 2. The data

The data sources providing information on flexible employment are the Labour Force Survey (LFS) administered by Statistics Netherlands and the Employment Register of the Netherlands (ER). LFS represents the main source of information on the labour market for official statistics. It produces information on employment and on the main aggregates of the job offer - profession, sector of economic activity, hours worked, type and duration of contracts, training. LFS is harmonized at the European level as established by the EU Regulation 2019/1700 of the European Parliament and the Council. In the Netherlands, the LFS has a rotating trimonthly scheme and is representative for the Dutch population older than 15 years. The survey was launched in 1987, and its longitudinal component was introduced in 1999. Since 1999, respondents are interviewed at 5 consecutive panel waves. The information that is collected refers to the moment of the interview, and the interviews are carried out during every week of the trimester. As described by (16), measurement errors in LFS may result from different sources. For example, they may depend on misreporting by respondents, mistakes in the recording of responses by interviewers and the use of proxy interviews.

The ER is a register dataset administered by the Institute for Employee Insurance (UWV), containing information on labour market and income for all insured workers in the Netherlands (2). The ER is constructed by collecting and matching information from various sources, i.e. the Tax Office, declarations from temporary work agencies and the Population Register. There is no missing data as the submission of tax-reporting statements is compulsory for employers. However, whereas the dataset contains monthly information, employers typically submit the relevant information only once per year. This may create possible mistakes for the period between two consecutive submissions. As introduced, additional sources of measurement error in ER may result from administrative delays, wrong registration, and erroneous administrative procedures.

For our research, we employ a MHMM with two indicators for the employment contract. These indicators come from linked data from the Labour Force Survey and the Employment Register of the Netherlands for the period 2016-2019. We work with trimester data from this period, for a total of 15 time points. In more detail, we selected all individuals that entered the LFS as from January 2016 until December 2019. From the LFS, we retained information from all the waves for which these individuals participated in the survey. For the same individuals, we also used monthly information from the ER covering all the months from January 2016 until December 2019. Information from the two datasets was linked at the individual level using a pseudonimised version of the Social Security Number (in Dutch: BRP).

DRAFT

## 3. The model and the estimation strategy

As introduced, the aim of this work is to use a mixed hidden Markov model to study the effect of measurement error on the role of flexible employment in the life course.

The true (latent) target variable of the model is the contract type $X_{it}$ at time $t$ for subject $i$, where $t = 0, ..., T$ and $i = 1, ..., N$. We have two measurements for the outcome variable, $C_{it}$ and $E_{it}$, denoting the observed contract type of person $i$ at time point $t$ according to the register and the survey. $C_{it}$, $E_{it}$, and $X_{it}$ can take on four values representing the categories of the contract type (permanent, fixed-term, temporary agency or on call, and other); we refer to a particular category of these variables by $c_t$, $e_t$, and $x_t$, respectively. The latent contract type $X_{it}$ follows a first-order Markov process; that is, the true contract at time point $t$, $X_{it}$, is independent of the contract at time point $t'$, $X_{it'}$, for $t' < t - 1$, conditionally on the state at $t - 1$, $X_{i(t-1)}$.

As indicated in the previous section, we use trimester data from 2007-2015, which means that $t$ runs from 0 to $T = 31$. The probability of following a certain observed path over the $T + 1$ months period can be expressed as follows (HMM):

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i) = \sum_{x_0=1}^{4} \sum_{x_1=1}^{4} ... \sum_{x_T=1}^{4} P(X_{i0} = x_0) \prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}) \prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t) \prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t)^{\delta_{it}} \quad (1)$$

The relevant probabilities are the initial state probabilities $P(X_{i0} = x_0)$, the time-specific transition probabilities $P(X_{it} = x_t | X_{i(t-1)} = x_{t-1})$, the measurement error probabilities for the register $P(C_{it} = c_t | X_{it} = x_t)$, and the measurement error probabilities for the survey $P(E_{it} = e_t | X_{it} = x_t)$. In the model, we assume that the observed states are independent of one another within and between time points, which is referred to as the local independence assumption or the assumption of independent classification errors (ICE). To deal with the fact that $E_{it}$ is observed only every third month for the rotating scheme of LFS, we use the indicator variable $\delta_{it}$ which equals 1 if the survey information is available for the month concerned and 0 otherwise.

The model in equation 2 has been extended to deal with more realistic assumptions in our context. First of all, since the ICE assumption is unrealistic, we used the covariate $\mathbf{V}_{it}$ introducing across-time correlation in the measurement error in the survey data. Furthermore, the model is further expanded with - possibly time-varying - observed variables affecting the initial state and latent transition probabilities. We denote these control variables by $\mathbf{Z}_{it}$. In addition, we used a finite mixture models with $K$ latent classes to account for additional unobserved heterogeneity in the initial latent state and in the latent transition probabilities that is not captured by $\mathbf{Z}_{it}$.

In our mixed hidden Markov model, the joint probability of having a particular observed state path conditionally on covariates $\mathbf{V}$, $\mathbf{Z}$ and $\mathbf{L}$ can be expressed as:

DRAFT

$$
\begin{aligned}
P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E_i} = \mathbf{e}_i | \mathbf{V}_i, \mathbf{Z}_i, \mathbf{L}_i) \quad = \quad & \sum_{k=1}^{K} \sum_{x_0=1}^{4} \sum_{x_1=1}^{4} ... \sum_{x_T=1}^{4} P(w_i = k|\mathbf{L}_i) P(X_{i0} = x_0 | w_i = k, \mathbf{Z}_{i0}) \\
& \prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}, w_i = k, \mathbf{Z}_{it}) \\
& P(C_{i0} = c_0 | X_{i0} = x_0) \\
& \prod_{t=1}^{T} P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1}) \\
& \prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t, \mathbf{V}_{it})^{\delta_{it}} \qquad (2)
\end{aligned}
$$

Equation 2 specifies a finite mixture model with $K$ latent classes to account for unobserved heterogeneity in the initial latent state and in the latent transition probabilities. $P(w_i = k|\mathbf{L}_i)$ is the probability of belonging to the latent class $k$ conditional on the respective covariates $\mathbf{L}_i$, $\mathbf{V}_{it}$ is the vector of covariates affecting the measurement error and $\mathbf{Z}_{it}$ is the vector of the covariates affecting the structural model.

Compared to equation 1, in equation 2, the error probabilities in the survey data are allowed to depend on covariates ($\mathbf{V}_{it}$). Moreover, the error probabilities in the register data are allowed to depend on the lagged observed and lagged true contract type. Note that $X_{i(t-1)}$ and $C_{i(t-1)}$ can take on 4 values, which implies that there are 16 ($4*4$) different sets of error probabilities in the register data, one for each possible combination of lagged observed and latent contract. Because it is not meaningful to estimate all these error probabilities freely, we used a more restricted model. More specifically, we define a constraint logit model when the same error is made between adjacent time points and otherwise being equal to 0. This model expresses that the likelihood of making a specific error depends on whether *the same error* was made at the previous time point. Similar restricted correlated error structures were used by (13) in a latent Markov model for retrospectively collected responses.

Maximum likelihood estimates of the model parameters are obtained using a variant of the Expectation-Maximization (EM) algorithm referred to as the forward-backward or Baum-Welch algorithm (4). We use an extension of this algorithm for MHMM with covariates as described - among others - in (20) and (17). This algorithm is implemented in the program Latent GOLD (19).

The analysis involves several steps and comparisons, requiring the estimation of several computationally-intensive models. In all steps of our strategy, we had to choose between models with different specifications. The choice between these models was done on the basis of model fit measures, namely the BIC, AIC and AIC3 as well as theoretical considerations on the relevance of the mixtures. As the models we estimate become computationally intensive when adding components, we decided to follow the two-step procedure that was suggested by (1).

## 4. Results

The final model was selected in 3 steps: first, we selected the model with the best-fitting specification of measurement error. Then, with this model at hand, we selected the model with the optimal number of trajectories (i.e. mixtures) by following the approach of (1). Finally, we tested whether adding covariates to the model improves model fit.

The final model has an extra error coefficient for the cases where the error made in time point $t-1$ could be repeated in $t$ in the LFS and in the ER. It has 7 Mixtures, and does not use other covariates (other than time) in the equation of latent transitions.

DRAFT

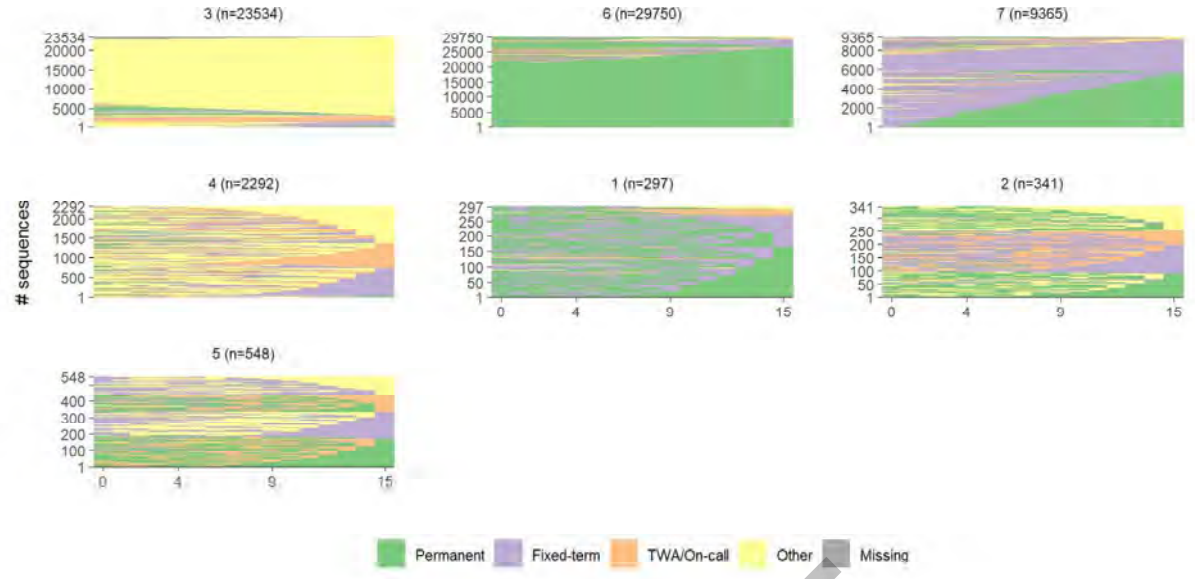



Figure 1: Employment trajectories with measurement-error correction

The results of the final model show that the largest group of individuals (trajectory 6 in Figure 1, 45% of the sample) included individuals that are mostly employed with a permanent contract throughout the observation period. The second largest group (trajectory 3 in Figure 1, 35.6%) belongs to a trajectory that is dominated by non-participation in paid employment. We should mention here that this does not only include unemployed individuals but also people in education and self-employment. The third group (trajectory 7, 14.2%) includes individuals whose employment includes long spells of fixed-term contracts. Some of them manage to get employment with a permanent contract, while others work with a fixed-term contract throughout the period of reference. Another, much smaller group (trajectory 4, 3.5% of the sample) includes careers with a lot of 'churning' between non-(paid) employment, and employment with a fixed-term contract, temporary agency work and on-call work. Clearly this seems to be much more disadvantaged group than other groups with people in employment. The last 3 Mixtures (trajectories 1, 2 and 5) are very small. One could easily classify them as 'residual' trajectories. However, these may also represent true but uncommon careers on the labour market. Mixture 1 (0.4%) includes individuals who switch often between employment with fixed-term and permanent employment. Mixtures 2 (0.5%) and 5 (0.8%) are not well separated as they include individuals that move often across all 4 states.

DRAFT

## 5. Conclusion and future work

In recent years, the debate on flexible employment has been at the center of political and scientific discussion in Europe. In the work, we implemented a MHMM with two indicators for the employment contract to show the effect of measurement error on the role of flexible employment in the life course. We used information coming from linked data from the Labour Force Survey and the Employment Register of the Netherlands for the period 2007-2015. The final model takes into account the longitudinal structure of the data, the measurement error structure of both data sources, and the heterogeneity coming from the population in the flexible employment dynamics.

An important parallel work concerns the comparison of the results obtained through the application of the MHMM to data for the Dutch territory with those for the Italian territory. The aim is to analyze the differences between the most suitable models in the two countries and thus what are the substantial differences in the labour market, particularly concerning the patterns of flexible employment careers.

## References


[1] Bakk, Z., Kuha, J.: Two-Step Estimation of Models Between Latent Classes and External Variables. Psychometrika. **83(4)**, 871–892 (2018)

[2] Bakker, B.F.M., Van Rooijen, J., Van Toor, L.: The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. Stat. J. IAOS. **30(4)**, 411–424 (2014)

[3] Bassi, F., Hagenaars, J.A., Croon, M.A., Vermunt, J.K.: Estimating True Changes when Categorical Panel Data are Affected by Uncorrelated and Correlated Classification Errors: An Application to Unemployment Data. Sociol. Methods Res.. **29(2)**, 230–268 (2000)

[4] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. Ann. Math. Stat.. **41**, 164–171 (1970)

[5] Biemer, P.P.: Latent Class Analysis of Survey Errors. John Wiley & Sons, New Jersey, USA (2011)

[6] Filipponi, D., Guarnera, U., Varriale, R.: Hidden markov models toestimate italian employment status. NTTS 2019 Bruxelles (2019)
https://coms.events/ntts2019/data/x_abstracts/x_abstract_141.pdf. Cited28Feb2023

[7] Filipponi D., Guarnera U., Varriale R.: Latent Mixed Markov Models for the Production of Population Census Data on Employment. In: Perna, C., Salvati, N., Schirippa Spagnolo, F. (eds.) Book of short papers SIS 2021, pp 112-117. Pearson (2021)

[8] Han, Y., Liefbroer, A., Elzinga, C.: Comparing methods of classifying life courses: sequence analysis and latent class analysis. Longitud. Life Course Stud. **8(4)**, 319–341 (2017)

[9] Han, S.Y., Liefbroer, A.C., Elzinga, C.H.: Mechanisms of family formation: an application of Hidden Markov Models to a life course process. Adv. Life Course Res. **43**, 100265 (2020)

[10] Helske, S., Helske, J., Eerola, M.: Combining Sequence Analysis and Hidden Markov Models in the Analysis of Complex Life Sequence Data. In: Ritchard, G., Studer, M. (eds.) Life Course Research and Social Policies, pp 185-200. Springer (2018)

[11] Latner, J.P.: Temporary employment in Europe: stagnating rates and rising risks. European Societies (2022) doi: 10.1080/14616696.2022.2072930
https://doi.org/10.1080/14616696.2022.2072930.Cited31Aug2022

[12] Latner, J.P., Saks, N.: The wage and career consequences of temporary employment in Europe: Analysing the theories and synthesizing the evidence. J. Eur. Soc. Policy. (2022) doi: 10.1177/09589287221106969
https://doi.org/10.1177/09589287221106969.Cited31Aug2022

[13] Manzoni, A., Vermunt, J.K., Luijkx, R., Muffels, R.: Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error. Sociol. Methodol. **40(1)**, 39–73 (2010)

[14] OECD: OECD Statistics (2022)
https://stats.oecd.org/#.Cited31Aug2022

[15] Pankowska P., Bakker B., Oberski D., Pavlopoulos D.: Dependent interviewing: a remedy or a curse for measurement error in surveys?. Surv. Res. Methods. **15(2)**, 135–146 (2021)

[16] Pavlopoulos, D., Vermunt, J.K.: Measuring Temporary Employment. Do Survey or Register Data Tell the Truth?. Surv. Methodol. **41(1)**, 197–214 (2015)

[17] Pavlopoulos, D., Muffels, R., Vermunt, J.K.: How Real is Mobility Between Low Pay, High Pay and Non-employment. J. R. Stat. Soc., A: Statistics in Society. **175(3)**, 749–773 (2012)

[18] Vermunt, J.k.: Longitudinal Research Using Mixture Models. In Montfort, K., Oud, J.H.L., Satorra, A. (eds) Longitudinal Research with Latent Variables, pp. 119-152. Springer, Berlin/Heidelberg (2010)

[19] Vermunt, J. K., Magidson, J.: Technical guide for Latent GOLD 5.1:Basic, advanced, and syntax. Stat. Innov. Inc., Belmont, MA V (2016)

[20] Vermunt, J.K., Tran, B., Magidson, J.: Latent Class Models in Longitudinal Research. In: Menard, S. (ed.) Handbook of Longitudinal Research: Design, Measurement, and Analysis, pp 373-385. Elsevier, Burlington, MA (2008)


DRAFT

# Staying or leaving? A nonlinear framework to explore the role of employee well-being on retention

Kocollari, U.[a], Demaria, F.[b], and Cavicchioli, M.[a]

[a]University of Modena and Reggio Emilia, Department of Economics "Marco Biagi";
ulpiana.kocollari@unimore.it, maddalena.cavicchioli@unimore.it
[b]Marco Biagi Foundation, University of Modena and Reggio Emilia;
fabio.demaria@unimore.it

**Abstract**

Employee well-being has gained the attention of scholars and practitioners over the past two decades. However, despite the increasing number of theoretical works on the topic, empirical studies are still limited. In this study, we aim to define the well-being construct through an exploratory and data-driven approach, and examine its impact on employee retention. We first used a nonlinear dimensionality reduction technique for categorical variables to identify the four main dimensions of work-related well-being. Then, we analyzed these dimensions in a fractional regression framework to predict employee retention. The empirical results suggest that the most significant aspects that discourage employee turnover are related to career growth opportunities, job satisfaction, and interpersonal relationships among coworkers.

***Keywords:*** employee well-being, retention, turnover, nonlinear PCA, fractional logistic regression.

DRAFT

## 1. Introduction

The roots of human well-being can be traced back to Ancient Greece, where philosophers such as Aristippus and Aristotle first began to explore the nature of happiness [19]. Recently, well-being has emerged as a central theme in the organizational context, and it is considered a tool to enhance employee engagement and, ultimately, to improve corporate performance [7]. In addition, over the past two decades, well-being has gained increasing attention on the agenda of public institutions as a key aspect of social sustainability and sustainable development, as reflected in Sustainable Development Goals (SDGs) 3 and 8 [21]. In particular, the latter seeks to promote sustainable economic growth and decent working conditions for all, while SDG 3 aims to improve global health and well-being. The COVID-19 pandemic has threatened progress on both fronts [12], emphasizing the critical importance of coordinated efforts to achieve these goals. In today's globally competitive environment, organizations must prioritize attracting and retaining talented human capital, often at the cost of its well-being. This tendency leads to increased job stress and higher turnover rates, incurring substantial costs for the organization [1; 15; 16]. The objective of this study is to operationally define an individual-level well-being construct using data-driven methods, and to examine its effect on employee retention. Integrating well-being goals into human resource policies may contribute to improving employee retention, optimizing human capital investments, and allowing firms to achieve sustainable development goals, such as SDG 3 and 8. This paper is organized as follows. First, we review the concept of well-being. Second, we present the methodology used to define a new well-being construct. Then, the main results and findings are discussed. Finally, we draw conclusions and suggest managerial implications of the work.

## 2. Literature review

The concept of well-being in social science research has evolved since the 1960s, with the emergence of positive psychology [5]. In general, two distinct yet overlapping perspectives can be distinguished: subjective well-being (SWB) and psychological well-being (PWB). SWB is largely determined by an individual's life satisfaction and evaluations, which are often influenced by objective factors [4]. The multidimensional model of SWB, proposed by Diener (1984) encompasses life satisfaction, positive feelings, and the absence of negative feelings, reflecting the hedonic orientation to happiness [19]. On the other hand, PWB refers to self-realization and has been operationalized as a six-component model, including self-acceptance, positive relations with others, autonomy, environmental mastery, purpose in life, and personal growth [20]. In organizational research, PWB is typically operationalized as job satisfaction or eudaimonic aspects, such as engagement and meaning [9]. From an organizational perspective, well-being has primarily been studied at the individual level, with positive impacts identified for both employees and organizations [6; 25]. The concept of well-being can be defined as "the overall quality of an employee's experience and functioning at work" [22] and is commonly measured through attitudes such as work engagement, job satisfaction, and affective organizational commitment [20; 6]. However, there is a lack of consensus on how to measure subjective work-related well-being due to its multifaceted nature [17].

This work aims to empirically define employee well-being through a data-driven approach and nonlinear methodology. In this regard, we define employee well-being as the outcome of five major areas encompassing both SWB and PWB: economic aspect, interpersonal relationships, personal and professional growth, embeddedness, and innovative behavior. Then, in the second part of our work, we explore the relationship between work-related well-being and retention to guide practitioners in the development of HR retention-oriented strategies.

DRAFT

## 3. Data and methods

The data for this study was collected by the research group through an online survey sent to employees of 12 small and medium enterprises (SMEs) in Northern Italy. The survey consisted of two sections: the first collected demographic information, while the second measured the work-related experiences of the participants through 40 items retrieved from the literature on a 5-point Likert scale [23]. The sample consisted of 196 employees, with roughly equal representation of male and female participants and a majority of respondents between 36 and 60 years old with a university degree $(45\%)$. Approximately $90\%$ of the participants had a full-time contract, and the average organizational tenure was nine years. Finally, we used demographic information to define our proxy variable for employee retention as the ratio between the organizational tenure of each individual, and their total number of working years. The resulting variable is continuous and takes values within the bounded range $[0, 1]$, where 1 identifies employees that spent their whole working career in the same company.

### 3.1 Methodology

To define the well-being construct, we employed Categorical Principal Component Analysis (CATPCA). This methodology was used to reduce the dimensionality of a dataset in an interpretable way while retaining most of its variability [13; 11]. Unlike linear PCA, which is limited to continuous numeric variables, CATPCA allows for the introduction of categorical variables (i.e., nominal and ordinal) through optimal scaling, which is used to quantify category labels [14]. The scaling level was defined as ordinal, based on monotonic transformations that maintain the original categories' order, as the questionnaire items were measured on a 5-point Likert scale. Once the scaling level and the variable weight is defined, the final solution is generated by an iterative algorithm that alternates optimal scaling and dimension reduction, through the minimization of a least squares loss function. For an extensive technical review of CATPCA we refer to Gifi [8]. Then, we used the objects scores extracted from CATPCA as explanatory variables in a fractional logistic regression to predict the employee retention. Fractional logistic regression is a nonlinear regression model designed for fractional response variables, with observations lying inside the unit interval. Given a sample of '$n$' observations, let '$y$' be the dependent variable $(0 \leq y \leq 1)$, and '$x$' the explanatory variable, for each $i^{th}$ observation $(i = 1, \ldots, n)$, we specify the fractional model as follows:

$$E(y_i|x_i) = G(x_i\beta) \tag{1}$$

where $G(.)$ is the logistic function satisfying $0 < G(z) < 1$ for each $z \in \mathbb{R}$. The unknown parameters are estimated using Quasi-Maximum Likelihood Estimation (QMLE), an efficient method under the Generalized Linear Model assumptions [18].

## 4. Results

The first step of our analysis consisted in the extraction of the latent dimensions underlying employee well-being. In order to check for multicollinearity between variables and justify the adoption of Categorical PCA, the association between variables was checked by using Kendall's Tau-b, which was performed for pairs of variables between groups [3]. All the correlations are within the range .204 to .356, suggesting that more than one component is needed to summarize the information in the data. This assumption is also confirmed by the results of Bartlett's test of sphericity, which provided a test statistic of 5101.06, statistically significant at a $1\%$ level

DRAFT

($p < .001$). All the Likert-scale variables were scaled at ordinal level, so that the transformed categories respect the rank order of the original variables. Transformation plots showed monotonic and non-decreasing curves, so the ordinal treatment was appropriate. The whole analysis was run considering a weight $= 1$ for all the variables. To maximize Variance Accounted For (VAF) across principal components while keeping the orthogonal constraint, Varimax rotation was chosen. Following the guidelines of Linting et al. [13], we selected four principal components on the basis of the VAF and interpretability. Besides, the comparison between the most (i.e., numeric) and the least (i.e., nominal) restrictive analysis produced a $2\%$ increase, meaning that accounting for nonlinearity improves the proprortion of variance explained.

**Table 1:** CATPCA model summary

| | **Variance Accounted For** | | |
|---|---|---|---|
| Dimension | Cronbach's Alpha | Total (Eigenvalue) | % of Variance |
| 1 | .955 | 8.445 | 21.653 |
| 2 | .950 | 6.765 | 17.346 |
| 3 | .936 | 4.691 | 12.027 |
| 4 | .842 | 2.736 | 7.014 |
| **Total** | **.981** | **22.636** | **58.040** |

**Note(s)**: Total Cronbach's Alpha is based on the total Eigenvalue.

The model summary indicates that the four components extracted account for the $58.04\%$ of the quantified variables. Cronbach's Alpha scores confirm the good internal consistency between items, with values greater than $.842$. We only retained in the analysis variables with a VAF greater than $.25$, that is, at least $25\%$ of the variance in a quantified variable is explained across the principal components. Based on the common traits of the variables selected for each dimension, we labelled them as follows. Dimension $1$, which explains $21.65\%$ of variance, is related to self-realization through work and skills and is labelled 'Realization'. Dimension $2$ captures $17.35\%$ of variance and was labelled 'Equity' as it refers to financial reward and career advancement, traditional HR management, and work-life balance. The third dimension, 'Responsibility', covers $12\%$ of variance and includes non-ordinary matters management, critical issues, and teamwork aspects. Finally, the fourth dimension captures $7\%$ of observed variance and represents interpersonal relationships aspects ('Connections').

DRAFT

## 4.1 Well-being and retention

After the dimensionality reduction performed by CATPCA, we fitted a predictive model to investigate the impact of the four principal components on employee retention. Since the outcome variable is measured as a fraction taking values within the interval $[0, 1]$, we employed a fractional logistic regression. Prior to fitting the model, the Ramsey's regression specification error test (RESET) was performed to identify any general functional form misspecification [24]. The RESET statistic yielded a value of $.40$ with an associated *p-value* of $.672$, indicating that the fractional model is correctly specified. The fractional model equation can be expressed as:

$$E(retention|x_k) = G(\beta_0 + \beta_1 Realization + \beta_2 Equity + \beta_3 Responsibility + \beta_4 Connections) \quad (2)$$

where $x_k$ is the $k^{th}$ predictor $(k = 1, \ldots, p)$ and $G(.)$ is the logistic function ensuring predictions to be within the unit interval. We estimated the nonlinear model in equation (1) using the logit link function and QMLE. In the following table, we report the baseline model estimates

and their relative average marginal effects (AME).

**Table 2:** Fractional logistic regression

| | **Model estimates** | | **AME** | |
|---|---|---|---|---|
| | Coefficient | Std.Error[a] | Coefficient | Std.Error[a] |
| *constant* | .780 | .095 | | |
| Realization | .304*** | .097 | .065*** | .021 |
| Equity | .144 | .108 | .031 | .023 |
| Responsibility | .062 | .089 | .013 | .019 |
| Connections | .162** | .083 | .034** | .018 |
| $R^2 = .052$ | | | | |

**Note(s)**: [a]Robust standard errors. * $p < .1$; ** $p < .05$; *** $p < .01$.

As can be seen from Table 2, all the dimensions are positively related to employee retention. In particular, 'Realization' and 'Connections' are statistically significant at .01 and .05 significant level, respectively. Thus, model estimates suggest that the most relevant elements encouraging employee retention relate to personal and professional growth opportunities, perceived job satisfaction, social capital, and relational aspects. In the second column, AME, or partial effects, indicate the impact of the four PCs on the response variable for every unit increase. In fact, because of its nonlinearity, partial effects should be preferred in the interpretation of the model [10]. For example, employee retention is expected to decrease, on average, by .65 with a marginal change in the 'Realization' dimension, holding the other predictors constant. Finally, since the Breusch-Pagan test underlined heteroscedastic disturbances in the model ($BP = 9.612$, $p - value = .047$), robust (i.e., White) standard errors were computed based on a heteroskedasticity-consistent covariance matrix.

DRAFT

## 5. Conclusions

In recent years, work-related well-being has attracted the attention of scholars and practitioners, becoming a primary focus with the onset of the Covid-19 pandemic. In our paper, we adopt an exploratory and data-driven approach to investigate the main components of employee well-being and their impact on employee retention. Retaining talented workers is a strategic element in maximizing the organization's investment in human capital and generating a competitive edge. Data collection for this study was conducted through an online survey, which consisted of 40 variables designed to measure various dimensions of work-related well-being. The sample consisted of 196 employees from 12 SMEs. The results of the nonlinear PCA revealed four latent dimensions of well-being, which refer to personal growth ('Realization'), monetary and non-monetary rewards ('Equity'), employee autonomy and critical issues management ('Responsibility'), and relationships among coworkers ('Connections'). The fractional logistic regression employed to predict employee retention revealed that all four dimensions had a positive impact on the response variable, with 'Realization' and 'Connections' being the most significant aspects in reducing employee turnover. These findings add to the existing literature on employee well-being by proposing a new data-driven construct. Results suggest that HR professionals should focus on providing professional growth opportunities and creating a positive and supportive work environment that fosters social connections to retain employees. Besides, embedding well-being objectives in HR strategies may contribute to meeting sustainable development targets, as reported in SDG 3 and 8.

## References


[1] Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining Talent: Replacing Misconceptions With Evidence-Based Strategies. *Academy of Management Perspectives*, 24(2), 48-64.

[2] Bartlett, M. S. (1951). The effect of standardization on a $\chi^2$ approximation in factor analysis. Biometrika, 38(3-4), 337-344.

[3] Chen, P., & Popovich, P. (2002). *Correlation*. SAGE Publications, Inc.

[4] Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95(3), 542-575.

[5] Diener, E. (2009). *The Science of Well-Being* (E. Diener, Ed.; Vol. 37). Springer Netherlands.

[6] Fisher, C. D. (2010). Happiness at Work. *International Journal of Management Reviews*, 12(4), 384-412.

[7] Gaston-Breton, C., Lemoine, J. E., Voyer, B. G., & Kastanakis, M. N. (2021). Pleasure, meaning or spirituality: Cross-cultural differences in orientations to happiness across 12 countries. *Journal of Business Research*, 134, 1-12.

[8] Gifi, A. (1990). Nonlinear multivariate analysis. Wiley.

[9] Grant, A. M., Christianson, M. K., & Price, R. H. (2007). Happiness, Health, or Relationships? Managerial Practices and Employee Well-Being Tradeoffs. *Academy of Management Perspectives*, 21(3), 51-63.

[10] Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, 28(4), 331-343.

[11] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

[12] Leal Filho, W., Brandli, L. L., Lange Salvia, A., Rayman-Bacchus, L., & Platje, J. (2020). COVID-19 and the UN Sustainable Development Goals: Threat to Solidarity or an Opportunity? *Sustainability*, 12(13), 5343.

[13] Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Koojj, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12(3), 336-358.

[14] Meulman, J., van der Kooij, A., & Heiser, W. (2004). Principal Components Analysis With Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 50-71). SAGE Publications, Inc.

[15] Michele Kacmar, K., Andrews, M. C., Van Rooy, D. L., Chris Steilberg, R., & Cerrone, S. (2006). Sure Everyone Can Be Replaced...But At What Cost? Turnover As A Predictor Of Unit-Level Performance. *Academy of Management Journal*, 49(1), 133-144.

[16] Mosadeghrad, A. M. (2013). Occupational Stress and Turnover Intention: Implications for Nursing Management. *International Journal of Health Policy and Management*, 1(2), 169-176.

[17] Page, K. M., & Vella-Brodrick, D. A. (2009). The 'What', 'Why' and 'How' of Employee Well-Being: A New Model. *Social Indicators Research*, 90(3), 441-458.

[18] Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6), 619-632.

[19] Ryan, R. M., & Deci, E. L. (2001). On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology*, 52(1), 141-166.

[20] Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57(6), 1069-1081.

[21] United Nations General Assembly. (2015). *Transforming our world: The 2030. Agenda for sustainable development*.

[22] Warr, P. (1987). *Work, unemployment, and mental health*. Oxford University Press.

[23] Warr, P. (2007). *Work, Happiness, and Unhappiness*. Psychology Press.

[24] Wooldridge, J. (2012). *Introductory econometrics: A modern approach* (5th ed.). Cengage Learning.

[25] Wright, T. A., & Cropanzano, R. (2004). The Role of Psychological Well-Being in Job Performance: A Fresh Look at an Age-Old Quest. *Organizational Dynamics*, 33(4), 338-351.


DRAFT

# The CAP instruments impact on GVA and employment: a multivalued treatment approach

Montezuma Dumangane[a] and Marzia Freo[a]

[a]European Commission, Joint Research Centre (JRC), Ispra;
Montezuma.DUMANGANE@ec.europa.eu, Marzia.FREO@ec.europa.eu

**Abstract**

This study provides an EU NUTS3 level evaluation of the CAP impact on employment and Gross value added by assessing the relative merits of its instruments. The policy implementation is interpreted as a discrete set of policy mixes, and modelled as a multivalued discrete treatment. The treatment levels represent the prevalence of the three main types of CAP funds -Market Measures, Direct Payments, and Rural Development- and a reference Low CAP level. The Treatment Effects estimates confirm the role of all CAP instruments, particularly Market Measures, in preserving jobs in the agricultural sector and the effectiveness of decoupling Direct Payments.

***Keywords:*** Generalised Propensity Score, CAP, Counterfactual Impact Evaluation

DRAFT

## 1. Introduction

Evaluating the CAP instruments has been the object of empirical studies using a diverse range of methodological approaches (see (9) for a comprehensive review of the literature on the CAP impact). This literature has become increasingly rich in analysing the effect of Pillar 1 instruments and Pillar 2 vast array of measures under Rural Development funding. However, the individual evaluation of its instruments does not address the relevant issue of assessing their relative effectiveness nor provides guidance on the efficient allocation of resources. Treating the CAP as a multidimensional policy and acknowledging the quantitative and qualitative differences in its implementation instead provides such insight to the policymaker. Furthermore, few studies address the challenge of robustly identifying the causal effect of the policy using the methods from the Treatment-Effects literature (8), while the region's economic and agricultural profile are potential determinants of the funds' allocation and outcomes and the evaluation of CAP causal effect need a Counterfactual Impact Evaluation (CIE) method. The existing empirical literature on the CAP impact with an EU regional dimension provides evidence of the policy performance that, however, is not derived from the toolkit of CIE methods. On the other side, most counterfactual studies are focused on a single country and the limited geographical coverage makes the results scarcely generalisable. Thus, the literature on the CAP evaluation lacks a counterfactual impact evaluation analysis that: provides an EU-level performance of the CAP, and inspects the effectiveness of the multiple CAP instruments. The present study attempts to address these issues by focusing on regional output and employment growth.

## 2. The CAP as a multivalued treatment

This study adopts an identification strategy to evaluate the CAP with causal methods while preserving its multidimensionality. The main feature of the identification strategy is that each region is treated

by a policy mix composed of three types of subsidies: Market Measures, Direct Payments, and Rural Development expenditure. As the policy mixes are as many as the regions, they are clustered into a discrete number of mutually exclusive groups. These are chosen to preserve their similarities within the groups as much as possible. The groups of policy mixes become distinctive of the treatment levels of a multivalued treatment design. This procedure reduces the dimensionality of a continuous multivariate treatment while preserving the qualitative aspects associated with the different policy mixes.

Consequently, regions within the same treatment level do not display perfectly homogeneous policy mixes. In this regard, the empirical literature on impact evaluation often neglects the inner composition of the treatment levels. While it is common practice to discretise continuous treatments or aggregate different policy instruments, to the best of our knowledge, very few analyses explicitly state that treatment levels should often be heterogeneous combinations of essential components with different intensities: (4) highlights that "the effective intervention plan usually contains multiple treatment components," and (10) acknowledges that "many applications involve simultaneous intervention on multiple variables". Of course, this only happens for practical reasons. Indeed, under continuous multivariate treatment, it is unfeasible to identify groups with units treated with the same level of each intervention. Allowing for heterogeneity within each treatment level makes the impact analysis feasible, as different units under the same (heterogeneous) treatment can be grouped.

The Generalised Propensity Score (GPS) method of (7) is applied to the discrete multivalued representation of the CAP to eliminate the selection bias. The GPS is estimated using the Covariate Balancing Propensity Score method of (6) (CBPS), which models the treatment assignment while optimising the covariate balance, and mitigates the consequences of the possible misspecification of propensity score models. Given an estimate of the propensity score, the Weighted Least Squares regression estimator (11), as a variant of the Horvitz and Thompson (5), is implemented.

DRAFT

## 3. Empirical analysis

The analysed period, 2011-2015, corresponds to the implementation of the 2009 Health check reform, which reinforced the funding of Rural Development programmes and further promoted the decoupling of Direct Payments. The analysis is based on a sample that aggregates some of the original NUTS3 units to produce a more homogenous territorial representation of the EU28. The combined effect of several types of CAP funds is considered: the Pillar 1 subsidies are grouped into three categories: Market Measures, granted to farmers to stabilise agricultural markets, prevent market crises from escalating, boost demand and help EU agricultural sectors to better adapt to market changes; Direct Payments, which include all forms of payments linked to the production of specific products or granted to farmers in advance and detached from the production level; and Rural Development Programmes, which payments do not target specifically the agricultural sector but aim to improve the quality of life in rural areas and achieve a balanced territorial development of rural areas and communities. Finally, the CAP funds' are measured as intensities (1; 2), the Pillar 1 funds as the ratio to the average GVA of agriculture and the Rural Development funds as the ratio to the average total GVA, computed from the two years previous to the beginning of the CAP implementation.

The analysis starts by identifying the treatment design by grouping the regions' mixes of Direct Payments, Market Measures and Rural Development Programmes intensities according to a two-step procedure. First, since all regions are CAP recipients, it is convenient to define a reference treatment level by identifying those with low intensity in all CAP funds. The resulting Low CAP group selected 98 regions with simultaneously all treatment level intensities of Pillar 1 and Pillar 2 below respectively the fifth and sixth decile. This treatment level is the natural counterfactual against which the impact of any other treatment level is to be evaluated. Second, the remaining regions are grouped using a hierarchical cluster algorithm on the three funds' intensities. The algorithm identified three groups, which, together with the Low CAP, form the treatment. Table 1 shows an index of the CAP funds' intensity in the treatment. The index measures the treatment level average intensity to the overall sample average intensity. The figures confirm Low CAP group of regions limited funding. These receive Market Measures, Direct Payments and Rural Development funds' intensities, respectively, equal to 15%, 43%,

Table 1: Treatment levels intensity index[a] across treatment designs

| | No. of Regions | Market Measures | Direct Payments | Rural Development |
|---|---|---|---|---|
| Low CAP | 98 | 15 | 43 | 28 |
| PRD | 224 | 72 | 71 | 211 |
| PDP | 342 | 65 | 150 | 71 |
| PMM | 132 | 301 | 62 | 40 |

[a] *Note*: Table entries show ratio of average treatment level intensity in treatment group to overall average treatment level intensity.

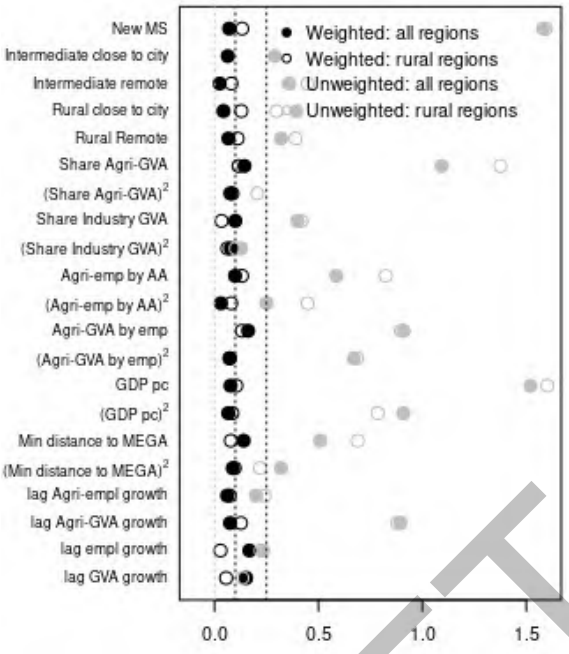
and 28% of the overall average. The clustering exercise produced three treatment levels. Because they are relatively more intense in each one of the CAP funds, the baseline treatment levels will be labelled with the prefix "Predominantly" to underline the prevalence rather than its homogeneity within the category. The Predominantly Rural Development (PRD) group has, on average, around twice the overall average intensity of these funds, the Predominantly Direct Payments (PDP) group, around 1.5 times, and the Predominantly Market Measures (PMM), exactly three times the overall average intensity. The treatment levels also exhibit a heterogeneous overall CAP intensity, computed as a weighted average of funds' intensities with weights equal to their share of the total CAP. On average, regions in the Low CAP group received 37% of an average EU region. In contrast, regions on the PMM and PRD receive, on average, respectively 64% and 96%, while the PDP receive 135%, reflecting the highest share of these funds in the total CAP.

Interest lies in measuring the CAP contribution to the EU regions' agricultural economy and their spillovers to the regional economy. Two economic indicators are considered: GVA at constant prices and Employment, in the agricultural sector and in the regional economy. Both are measured as growth rates, between one year before the beginning of the CAP period and one, two and three years after its end. This allows us to infer the impacts' time path while simultaneously providing a robustness check. To investigate it, particularly the CAP's effectiveness in promoting balanced territorial growth, a heterogeneity analysis is carried out in the sub-sample of rural regions here defined as the non-predominantly urban areas from the Eurostat classification (3). The empirical strategy relies on the ability to identify a set of pre-treatment variables (see Figure 1) that satisfy the (weak) unconfoundedness assumption (7). The CBPS has been estimated over a set of control variables, describing the regions in several dimensions that influence both the potential outcomes and the treatment levels' assignment, one year before the policy period, i.e., in 2010. Except for the dummy indicators and the lagged outcomes, squares of the pre-determined variables were added, thus enriching the balancing property by extending it to the second moments. Figure 1 shows the absolute standardised maximum mean differences for assessing covariate balancing in the treatment. The figure illustrates how the procedure successfully balanced the means and variances of the covariates across the treatment levels.

Figure 2 shows the ATEs estimates for the entire sample and the sub-sample of rural regions. In both representations, $t+1$, $t+2$ and $t+3$ correspond to growth rates between 2011 and respectively 2016, 2017, and 2018. The estimated Low CAP Expected Potential Outcomes (EPO) growth at $t+1$ and $t+3$ (here not presented) is respectively, -14% and -17%. They reflect the negative trend of employment in the sector between 2011 and 2016 (2018). The same is observed in GVA in agriculture growth, where the EPOs at $t+1$ and $t+3$ are respectively, -5.9% and -6.3%. The ATEs show how all forms of CAP support contributed to attenuating the job losses and GVA decline in the agri-sector. The ATEs of regions treated with a high intensity of Market Measures reveal that this type of support contributed to preserving most agricultural jobs. The remaining CAP treatments only attenuate the negative trend in agri-employment, with the PDP group exhibiting a higher impact than the PRD. Conversely, while all CAP support positively impacts GVA in agriculture growth above the Low CAP EPO, PRD mixes induce the highest medium term growth. In fact, the policy mix predominantly funded by Rural Development measures is the only one to display an ATE that induces a non-negative expected growth GVA in agriculture. These results reveal that the CAP stimulates the agri-economy by safeguarding jobs through interventions that support the farmers' activity (PMM) and direct income support while Rural

DRAFT

Figure 1: Maximum absolute standardised mean differences before and after weighting: all and rural regions.



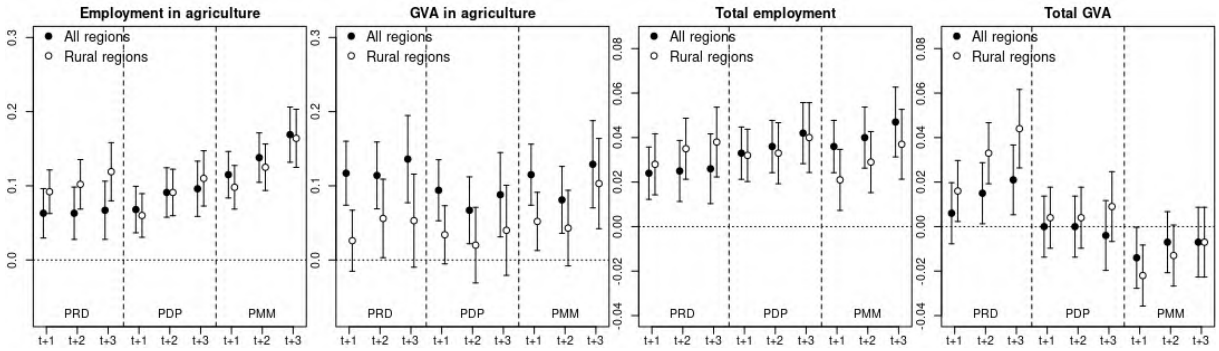

DRAFT

Development expenditure promotes consistent growth in the agricultural sector output. The results in the sub-sample of rural regions show that, there are no substantial differences in the negative trend of the agri-employment nor on how the Pillar 1 funds sustain its growth. Instead, the PRD interventions are more effective in safeguarding jobs in rural regions. The Low Ates estimates on GVA in agriculture are associated to a non significant impact of the PRD and PDP mixes and to smaller ATEs of PMM. Thus these results suggest that the structural changes induced under the PRD interventions do not close the GVA growth gap in the rural areas' agri-sector.

The last two graphs of Figure 2 show the CAP impact on the local economy outcomes. These (spillover) effects are, as expected, significantly smaller and are, under Low CAP support, associated with less negative trends in employment and a positive trend in GVA growth. However, while all CAP treatments induced growth in employment of around 3 percentage points above the Low CAP, only the PRD group exhibited a positive but negligible impact on total GVA growth. This treatment effect is significantly stronger in the sub-sample of rural regions. The longer term $t+3$ ATEs imply a yearly total GVA growth above the Low CAP of 0,7 percentage points, showing that although PRD support for the agricultural sector was relatively less effective in these regions, it contributes to the convergence of the rural regions' economy.

The paper's main findings shed light on how the different CAP instruments promote regional growth and safeguard jobs in the EU regions. All three forms of CAP support contribute to attenuating the job losses in the regional agricultural with limited spillovers to the local economy. However, although their positive contribution to GVA in agriculture growth does not unequivocally extend to the rural regions, Rural Development expenditure promotes growth and convergence of the rural economies. In fact, the policy mix predominantly based on Rural Development measures positively and significantly impacts all other outcomes. Despite being the least funded, the CAP mix intensive on Market Measures is found to be the most effective in promoting agricultural employment, with significant spillovers on regional jobs. Conversely, policy mixes based on Direct Payments, despite being more intensively financed, are found not to provide the highest impacts in any outcome.

Figure 2: Average treatment effects estimates: Full sample and rural regions




## 4. Conclusions

A recent review of the empirical literature on the socioeconomic impacts of the CAP causal impact found around 60 contributions. These analyses range from considering the whole CAP to analysing specific policy instruments. Most studies focused on a single country, or a part of it, providing limited geographical coverage, which hinders external validity given the regional differences across the EU regions. Furthermore, most existing evaluations of the CAP use methods that do not allow inferring the causality link between the policy and the outcomes. The present paper evaluates the CAP impact on regional employment and GVA growth in the agricultural sector and total local economy. The analysis provides two important contributions. First, the evaluation covers the EU28 regions and brings insight into the nature of the impacts in the rural areas. This is made possible by using a NUTS3 level dataset with disaggregated data on the different CAP funds. Secondly, the study proposes a procedure to investigate policies characterised by multiple joint interventions. Indeed the detailed info on the CAP is exploited to study the combined impact of different funds' categories using the Generalised Propensity Score framework of Imbens (7).

DRAFT

## References

[1] Camaioni, B., Esposti, R., Lobianco, A., Pagliacci, F., Sotte, F.: How rural is the eu rdp? an analysis through spatial fund allocation. Bio-based and Appl. Econ. (2013) doi: 10.13128/BAE-13092

[2] Crescenzi, R., Giua, M. The EU cohesion policy in context: Does a bottom-up approach work in all regions? Environ. and Plan. A: Econ. and Space (2016) doi: 10.1177/0308518x16658291

[3] European Commission Methodological manual on territorial typologies: 2018 edition. Publications Office of the European Union, Luxembourg (2019) doi: 10.2785/228845

[4] Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., Li, X.-S. Generalized propensity score for estimating the average treatment effect of multiple treatments. Stat. in Med. **31**, 681-697 (2011)

[5] Horvitz, D., Thompson, D. A generalization of sampling without replacement from a finite universe. J. of the Am. Stat. Assoc. **47**, 663-685 (1952).

[6] Imai, K., Ratkovic, M. Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76**, 243-263 (2014)

[7] Imbens, G. W. The role of the propensity score in estimating dose-response functions. Biometrika **87**, 706-710 (2000)

[8] Imbens, G. W., Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. J. of Econ. Lit. **47**, 5-86 (2009)

[9] Lillemets, J., Fert?o, I., Viira, A.-H. The socioeconomic impacts of the CAP: Systematic literature review. Land Use Policy (2022) doi: 10.1016/j.landusepol.2021.105968

[10] Qian, Z., Curth, A., Schaar, M. v. d. Estimating multi-cause treatment effects via single-cause

perturbation. Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021) (2021)
[11] Robins, J., Hernan, M., Brumback, B. Marginal structural models and causal inference in epidemiology. Epidemiol., 550-560 (2000)

DRAFT

# The determinants of leaving the parental home in Italy: 2012-18

Ilaria Rocco[a] and Gianpiero Dalla Zuanna[b]

[a] Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy; Veneto Lavoro, via Ca' Marcello 67/b, 30173 Mestre-Venezia, Italy ilaria.rocco@uniroma1.it

[b] University of Padova, via Battisti 241, Padova 35121, Italy; gpdz@stat.unipd.it

**Abstract**

Using the longitudinal structure of the Istat Italian Labor Force Survey, we follow for 15 months the 77.700 young people aged 19-34 who, at the first interview during 2012-18, lived in their parental home, dividing them into four groups (Italian men, Italian women, foreign men, foreign women). In the fourth interview, 15 months after the first one, 16% of them are no longer resident with their parents. The main objective of this work is to model the probabilities of leaving the parental home with the characteristics of young people at the first interview, using logistic regression. The young Italians most likely to leave their parental home are those employed full-time, regardless of gender and type of contract (fixed-term or open-ended). Students remain more frequently with parents irrespective of gender and citizenship. Only for Italians, leaving home is faster if the family of origin is large and the parents are more educated.

**Keywords:** leaving home, young people, employment, gender, citizenship

DRAFT

## 1. Introduction

Leaving parental home is a key marker in the transition to adulthood [12, 29] and can have important consequences for young adults' life course outcomes. If on one hand a late exit from parental home may prolong the transition to adult roles [15, 17, 19], on the other one an early exit may instead be associated with earlier experiences of other events, such as exit from education, entry into work or forming a family, that could prevent young people from acquiring adequate human capital for later in life [24, 26].

A large demographic research literature has therefore addressed a broad array of individual, parental, and contextual characteristics shaping the decision of young adults to leave parental home in Western societies [1, 2, 9, 11, 12, 16, 20, 21, 22, 27, 30].

Among the numerous factors influencing this decision, difficulties in labour market integration and unstable working conditions may play an important role. Several studies showed the negative effects of job precariousness on the propensity of youth to leave the parental household [6, 16, 22, 32], that can vary depending on the institutional and cultural context, as well as according to gender and reason for leaving home [2, 16, 27]. Not only unemployment or inactivity may limit young people's opportunities to leave their parental home [2], but also the instability of employment [4, 5, 14], irrespective of the level of income received by young adults. An Italian study [3], confirmed by a comparative analysis between Italy and Spain [8], observed that being employed is a prerequisite to exit from the family household more important than income itself. However this results true only for men: being employed, unemployed or inactive does not influence women's propensity of leaving their parents.

From another body of literature it emerges that also the family background, that is linked to young adults' opportunity and need structure, may influence the transition out of the parental home. In particular

the parental education, that is a predictor of young adults' occupational achievement stronger than parents' occupational status and income [13], seems affect the probability of exiting from family home [10, 18, 23, 28, 31] depending on the modality of leaving home (i.e. with or without a partner) and young adults' gender [7].

This study focuses on Italy where the youth unemployment rate is one of the highest among the EU members and leaving home traditionally occurs at later ages than in Northern or Western Europe. In Southern-European countries, characterised by a familist welfare model, national policies to support individuals from job loss or in housing costs are less generous than those of Scandinavian countries and therefore the relationship between the labour market condition and the transition to residential autonomy tends to be stronger [25].

Our aim is to investigate the factors associated to the exit from the family household in recent years, considering both young adults' characteristics, their working condition and their familiar background. We intend to pay special attention to the gender dimension, and also introduce a citizenship perspective, that is still little explored.

## 2. Data and methods

The analysis is based on microdata from the Italian Labour Force Survey (ILFS) carried out by Istat. The ILFS is the largest survey conducted in Italy to monitor the quarterly dynamics of the labour market and it provides data on employment, wages, and workforce. This survey follows a rotating sample design, where households participate for two consecutive quarters; they exit for the following two quarters, and come back into the sample for a further two consecutive quarters. Therefore, 50% of the households interviewed in the course of a quarter are re-interviewed after 3 months, 50% after 12 months, 25% after 9 and 15 months, respectively. If an individual leaves the sampled family, he/she is not re-interviewed, but it is possible to know – in fact – that he/she no longer belongs to that family unit.

The access to this dataset with wider information than the free version was possible thanks to a data sharing agreement between Istat and the University of Padua.

The present work uses data for the period 2012-18 (first interview) and 2013-19 (fourth interview), and it focuses on the around 77 thousand individuals aged 19-34 living with their parents when the household was enrolled in the survey. We were interested to observe which subjects left the parental home during the 15-month follow-up in order to identify their main characteristics. We divided young respondents in four groups according to their gender and citizenship, and then we ran four separate logistic regressions to highlight the different factors associated to the propensity to leave the parental home in each group.

We estimated three models for each group by including explanatory variables according to three thematic blocks: socio-demographic characteristics (i.e. age, geographic repartition and year of enrolment); employment status (i.e. temporary/permanently fully/partially employed, students, jobseeker, inactive); household characteristics (i.e. educational level of parents and number of household members). Therefore, the first model (that will be called *m1* in Table 2 of the results section) includes only the socio-demographic characteristics of the individual; in the second one (*m2*) we added the variable on the employment status; finally in the last model (*m3*) we included all the three blocks of variables.

## 3. Results and preliminary conclusion

Among the 77.711 individuals aged 19-34 living with parents when their household was enrolled in the survey 6% have foreign citizenship (Table 1). The percentage of men is greater than that of women (55% among Italians and 58% among foreigners). Although the foreign subjects are younger than Italians, they are less involved in the educational process and more employed (also with open-ended contracts).

The foreign presence is concentrated in central-northern regions. Focusing on family characteristics, about one third of foreigners live with at least 4 people and nearly half of them have poorly educated parents.

During 15-month follow-up 16% of the total sample left the parental home: this percentage is higher among Italians and among women of both citizenships.

The logistic regression shows that age has a positive and significant effect on leaving the parental home among Italian males and females, but not for foreigners. Moreover, in comparison with Italians living in the Centre, those living in the South and the Islands are more likely to exit from the household during the

DRAFT

follow-up; among Italian females also living in the North-east has a positive effect on the propensity to leave the parental home. For the foreigners, there is no significant effect of the geographic repartition of residence.

Table 1: Sample description (weighted frequencies)

| | Italians | | | Foreigners | | |
|---|---|---|---|---|---|---|
| | Males | Females | Total | Males | Females | Total |
| Sample size | 39.536 | 33.509 | 73.045 | 2.684 | 1.982 | 4.666 |
| Age, mean (std. dev.) | 25,3 (4,3) | 24,5 (4,1) | 24,9 (4,2) | 24,1 (4,1) | 23,7 (3,8) | 23,9 (4,0) |
| Job condition | | | | | | |
| Open-ended, FT[a] | 20% | 10% | 15% | 24% | 11% | 18% |
| Open-ended, PT[b] | 3% | 5% | 4% | 5% | 11% | 7% |
| Fixed-term, FT | 11% | 8% | 9% | 13% | 7% | 10% |
| Fixed-term, PT | 3% | 5% | 4% | 3% | 6% | 4% |
| Self-employed | 9% | 6% | 8% | 6% | 3% | 5% |
| Student | 26% | 37% | 31% | 18% | 30% | 23% |
| Jobseeker | 16% | 15% | 15% | 19% | 17% | 18% |
| Inactive | 13% | 14% | 14% | 12% | 16% | 14% |
| Geographic repartition | | | | | | |
| Centre | 19% | 19% | 19% | 30% | 27% | 28% |
| Islands | 13% | 13% | 13% | 4% | 3% | 4% |
| North-east | 17% | 17% | 17% | 26% | 26% | 26% |
| North-west | 23% | 22% | 23% | 32% | 35% | 33% |
| South | 29% | 29% | 29% | 9% | 10% | 9% |
| Household members | | | | | | |
| 2 | 9% | 7% | 8% | 12% | 12% | 12% |
| 3 | 34% | 32% | 33% | 24% | 25% | 24% |
| 4 | 42% | 44% | 43% | 32% | 31% | 31% |
| 5+ | 16% | 17% | 16% | 33% | 33% | 33% |
| Parental education | | | | | | |
| Up to middle school | 42% | 40% | 41% | 48% | 42% | 45% |
| Secondary school | 43% | 44% | 43% | 43% | 47% | 44% |
| University | 15% | 16% | 16% | 9% | 12% | 10% |
| Left during the follow-up | | | | | | |
| No | 84% | 82% | 83% | 85% | 84% | 84% |
| Yes | 16% | 18% | 17% | 15% | 16% | 16% |

[a] Full time; [b] Part time

Focusing on the job condition, it emerges that students remain in the parental home more frequently than other occupational groups, regardless of their gender and citizenship. The same results are observed also for self-employed people, but only among Italians. Moreover, among Italian men and women what matters is not so much the duration of the contract but the working time: people with a part-time job have indeed a lower probability to leave than full-time workers while the duration of the contract has a not significant effect. The duration of the contract and the working time seem not to affect the propensity of foreign males of leaving their family; foreign females that work part-time with an open-ended contract are less likely to exit from the parental home than those with the same contract but a full-time position.

DRAFT

Table 2: Propensity to leave the parental household during 15-month follow-up. Estimated odds ratios from logit models

| | Italian Females | | | Italian Males | | | Foreign Females | | | Foreign Males | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | m1 | m2 | m3 | m1 | m2 | m3 | m1 | m2 | m3 | m1 | m2 | m3 |
| Intercept | 0,017 | 0,026 | 0,019 | 0,015 | 0,016 | 0,013 | 0,229 | 0,377 | 0,289 | 0,163 | 0,223 | 0,223 |
| Age | 1,101* | 1,096* | 1,097* | 1,096* | 1,100* | 1,101* | 1,008 | 0,993 | 0,993 | 1,019 | 1,008 | 1,007 |
| Geographic repartition | | | | | | | | | | | | |
| Centre | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. |
| Islands | 1,224* | 1,321* | 1,377* | 1,377* | 1,566* | 1,659* | 1,651 | 1,631 | 1,742 | 1,711* | 1,570 | 1,586 |
| North-east | 1,211* | 1,158* | 1,159* | 1,062 | 1,015 | 1,030 | 1,370 | 1,317 | 1,327 | 0,893 | 0,911 | 0,905 |
| North-west | 1,095 | 1,051 | 1,068 | 1,054 | 1,025 | 1,049 | 1,086 | 1,099 | 1,122 | 1,011 | 1,006 | 1,005 |
| South | 1,120* | 1,193* | 1,223* | 1,240* | 1,355* | 1,407* | 1,307 | 1,314 | 1,344 | 1,404 | 1,333 | 1,342 |
| Year of enrolment | | | | | | | | | | | | |
| 2012 | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. | ref. |
| 2013 | 1,020 | 1,026 | 1,014 | 1,033 | 1,051 | 1,045 | 0,667 | 0,672 | 0,673 | 0,808 | 0,842 | 0,839 |
| 2014 | 1,083 | 1,102 | 1,086 | 1,138* | 1,172* | 1,160* | 0,890 | 0,919 | 0,896 | 0,809 | 0,823 | 0,823 |
| 2015 | 1,071 | 1,088 | 1,075 | 1,063 | 1,093 | 1,078 | 0,376* | 0,366* | 0,361* | 0,499* | 0,513* | 0,511* |
| 2016 | 1,022 | 1,033 | 1,013 | 1,003 | 1,023 | 1,001 | 0,512* | 0,528* | 0,524* | 0,480* | 0,496* | 0,496* |
| 2017 | 1,073 | 1,080 | 1,060 | 1,011 | 1,024 | 0,997 | 0,603* | 0,579* | 0,575* | 0,609* | 0,621* | 0,624* |
| 2018 | 1,013 | 1,011 | 0,999 | 0,975 | 0,979 | 0,961 | 0,557* | 0,558* | 0,556* | 0,447* | 0,480* | 0,483* |
| Job condition | | | | | | | | | | | | |
| Open-ended, FT[a] | | ref. | ref. | | ref. | ref. | | ref. | ref. | | ref. | ref. |
| Open-ended, PT[b] | | 0,629* | 0,630* | | 0,639* | 0,630* | | 0,517* | 0,508* | | 1,241 | 1,230 |
| Fixed-term, FT[a] | | 1,022 | 0,973 | | 1,021 | 0,993 | | 1,765* | 1,678 | | 0,704 | 0,704 |
| Fixed-term, PT[b] | | 0,725* | 0,704* | | 0,677* | 0,651* | | 0,695 | 0,665 | | 1,561 | 1,554 |
| Self-employed | | 0,742* | 0,684* | | 0,790* | 0,752* | | 0,609 | 0,580 | | 0,604 | 0,605 |
| Student | | 0,729* | 0,630* | | 0,938 | 0,785* | | 0,619* | 0,583* | | 0,559* | 0,553* |
| Jobseeker | | 0,604* | 0,597* | | 0,633* | 0,623* | | 1,153 | 1,144 | | 1,010 | 1,010 |
| Inactive | | 0,590* | 0,586* | | 0,495* | 0,488* | | 1,114 | 1,114 | | 1,418 | 1,408 |
| Household members | | | | | | | | | | | | |
| 2 | | | ref. | | | ref. | | | ref. | | | ref. |
| 3 | | | 1,083 | | | 1,015 | | | 1,146 | | | 1,056 |
| 4 | | | 1,175* | | | 1,026 | | | 1,055 | | | 0,905 |
| 5+ | | | 1,203* | | | 1,185* | | | 1,158 | | | 1,020 |
| Parental education | | | | | | | | | | | | |
| Up to middle school | | | ref. | | | ref. | | | ref. | | | ref. |
| Secondary school | | | 1,242* | | | 1,240* | | | 1,342* | | | 1,070 |
| University | | | 1,804* | | | 1,886* | | | 1,411 | | | 1,059 |

[a] Full time; [b] Part time; * $p<0.05$

Finally, the inclusion of variables on family size and parental education does not modify the effects previously described. We observe a significant effect of the number of household members and the parental education exclusively among Italian men and women: individuals living in large families and with parents having a middle/high education show a greater likelihood of leaving the parental home during the 15 months following the first interview. Since in Italy income is strongly and positively connected with the education, and since income is not a variable available for all the families interviewed, this result could also indicate that leaving the parental home is easier for the children of wealthier parents.

The present work contributes to the study of transition out of the parental home by differentiating according to young adults' gender and citizenship. Unfortunately, the information collected by ILFS is less

DRAFT

rich than that available for other dedicated surveys (such as Family and Social Subjects), and we do not have the characteristics of young people after they leave home (for example, we do not know their working conditions and their living arrangement). However – unlike what happens for dedicated surveys – ILFS has the advantage of being available every quarter of year, a few months after the survey, with large, statistically controlled samples, because they are used to measure employment and unemployment rates. It is therefore our intention to also analyse the data for subsequent years, for example by observing the effect on leaving home of the Covid-19 epidemic of 2020-21, and the subsequent recovery in employment.

## References


[1] Aassve, A., Arpino, B., Billari, F.C.: Age norms on leaving home: Multilevel evidence from the European Social Survey. Environ. Plan A **45**(2), 383--401 (2013)

[2] Aassve, A., Billari, F.C., Mazzuco, S., Ongaro, F.: Leaving home: A comparative analysis of ECHP data. J. Eur. Soc. Policy **12**(4), 259--275 (2002)

[3] Aassve, A., Billari, F.C., Ongaro, F.: The impact of income and employment status on leaving home: Evidence from the Italian ECHP sample. Labour **13**(3), 501--529 (2001)

[4] Barbieri, P., Cutuli, G., Scherer, S.: Giovani e lavoro oggi. Uno sguardo sociologico a una situazione a rischio. Sociologia del lavoro (Young people and work today. A sociological look at a risk situation. Sociology of work). Issue 136, pp. 73-98 (2014)

[5] Becker, S.O., Bentolila, S., Fernandes, A., Ichino, A.: Youth emancipation and perceived job insecurity of parents and children. J. Popul. Econ. **23**(3), 1047--1071 (2010)

[6] Bertolini, S., Bolzoni, M., Ghislieri, C., Goglio, V., Martino, S., Meo, A., Moiso, V., Musumeci, R., Ricucci, R., Torrioni, P.M.: Labour market uncertainty and leaving parental home in Italy. In: Baranowska-Rataj, A., Bertolini, S., Goglio, V. (eds.) Country level analyses of mechanisms and interrelationships between labour market insecurity and autonomy. EXCEPT Working Paper No. 11, pp. 16-40. Tallin University, Tallin (2017)

[7] Bertolini, S., Goglio, V.: Job uncertainty and leaving the parental home in Italy: Longitudinal analysis of the effect of labour market insecurity on the propensity to leave the parental household among youth. Int. J. Sociol. Soc. 39(7/8), 574--594 (2019)

[8] Billari, F.C., Castiglioni, M., Casto Martin, T., Michielin, F., Ongaro, F.: Household and Union Formation in a Mediterranea Fashion: Italy and Spain. In: Klijzin, E., Corijn, M. (eds.) Dynamics of fertility and partnership in Europe, vol. II, 1-16, UNECE, New York and Geneva, United Nations (2002).

[9] Billari, F.C., Liefbroer, A.C.: Towards a new pattern of transition to adulthood? Adv. Life Course Res. **15**(2–3), 59--75 (2010)

[10] Blossfeld, H.-P., Huinink, J.: Human Capital Investments or Norms of Role Transition? How Women's Schooling and Career Affect the Process of Family Formation. Am. J. Sociol. **97**(1), 143--168 (1991)

[11] Chiuri, M.C., Del Boca, D.: Home-leaving decisions of daughters and sons. Rev. Econ. Househ. **8**(3): 393--408 (2010)

[12] Corijn, M., Klijzing, E. (eds.) Transition to adulthood in Europe. Kluwer Academic Publishers, Dordrecht (2001)

[13] Erola, J., Jalonen, S., Lehti, H.: Parental education, class and income over early life course and children's achievement. Res. Soc. Stratif. Mobil. **44**, 33--43 (2016)

[14] Fernandes, A., Becker, S.O., Bentolila, S., Ichino, A.: Income Insecurity and Youth Emancipation: A Theoretical Approach. B.E. J. Econ. Anal. Policy. **8**(1), (2008)

[15] Furstenberg, F. F. Jr.: On a new schedule: transitions to adulthood and family change. Future Child. **20**(1), 67--87 (2010)

[16] Iacovou, M.: Leaving home: Independence, togetherness and income. Adv. Life Course Res. **15**(4).147--160 (2010)

[17] Krahn, H.J., Chai, C.A., Fang, S., Galambos, N.L., Johnson, M.D.: Quick, uncertain, and delayed adults: timing, sequencing and duration of youth-adult transitions in Canada. J. Youth Stud. **21**(7), 905--921 (2018)

[18] Liefbroer, A.C., Billari, F.C.: Bringing norms back in: A theoretical and empirical discussion of their importance for understanding demographic behaviour. Popul. Space Place **16**(4), 287--305 (2010)

[19] Liefbroer, A.C., Toulemon, L.: Demographic perspectives on the transition to adulthood: An introduction. Adv. Life Course Res. **15**(2), 53--58 (2010)

[20] Mazzuco, S., Ongaro, F.: Parental separation and family formation in early adulthood: Evidence from Italy. Adv. Life Course Res. **14**, 119--130 (2009)

[21] Meggiolaro S., Ongaro F.: Leaving Home over the Recent Cohorts in Italy: Does Economic Vulnerability Matter? SocArXiv (2022)

[22] Mulder, C.H., Clark, W.A.V.: Leaving home and leaving the State: Evidence from the United States. Int. J. Popul. Geogr. **6**(6), 423--437 (2000)


DRAFT

[23] Mulder, C.H., Hooimeijer, P.: Leaving home in the Netherlands: Timing and first housing. J. Hous. Built Environ. **17**(3), 237--268 (2002)
[24] Osgood, D.W., Ruth, G., Eccles, J.S., Jacobs, J.E., Barber, B.L.: Six paths to adulthood: Fast starters, parents without careers, educated partners, educated singles, working singles, and slow starters. In: Settersten, R.A., Furstenberg, F.F., Rumbaut R.G. (eds.). On the frontier of adulthood: Theory, research, and public policy. pp. 320-355. University of Chicago Press, Chicago (2005)
[25] Ranci, C., Brandsen, T. Sabatinelli, S.: Social Vulnerability in European Cities: The Role of Local Welfare in Times of Crisis. Palgrave Macmillan (2014)
[26] Schwanitz, K.: The transition to adulthood and pathways out of the parental home: A cross-national analysis. Adv. Life Course Res. **32**, 21--34 (2017)
[27] Schwanitz, K., Mulder, C.H., Toulemon, L.: Differences in leaving home by individual and parental education among young adults in Europe. Demographic Res. **37**, 1975--2010 (2017)
[28] Settersten, R.A., Furstenberg, F.F., Rumbaut, R.G.: On the frontier of adulthood: Theory, research, and public policy. University of Chicago Press, Chicago (2005)
[29] Shanahan, M.J.: Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective, Annu. Rev. Sociol. **26**(1), 667--692 (2000)
[30] Tosi, M.: Age norms, family relationships, and home-leaving in Italy. Demographic Res. **36**(9), 281--306 (2017)
[31] Ward, R.A., Spitze, G.D.: Nestleaving and coresidence by young adult children: The role of family relations. Res. Aging **29**(3). 257--277 (2007)
[32] Whittington, L.A., Peters, H.E.: Economic incentives for financial and residential independence. Demography **33**(1), 82--97 (1996)

DRAFT

# A Bayesian weather–driven spatio–temporal model for $PM_{10}$ in Lombardy

Michela Frigeri[1], Alessandra Guglielmi[1], and Giovanni Lonati[2]

[1]Department of Mathematics, Politecnico di Milano, Milano, Italy;
michela.frigeri@polimi.it, alessandra.guglielmi@polimi.it
[2]Department of Civil and Environmental Engineering, Politecnico di Milano, Milano, Italy;
giovanni.lonati@polimi.it

DRAFT

**Abstract**

Po Valley is well known to be one of the most polluted areas in Italy, because of its large population density, its shape and climate. Thus, there is an obvious interest in monitoring the air quality in several stations scattered across the whole territory. In this work, we develop a Bayesian spatio–temporal model describing the $PM_{10}$ pollution in Lombardy to assess how station features and weather factors affect the $PM_{10}$ concentration. We will rely on *Stan* for posterior inference.

***Keywords:*** Bayesian inference, Gaussian processes, auto-regressive hierarchical model

## 1. Introduction

*Air pollution* is defined as the release by human activities of gases and particulate matter (PM) into the atmosphere. Due to their impact on human health, we are interested in studying fine dusts, which are characterized by their micrometre size $PM_{10}$ to $PM_{2.5}$. *Po Valley* is a well-known hotspot for PM pollution in Europe (3). Many factors might be responsible: for instance the high population density, together with a high level of urban and industrial areas, its geographic shape and climate. In this work we consider air quality data of the Po valley, specifically focusing on $PM_{10}$ in Lombardy. $PM_{10}$ ambient concentration data are collected as daily averages by fixed monitoring stations at 64 different sites, which are distributed across the whole region. Given the well established relation between meteorological conditions and air pollution, we considered also weather-related data collected by a dedicated monitoring network in Lombardy. Both datasets store spatio-temporal information thanks to the geolocated monitoring stations providing multiple time series, as for each location we have both air quality and meteorological data. The spatial dependence characterizing point-referenced data is usually modeled via Gaussian processes (GP), though there is a plenty of different models for (multiple) time series analysis in the Bayesian setting (see (8)). In this paper we propose a Bayesian hierarchical model for the analysis of $PM_{10}$ spatio-temporal data in the Po valley, including also the weather as influential factor. The model also accounts for clustering of the monitoring stations, since we want to assess if there are differences in the magnitude of $PM_{10}$ dependence on past values across different locations. The paper is organized as follows: in Section 2. we briefly describe the two datasets we analyze; then in Section 3. we present the Bayesian model for $PM_{10}$ data and in Section 4. we provide initial findings based on posterior inference along with possible future developments.

## 2. The data

The dataset we analyze has been recorded by the air quality monitoring network of the Po Valley, the area of interest. This network is managed autonomously by the Regional Agencies of Environmental Protection (ARPA) of Emilia Romagna, Lombardy, Piedmont and Veneto. Data about concentrations of pollutants are collected through fixed monitoring stations, distributed over the whole valley. Each station collects, through time, the concentration of many pollutants, in particular: nitrogen dioxide ($NO_2$), benzene ($C_6H_6$), ammonia ($NH_3$) and particulate matter ($PM_{10}$, $PM_{2.5}$). In particular, we focus on $PM_{10}$ values, which are recorded as daily averages in all regions.

The dataset also provides information about the $PM_{10}$ monitoring sites. For each monitoring station, we know its location (lat–long coordinates and altitude on sea level), whether or not the station is exposed to a specific emission source (traffic, industry or background station) and the level of urbanization of the surroundings (urban, suburban or rural area). Data also provide the altitude of the station, which could be a relevant factor affecting the $PM_{10}$ concentration. Exploratory data analysis of this dataset can be found in (5). It is well-known that meteorological conditions might drive $PM_{10}$ concentration (e.g. (7)). Consequently, focusing only on Lombardy, we have considered data from the meteorological network, which is also managed by ARPA and involves over 200 weather stations across the region. However, this second dataset is recorded at different sites and time scale than the $PM_{10}$ dataset. To overcome the first problem, we associated each $PM_{10}$ station to the closest meteorological station. Secondly, $PM_{10}$ values are provided as daily averages while meteorological factors are recorded every 10 minutes. Adopting proper summaries for the various phenomena (e.g. rainfall, wind speed, solar radiation etc.) we have defined average daily values also for weather information, making the two datasets at the same spatial and temporal scale.

## 3. The Bayesian model

We focus on modeling $PM_{10}$ daily concentrations collected by all stations in Lombardy during 2018. Since we know the exact location of each monitoring station (lat–long coordinates), we model the $PM_{10}$ time-series as point-referenced spatio-temporal data. In this context many interesting models can be found in literature, see (8) Chapter 7. Some of them may be generalized and tailored for the data we focus on. We denote by $y(\boldsymbol{s}, t)$ the logarithm of $PM_{10}$ concentration collected at location $\boldsymbol{s}$ at time $t$, for $\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^d$, with $d = 2$ and for $t \in [0, T] \subset \mathbb{R}^+$. In our analysis the spatial domain $\mathcal{D}$ is given by Lombardy territorial extension, while each time instant $t$ corresponds to a Julian day of the year. We rearrange our data as the collection of temporal independent spatial processes, one for each time instant, adopting the following vector notation:

$$\boldsymbol{Y}_t = [y(\boldsymbol{s}_1, t), \ldots, y(\boldsymbol{s}_n, t)]' \qquad t = 1, \ldots, T. \tag{1}$$

In particular, for our case study we set $n = 64$ (number of monitoring stations in Lombardy) and $T = 365$. We consider the following AR(1) model, specifying the auto regression on the centered random effects $\boldsymbol{O}_t$ and not directly on $\boldsymbol{Y}_t$ (see (1; 8; 9)):

$$\boldsymbol{Y}_t = \boldsymbol{O}_t + \boldsymbol{\epsilon}_t \qquad t = 1, \ldots, T \tag{2}$$

$$\boldsymbol{O}_t = diag(\boldsymbol{\rho})\boldsymbol{O}_{t-1} + X_t\boldsymbol{\beta} + \boldsymbol{w}_t \qquad t = 1, \ldots, T \tag{3}$$

where $\boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$ is the $n \times 1$ pure error term, $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]$ is the $1 \times n$ row vector of station-specific autoregressive parameters, $X_t$ is the $n \times p$ design matrix and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients. The *hierarchical centering* mean parameter $\boldsymbol{O}_t$ (see (6)) is the $n \times 1$ vector representing the *true* value corresponding to data $\boldsymbol{Y}_t$ and we assume an autoregressive structure for $\boldsymbol{O}_t$ as in (1). The spatial residual term $\{\boldsymbol{w}_t\}$ in (3) is assumed to be independent from the pure error terms $\{\boldsymbol{\epsilon}_t\}$ in (2).

DRAFT

For the spatial process $\{\boldsymbol{w}_t\}$ we assume the following Gaussian Process (GP), as in (8):

$$\boldsymbol{w}_t \overset{iid}{\sim} \mathcal{N}_n(\mathbf{0}, \sigma_w^2 H) \qquad t = 1, \ldots, T \tag{4}$$

$$H_{i,j} = exp\{ -\frac{1}{R} ||\boldsymbol{s}_i - \boldsymbol{s}_j||_e \} \qquad i, j = 1, \ldots, n. \tag{5}$$

Hence each $\boldsymbol{w}_t$ is modeled as a zero-mean GP having covariance matrix defined by the exponential correlation function, which depends on the Euclidean distance $||\cdot||_e$ between the monitoring sites. Moreover the spatial residuals are assumed to be temporally independent.

**Covariates** The design matrix $X_t$ is composed by spatio-temporal covariates, including spatial characteristics of the stations (fixed-time covariates) and information about meteorological phenomena recorded at time $t$. In particular we will consider as influential factors: the altitude of each station, a dummy variable specifying whether or not the station is located in a traffic or industrial zone, rainfall amount, wind speed, total solar radiation and a sinusoidal function with annual frequency in order to provide the mean of the response with the typical U-shaped behaviour of $\log(\text{PM}_{10})$ time-series.

**Prior** The model just described is then completed with suitable marginal prior distributions for each parameter involved in the spatio-temporal model, i.e. $\boldsymbol{\rho}$, $\boldsymbol{\beta}$, $\sigma_\epsilon^2$, $\sigma_w^2$. For the moment, we focus on a model that assumes all component of $\boldsymbol{\rho}$ as given by $\widetilde{\rho}$, a common random parameter, with $\widetilde{\rho} \sim \mathcal{N}(0, 1)$. In the Bayesian framework, there is no need to impose strict constraints about the autoregressive parameter domain, since we are not assuming a priori the stationarity (or non-stationarity) of the series. The rest of the parameters are given the following marginal priors:

$$\beta_j \overset{iid}{\sim} \mathcal{N}(0, 1) \quad j = 1, \ldots, p$$
$$\sigma_\epsilon^2 \overset{iid}{\sim} InvGa(a, b)$$
$$\sigma_w^2 \overset{iid}{\sim} InvGa(a, b)$$

It is a common choice to provide $\mathbb{E}[\sigma_\epsilon^2] = \mathbb{E}[\sigma_w^2] = 1$ and $Var[\sigma_\epsilon^2] = Var[\sigma_w^2] = 2$, leading to the prior hyperparameters fixed to $a = 3$, $b = 2$. Note that, in this context, the standard Gaussian prior for the regression parameters turns out to be sufficiently vague, since $\log(\text{PM}_{10})$ ranges between 0 and 5 and all the numerical covariates have been standardized. We complete prior specification assuming:

$$\boldsymbol{O}_0 \sim \mathcal{N}_n(\boldsymbol{m}_0, \sigma_0^2 H)$$
$$\boldsymbol{O}_t \sim \mathcal{N}_n(\widetilde{\rho}\, \boldsymbol{O}_{t-1} + X_t \boldsymbol{\beta}, \sigma_w^2 H) \qquad t = 1, \ldots, T.$$

We fix $\boldsymbol{m}_0$ to be a $n \times 1$ vector having all components equal to the overall mean $\log(\text{PM}_{10})$ value, while $\sigma_0^2 \sim InvGa(3, 2)$ providing $\mathbb{E}[\sigma_0^2] = 1$ and $Var[\sigma_0^2] = 2$. Block of parameters are assumed a priori independent. Finally, we fix the spatial range parameter in expression (5) to $R = 70km$ (estimated through empirical variogram techniques).

**BNP Clustering** Assuming a convenient BNP prior for $\rho_1, \ldots, \rho_n$, we can cluster together stations showing a similar trend in their time-series, providing interesting information about the data outline. We assume a Dirichlet process mixture (DPM) model (see (4)) for $\rho_1, \ldots, \rho_n$ without modifying any of the other terms in model (2)-(3). The DPM model is then defined as:

$$\rho_i \overset{iid}{\sim} \sum_{g=1}^{+\infty} \eta_g \mathcal{N}(m_g, s_g^2) \quad i = 1, \ldots, n. \tag{6}$$

DRAFT

We also assume the mixture hyperparameters as $(m_g, s_g^2) \overset{iid}{\sim} P_0$ , $g = 1, \ldots, G$ , where $P_0$ is a suitable distribution on $\mathbb{R} \times (0, +\infty)$. The mixture weights $\{\eta_g\}$ come from the truncated stick-breaking representation of a Dirichlet process and can be written as:

$$\eta_1 = v_1 \tag{7}$$

$$\eta_g = v_g \prod_{j=1}^{g-1} (1 - v_j) \qquad g = 2, \ldots, G \tag{8}$$

$$v_g \overset{iid}{\sim} Beta(1, \alpha) \qquad g = 1, \ldots, G \tag{9}$$

where G is fixed and the total mass parameter is $\alpha \sim Gamma(c, d)$. It can be proven that the weights sum to one, i.e. $\sum_{g=1}^{G} \eta_g = 1$.

## 4. Preliminary data analysis and posterior inference

The dataset collected in Lombardy during 2018 contains daily average concentrations of $PM_{10}$ at the 64 monitoring stations for every day of the year ($T = 365$). In addition we also have weather information coming from the nearest meteorological station, providing data for each day of the year. Specifically, we consider as influential factors the daily averages of *solar radiation intensity*, *humidity* and *temperature*. The information about the presence of *rain* or *wind* is instead summarized by two binary variables: one indicating very rainy days (more than 8 over 24 hours raining) and the other indicating days of calm (more than 8 over 24 hours without wind). These are the time-varying covariates $X_t$ in (3). We assume as regressors of interest also the *altitude* of each $PM_{10}$ station and a binary variable indicating the proximity of *traffic* or *industrial* sources of pollution. These latter are fixed-time covariates. We fix prior hyperparameters as follows: $a = 3$, $b = 2$, $c = d = 3$. We rely on *Stan* probabilistic programming language (2) for posterior simulation, i.e for computing the posterior of all parameters involved in model (2)-(9). We run two parallel MCMC chains in *Stan*, assuming $1,000$ burn-in iterations and $1,000$ sampling iterations. Figure 1 displays $95\%$ posterior credible intervals for the regression coefficients. Note that all regressors are quite significant. As expected, air quality improves with altitude, while traffic and industrial areas are associated to higher levels of pollution. As for as weather covariates are concerned, we see that rainy days and high temperatures are associated to lower $PM_{10}$ concentrations, while the absence of wind, together with high solar radiation and humidity, contributes to raise the pollution level.

DRAFT

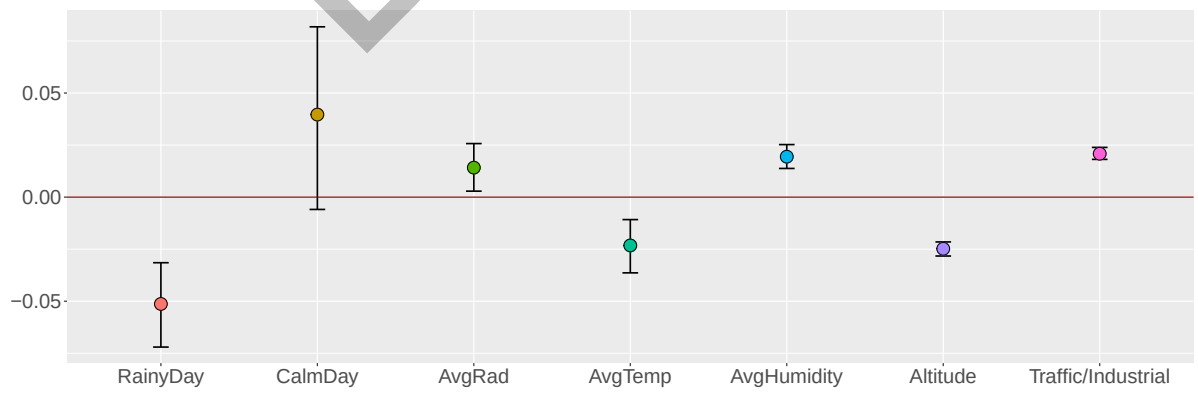



Figure 1: 95% marginal posterior credible intervals (CI) for regression coefficients $\boldsymbol{\beta}$.

Figure 2 shows the model $95\%$ credible intervals of the posterior predictive distribution for $\log(\ PM_{10})$ at a specific station (*Pavia-Piazza Minerva*) where the last 30 days of the year have been removed from training set. The model provides reasonable prediction for the $\log(\ PM_{10})$ values in these latter days,

since all the true values (in blue) are fully inside the 95% predicted credible intervals. We obtain similar results for the majority of locations, but few stations seem to follow a different trend. Figure 3 shows the 95% credible intervals of the posterior predictive distribution for $\log(\ PM_{10})$ at *Moggio*, one of the stations whose posterior predictive distributions seem different from all the other ones. As in Figure 2, we removed the last 30 days from the training set and computed the posterior predictive distribution of $\log(\ PM_{10})$ in these latter days, using the model here described. It is clear that most of the datapoints (blue line) are not included in the associated 95% credible intervals. These patterns suggest the presence of at least two different temporal trends among the stations under study, corroborating the usefulness of a Bayesian cluster model able to find all the possible different scenarios. Future work includes implementing in *Stan* the Bayesian non-parametric clustering presented in Section 3.

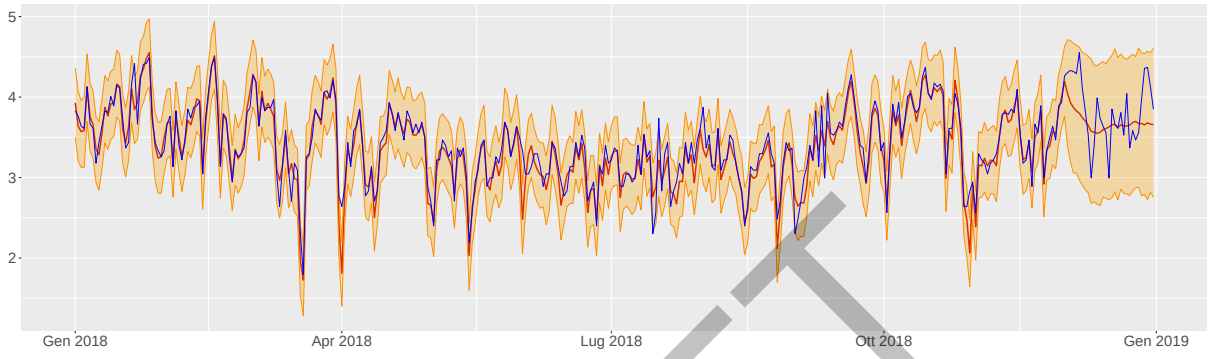



Figure 2: Observed values (in blue) , 95% credible intervals (in orange) and median (in red) of the posterior predictive distribution for $PM_{10}$ log-concentrations at Pavia - Piazza Minerva.

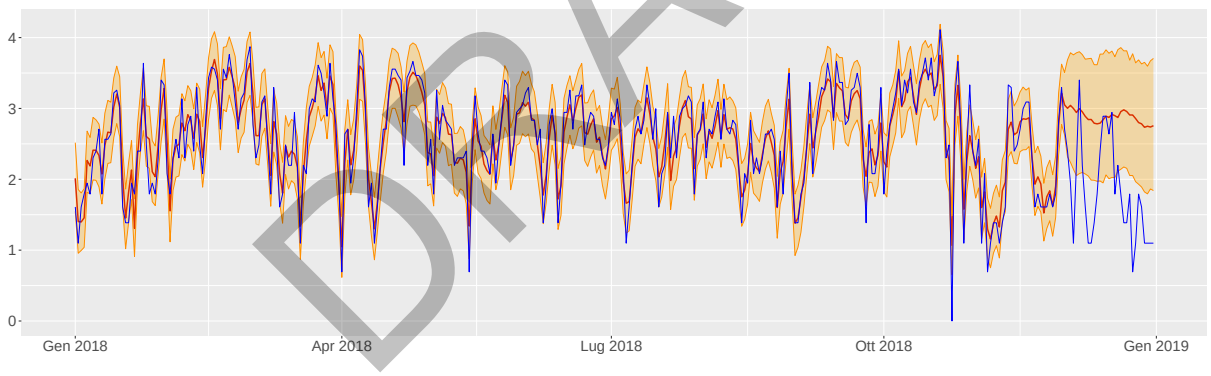



Figure 3: Observed values (in blue) , 95% credible intervals (in orange) and median (in red) of the posterior predictive distribution for $PM_{10}$ log-concentrations at Moggio.

DRAFT

## References


[1] Bakar, K.S., Sahu, S.K.: spTimer: Spatio-temporal bayesian modeling using R. Journal of Statistical Software **63**(15), 1–32 (2015) doi: 10.18637/jss.v063.i15

[2] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017) doi: 10.18637/jss.v076.i01

[3] EEA: Air quality in Europe - 2019 Report: Technical Report. Tech.rep., European Environmental Agency (EEA) (2019)

[4] Escobar, M., West, M.: Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association **90**(430), 577–588 (1995) doi: 10.2307/2291069

[5] Frigeri, M.: Spatio-Temporal Models for Particulate Matter in the Po Valley. Thesis, M.Sc. in Mathematical Engineering, Politecnico di Milano (2022).
http://hdl.handle.net/10589/195435
[6] Gelfand, A. E., Sahu, S. K., Carlin, B. P.: Efficient Parametrisations for Normal Linear Mixed Models. Biometrika **82**(3), 479–488 (1995).
https://doi.org/10.2307/2337527
[7] Grange, S.K., Carslaw, D.C., Lewis, A.C., Boleti, E., Hueglin, C.: Random forest meteorological normalisation models for Swiss $PM_{10}$ trend analysis. Atmospheric Chemistry and Physics **18**(9), 6223–6239 (2018) doi: 10.5194/acp-18-6223-2018
[8] Sahu, S.K.: Bayesian modeling of spatio-temporal data with R. Chapman and Hall/CRC (2022)
https://doi.org/10.1201/9780429318443
[9] Sahu, S.K., Gelfand, A.E., Holland, D.M.: High-resolution space-time ozone modeling for assessing trends. Journal of the American Statistical Association **102**(480), 1221–1234 (2007)
doi: 10.1198/016214507000000031.

DRAFT

# A preliminary study on shape descriptors for the characterization of microplastics ingested by fish

Greta Panunzi[a], Tommaso Valente[b,c], Marco Matiddi[c], and Giovanna Jona Lasinio[a]

[a]Department of Statistical Sciences,University of Rome "La Sapienza";
greta.panunzi@uniroma1.it, giovanna.jonalasinio@uniroma1.it
[b]Department of Environmental Biology, University of Rome "La Sapienza";
tommaso.valente@uniroma1.it
[c]ISPRA, Italian National Insitute for Environmental Protection and Research.
marco.matiddi@isprambiente.it

DRAFT

**Abstract**

In this study we investigate the reliability of shape descriptors in providing a more detailed characterization of microplastics ingested by fish. In particular, a new approach based on shape descriptors is compared with the traditional approach, in the explanation of species-specific selection mechanisms in the ingestion of MPs. In accordance with the traditional information available, the first question is whether the rate of ingestion depends on the species. Correspondence analysis was then performed to highlight possible differences in the types of MPs ingested by the three species in terms of shape categories, color and size classes. The traditional approach does not provide enough information on the selection mechanism of MPs. New quantitative shape descriptors were therefore used to identify a relationship between MPs characteristics and species. PCA technique and Kruskal-Wallis test were used.

***Keywords:*** Hurdle model, Correspondence analysis, Kruskall-Wallis, Microplastics, Fish

## 1. Introduction

Microplastics (MPs) are commonly defined as little pieces of plastic less than 5 mm in size. MPs are documented as pervasive and ubiquitous environmental contaminants (Arthur et al., 2009). MP pollution is one of the main environmental issues of current times (Horton et al. , 2017). Most studies about MP emphasize contamination of oceans and seas, paying particular attention to the ingestion of these particles by marine animals (Gouin , 2020). Despite the low toxicity associated to most plastic polymers, MPs are noxious particles that can negatively affect the physiology of marine organisms by determining physical damages (such as the obstruction of the gastrointestinal tract) or acting as carriers of hazardous chemicals like plastic additives (e.g., plasticizers, flame retardants, and dyes) or other adsorbed pollutants (Rai et al., 2021). Since MPs overlay the size range of prey of many marine species, MP ingestion may happen either accidentally, or intentionally (i.e., by confusing plastic particles with natural or potential preys), as well as due to trophic transfer (i.e., secondary ingestion of MPs already ingested by prey) (Fossi et al., 2018). In other words, MP ingestion maybe not depend only on its detectability, but likely also on the foraging decision-rules of organisms. The full comprehension of patterns that define the

environmental fate of MPs is a key step for assessing the impact of plastic pollution on marine biota and to drive effective management policies (Valente et al., 2020). Current methods to characterize MPs extracted from biological samples are based on visual identification and classification according to shape categories and size classes. The qualitative operator-based classification of MPs often implies loss of objective information and is subject to observer bias (Primpke et al., 2017). Furthermore, MP categories used in different studies are not always congruent and clearly defined, resulting in the lack of a standard, globally shared glossary (Miller et al., 2020). Then, the use of categorical variables to describe the shape and size of MPs may limits cross-studies consistency, and therefore our understanding of the fate of different MP types within ecosystems (Cowger et al., 2020).

Shape descriptors are quantitative measures describing the geometry of a given shape. Although a particle cannot be redrawn from shape descriptors, these should be sufficiently different to distinguish specific geometric features. Therefore, image processing approaches aimed at shape quantification are regarded as good candidates for developing a new MP characterization framework (Valente et al., 2022a). In this study we perform the first investigation on the reliability of shape descriptors in providing a more detailed description of MPs ingested by marine fish. To hit this goal, we analyzed MPs ingested by three different fish species from the upper slope (400-600 m depth) of the Tyrrhenian Sea (Western Mediterranean), namely the Shortnose greeneye *Chlorophthalmus agassizi*, the Mediterranean slimehead *Hoplostethus mediterraneus*, and the Hollowsnout grenadier *Coelorinchus caelorhincus*. Following the traditional approach, the ingested MPs were firstly counted and classified according to customary shape categories and size classes. Then, all the MPs were photographed using a camera-equipped dissecting microscope and the obtained pictures were processed to obtain a shape characterization based on the following shape descriptors: surface area, perimeter, major axis, minor axis, aspect ratio, solidity, and circularity. The information obtained from the two different MP characterization approaches were compared to verify the ability of shape descriptors in highlighting differences in the amount and diversity of MP types ingested by the three species.

DRAFT

## 2. Materials and Methods

**Sampling and laboratory analyses** A total of 90 fish (30 for each species) were sampled during a single fishing trip carried out in April 2021 out off Rome (41°20′24.5″ N; 012°17′20.6″ E), within a marine area representing an optimal pilot zone for studies on MP ingestion. Indeed, this is a prospective site of MP accumulation due to the discharge of the Tiber River, high coastal urbanization, and very variable current patterns creating unique conditions for the accumulation of waste from local inputs (Valente et al., 2019, 2022b). Laboratory analyses were performed following the guidelines provided in Matiddi et al. (2021).

**Characterization of microplastics and data reporting** Following the traditional approach, all the MPs extracted from the gastrointestinal contents were visually identified, counted, and classified according to the customary shape categories (*i.e.*, fiber, filament, film, fragment, foam, granule, and pellet) and size classes (S1: 100 - 330 $\mu$m; S2: 330 $\mu$m - 1 mm; S3: 1 - 5 mm) defined within the European research project INDICIT 2 (`https://indicit-europa.eu`). Therefore, the amount of MP ingested was expressed as no. of MPs ingested by each individual, while the diversity of ingested MP types was evaluated according to the combination of the categorical variables describing the shape and size of each particle. Thereafter, all the MPs were photographed using a ZEISS Stemi 2000-C dissecting microscope equipped with an Axiocam 208 color camera. Following the image analysis protocol developed in Valente et al. (2022b), the obtained pictures were processed using the open-source software ImageJ (`https://imagej.nih.gov.ij/`) to compute the following shape descriptors: surface area (i.e., no. of pixels in the particle), perimeter (no. of pixels in the boundary of the particle), aspect ratio (height-to-width ratio, $\frac{\text{major axis}}{\text{minor axis}}$), solidity (ratio of the area of an object to the area of the convex hull of the object, $\frac{\text{area}}{\text{convex area}}$), and circularity (ratio of the particle's area to the area of a circle with the same perimeter, $4\pi \cdot area \cdot perimeter^2$). Surface area measurements were used to both assess the amount of ingested MP and the size of the particles. All the other shape descriptors were considered to summarize

the geometric features of each MP and highlight the variety of MP types ingested by the three species.

**Statistical analyses** The workflow diagram in Figure 1 describe the experimental set-up developed to compare the two MP characterization approaches.

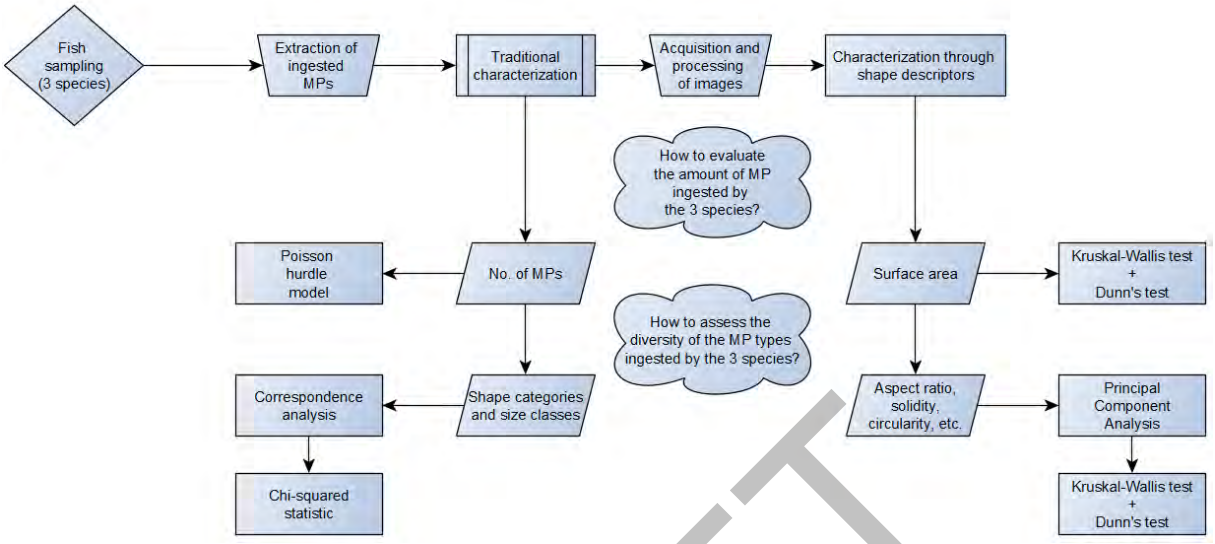



Figure 1: Workflow

According to the traditional approach, the amount of MP ingested by the three species was expressed as no. of MPs found in each individual. Given the large proportion of 0s coupled with a right-skewed distribution, the Poisson hurdle model (Mullahy, 1986) was used to test for differences in microplastic ingestion rates among the three species. Let's consider $Y_i$ the number of microplastics ingested by the subject $i = 1, \ldots, n$. The Hurdle model is given by

$$\begin{aligned} P(Y_i = 0) &= 1 - p, & 0 \leq p \leq 1 \\ P(Y_i = k) &= p \cdot \frac{\mu^k e^{-\mu}}{k!\{1 - e^{-\mu}\}} & k = 1, \ldots, \infty, \quad 0 < \mu < \infty \end{aligned} \tag{1}$$

With:

$$\begin{aligned} log\left(\frac{p}{1-p}\right) &= \alpha_0 + \mathbf{X}^T\alpha \\ log\left(\mu\right) &= \beta_0 + \mathbf{X}^T\beta. \end{aligned} \tag{2}$$

were $\mu$ is the mean of an untruncated Poisson distribution, $\mathbf{X}$ is a set covariates with linear effects $\alpha$ and $\beta$ to be estimated.

Correspondence analysis (Greenacre, 1984) was performed to highlight possible differences in the types of MPs ingested by the three species in terms of shape categories, color and size classes. Association between variables was tested through chi-square statistics. Shape descriptors were analyzed in parallel to test their ability in detecting possible differences in MP ingestion among the three examined species. PCA was performed to investigate the relationship between shape descriptors and species. Then, due to the invalidation of one of the assumptions for parametric analysis (namely homoscedasticity, as tested by Levene's test), the Kruskal-Wallis test for multiple comparisons (Kruskal et al., 1952) was used. Post-hoc multiple pairwise comparison based on Dunn's z-test-statistic with Bonferroni correction (Dunn, 1964) was finally performed whether significant differences among species were detected.

All statistical analyses were performed with R v4.2.2 (R Core Team , 2021). The significance level was set at $p<0.05$ for all analyses.

DRAFT

## 3. Results

MP ingestion was detected in all the three examined species, with an overall frequency of occurrence of 34.4% and an average number of microplastics per individual equal to 0.56 $\pm$ 1.03 (1.68 $\pm$ 1.11 considering only the individuals with ingested MP). Results by species are available in Table 1.

Table 1: Summary statistics by species

| | C. agassizi | H. mediterraneus | C. caelorhincus |
|---|---|---|---|
| Frequency of Occurrence | 26.7% | 40.0% | 36.7% |
| Average number of MPs ($\pm$ sd) | 0.53 $\pm$ 1.20 | 0.70 $\pm$ 1.09 | 0.50 $\pm$ 0.78 |
| Average number of MPs ($\pm$ sd) (individuals with ingested MP) | 2 $\pm$ 1.60 | 1.75 $\pm$ 1.06 | 1.36 $\pm$ 0.67 |

No significant differences in MP ingestion rates were detected among the three species. The estimated values for the Hurdle model parameters together with the standard deviations and the relevant p-values are shown in Table 2.

Table 2: Hurdle coefficients estimates

| **Count model coefficients** | mean | standard error | p-value |
|---|---|---|---|
| $\beta_0$ | 0.466 | 0.325 | 0.151 |
| $\beta_{C.caelorhincus}$ | -0.888 | 0.577 | 0.124 |
| $\beta_{H.mediterraneus}$ | -0.245 | 0.448 | 0.585 |
| **Zero hurdle model coefficients** | mean | std | p-value |
| $\alpha_0$ | -1.012 | 0.412 | 0.014 |
| $\alpha_{C.caelorhincus}$ | 0.465 | 0.560 | 0.407 |
| $\alpha_{H.mediterraneus}$ | 0.606 | 0.556 | 0.279 |

DRAFT

Correspondence analysis showed no significant association between species and the classical categorical variables "shape category", "color" and "size class". PCA performed on the computed shape descriptors well highlights that *C. caelorhincus* ingested the widest variety of MPs. In particular, it turns out that for the species *C.agassizi* and *H.mediterraneus* the main characteristic of the microplastics ingested is linked to the circularity and the roundness, while for *C.caelorhincus* influence the shape descriptors area and the Feret's diameter, all descriptors linked to the amplitude of the microplastic (Figure 2).
To confirm what the PCA showed, the Kruskall-Wallis test and the Dunn's test revealed that *C. caelorhincus* ingested lager MPs than *H. mediterraneus* and *C. agassizi* (Table 3).

Table 3: Dunn's test estimation

| Comparisons | Z | adjusted p-value |
|---|---|---|
| C.caelorhincus - H.mediterraneus | 2.65 | 0.012 |
| C.agassizi - C.caelorhincus | -1.95 | 0.045 |
| C.agassizi - H.mediterraneus | 0.54 | 0.88 |

Table 4 summarizes the main differences between the two approaches.

## 4. Discussion

The idea of the study is to compare a new approach (based on shape descriptors) with the traditional one in particularizing the ingestion of microplastics according to the species. This study further confirms

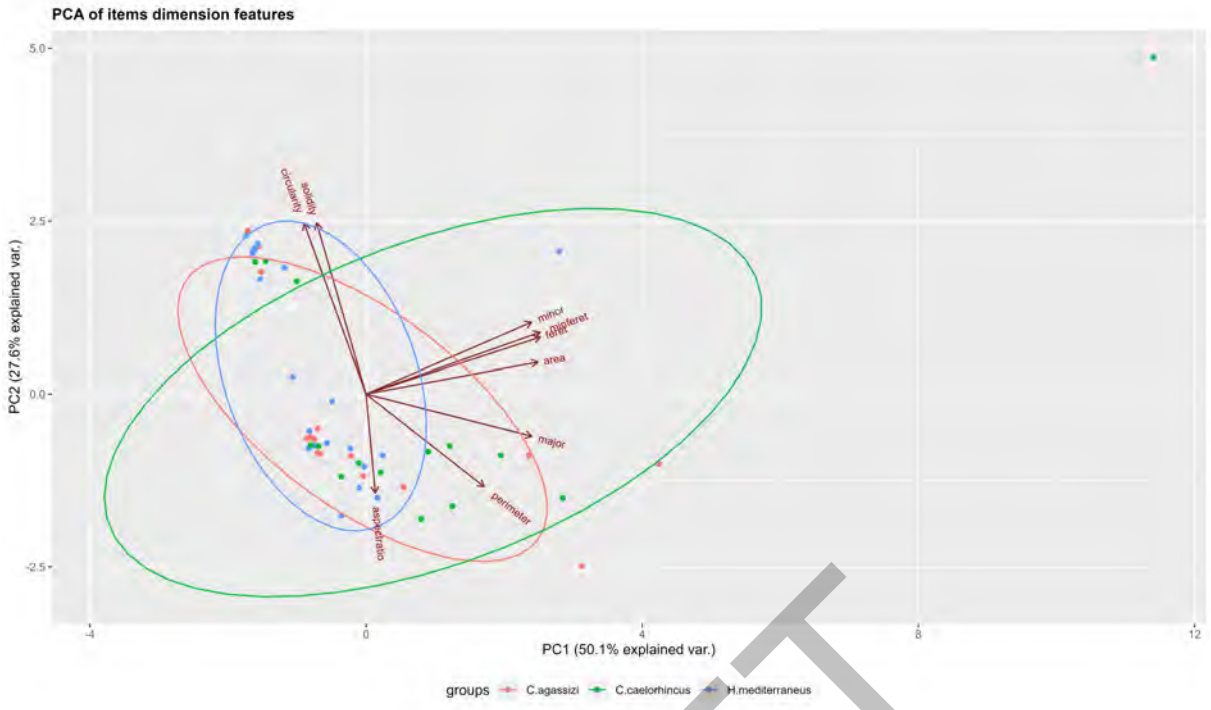



Figure 2: Principal Component analysis results.

Table 4: Differences between the two approaches

| Topic | Traditional approach | Shape descriptors | Detected differences (TA/SD) |
|---|---|---|---|
| Amount of ingested MPs | number of ingested MPs | surface area of ingested MPs | N/Y |
| Shape of MPs | shape categories (i.e. fiber, film, foam, fragment) | aspect ratio, circularity, solidity | N/N |
| Size of MPs | size classes | perimeter, major, minor | N/Y |

DRAFT

that MP ingestion by fish is widespread in deep-water habitats (Valente et al., 2019). Through a snapshot in time, our results highlight that MP ingestion rates show no remarkable differences according to the examined species, as previously highlighted in other marine compartments within the same area (Valente et al., 2022b). The use of the rate of ingestion and of the categorical variables of shape and color is not sufficient to discriminate the characteristics of the microplastics with respect to the species. However, differences among species emerge when considering the variety of ingested MP types using the quantitative shape descriptors. In particular, it was noted that *H. mediterraneus* ingesting smaller MPs than the other two species, and *C. caelorhincus* ingesting the highest diversity of MPs. The observed differences - likely due to the different feeding habits of the three species (Miller et al., 2020), show that the use of categorical variables to describe MPs limits the comprehension of the pathways that different MP types may follow through the marine food webs. According to our preliminary results, the characterization of MPs through shape descriptors could represent a simple way to translate categorical data into quantitative variables, avoiding the loss of information. In this view, a more detailed description of MPs could improve our knowledge about the diversity and distribution of MPs between and within environmental compartments. Although the shape cannot be redrawn from shape descriptors, these are sufficiently different to distinguish different shapes, allowing the clarification of the selection mechanisms that might

determine the uptake of these hazardous particles by marine organisms, such as size, circularity and solidity.

## References


Arthur, Courtney, Joel E. Baker, and Holly A. Bamford. "Proceedings of the International Research Workshop on the Occurrence, Effects, and Fate of Microplastic Marine Debris, September 9-11, 2008, University of Washington Tacoma, Tacoma, WA, USA." (2009).

Cowger, Win, et al. "Critical review of processing and classification techniques for images and spectra in microplastic research." *Applied spectroscopy* 74.9 (2020): 989-1010.

Dunn, Olive Jean. "Multiple comparisons using rank sums." *Technometrics* 6.3 (1964): 241-252.

Fossi, Maria Cristina, et al. "Bioindicators for monitoring marine litter ingestion and its impacts on Mediterranean biodiversity." *Environmental Pollution* 237 (2018): 1023-1040.

Gouin, Todd. "Toward an improved understanding of the ingestion and trophic transfer of microplastic particles: critical review and implications for future research." *Environmental Toxicology and Chemistry* 39.6 (2020): 1119-1137.

Greenacre, Michael J. "Theory and applications of correspondence analysis." (1984).

Horton, Alice A., et al. "Microplastics in freshwater and terrestrial environments: evaluating the current understanding to identify the knowledge gaps and future research priorities." *Science of the total environment* 586 (2017): 127-141.

Kruskal, William H., and W. Allen Wallis. "Use of ranks in one-criterion variance analysis." *Journal of the American statistical Association* 47.260 (1952): 583-621.

Libungan, Lísa Anne, and SnæbjBörn Pálsson. "ShapeR: an R package to study otolith shape variation among fish populations." PLoS One 10.3 (2015): e0121102.

Matiddi, M., et al. *Monitoring micro-litter ingestion in marine fish: a harmonized protocol for MSFD and RSCS areasEAS*. 2021.

Miller, Michaela E., Mark Hamann, and Frederieke J. Kroon. "Bioaccumulation and biomagnification of microplastics in marine organisms: A review and meta-analysis of current data." *PLoS One* 15.10 (2020): e0240792.

Mullahy, John. "Specification and testing of some modified count data models." *Journal of econometrics* 33.3 (1986): 341-365.

O'Connor, James D., et al. "Microplastics in freshwater biota: a critical review of isolation, characterization, and assessment methods." *Global challenges* 4.6 (2020): 1800118.

Primpke, Sebastian, et al. "An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis." *Analytical Methods* 9.9 (2017): 1499-1511.

Rai, Prabhat Kumar, et al. "Adsorption of environmental contaminants on micro-and nano-scale plastic polymers and the influence of weathering processes on their adsorptive attributes." *Journal of Hazardous Materials* (2021): 127903.

Team, R. Developement Core. "A language and environment for statistical computing." `http://www.R-project.org`(2009).

Valente, Tommaso, et al. "Exploring microplastic ingestion by three deep-water elasmobranch species: A case study from the Tyrrhenian Sea." *Environmental Pollution* 253 (2019): 342-350.

Valente, Tommaso, Umberto Scacco, and Marco Matiddi. "Macro-litter ingestion in deep-water habitats: is an underestimation occurring?." *Environmental Research* 186 (2020): 109556.

Valente, Tommaso, et al. "Image processing tools in the study of environmental contamination by microplastics: reliability and perspectives." *Environmental Science and Pollution Research* 30.1 (2023): 298-309.

Valente, Tommaso, et al. "One is not enough: Monitoring microplastic ingestion by fish needs a multispecies approach." *Marine Pollution Bulletin* 184 (2022): 114133.


DRAFT

# Artificial neural network in predicting odour concentrations: a case study*

V. Distefano[a] and G. Mazuruse[a,b]

[a]DES-Sect. of Mathematics and Statistics, University of Salento (Italy);
veronica.distefano@unisalento.it
[b]Marondera University of Agricultural Sciences and Technology (Zimbabwe);
gmazuruse@gmauast.ac.zw

**Abstract**

The reduction of emissions of unpleasant odour generated by treatment plants is one of the most crucial aspects in air quality monitoring. In this context, the estimation of odour concentrations from a multiparameter sensors system is a very complex procedure and in the literature most of the studies have applied multilinear regression methods to assess odour concentration. In this paper, an Artificial Neural Network (ANN) has been performed to estimate the odour concentrations from emission sources and identify an estimation function of the odour concentrations, starting from data collected by a multi-parametric set of 10 sensors of gaseous compounds.

***Keywords:*** Odour concentration, emission sources, multilayer perceptron

DRAFT

## 1. Introduction

In twenty-eight European Union (EU) countries, dangerous odour emissions are regulated through the Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on Industrial Emissions (the Industrial Emission Directive or IED). The IED establishes a general framework for determining limits, including odour limits for many industrial activities/processes intending to control odour emissions. The covered sectors include, for example, the energy industry, metals production and processing, waste management, chemical and mineral industry, and agriculture sectors such as animal production.

Olfactory pollution is one of the main reasons for citizens to complain about the environment since often it constitutes an important problem which negatively affects the quality of life (6). However, the monitoring and estimation of the odour emissions present complex aspects to deal, due to a) the nature of the phenomenon itself, which is the resulting of a mixture of numerous chemical substances, and b) the impossibility of carrying out continuous measurements of the odour concentration.

Many studies have been carried out to appreciate the odour concentration (1) and most of them were based on single or multilinear regression methods. However, more effort is still needed in this area. The most critical issue concerns the concrete difficult for a sensory system of acting to the olfactory stimulus with the same sensitivity as the human nose.

*The paper is one of results from the research project titled "Modelli intelligenti di correlazione per la valutazione dell'inquinamento olfattivo" (project code f3f76727) approved by the Regional Program "RIPARTI" of the Apulian Region.

On the basis of the above considerations, the present paper aims to illustrate the development of a multivariate statistical model which is able to estimate odour concentrations through the responses obtained from a multi-parametric system (IOMS - Instrumental Odour Monitoring System). An ANN will be properly implemented for the aim of the study and the main results will be discussed in the following sections.

## 2. The dataset

The analyzed data, which have been collected from a company specialized in environmental assistance and consultancy services for private and public enterprises, refer to the responses from 10 sensors of a multi-parametric system of odour monitoring. The sensor array is composed by 10 metal oxide semiconductors (MOS) used to detect the emitted volatile organic compounds (VOCs) at all the stages of the plants. These sensors reported in in Tab. 1 allow to detect specific odour compounds whose values are expressed in $mA$ (milliamperes). The data set consists of 354 measurements for 10 sensors at several stages of various treatment plants and the response variable is the odour concentration of gaseous compounds, expressed in $ou_E/m^3$ (European odour unit per $m^3$), The analyzed dataset does not contain missing values.

Table 1: Sensors description

| Label Sensor | Description |
|---|---|
| W2W | Are found in garlic and in crude oil.<br>Causes extreme global warming and acid rain |
| W1S | A colourless gas with a pungent odour.<br>Toxic to human and aquatic organisms. |
| W1C | Affect the skin, eyes, and respiratory tract.<br>Used in the production of paints and rubber |
| W5S | Produced by biogenic sources such as plants and yeasts.<br>Some are toxic, and all contribute to ozone. |
| W3C | A liquid that smells like gasoline and boils at 80 . |
| W6S | It is found in oil , human and animal waste, and<br>sewage treatment. Used for producing chemicals |
| W5C | Produced from crude oil refinement.<br>Causes headaches, dizziness, and even death. |
| W3S | A gas with important greenhouse gas<br>properties. Fuel production and engines. |
| W1W | A compound gas with distinctive aromatic flavours like citrus.<br>Prevent inflammatory diseases. |
| W2S | A poisonous gas. Originates from vehicle engines,<br>waste burning and forest wildfires. |

The exploration data analysis has been performed to detect the presence of possible extreme values (outliers) among the olfactometric measurements. Fig. 1 displays the box-plot for the distribution data before and after removing the outliers. The median absolute deviation (MAD) has been applied for removing outliers from the data set, which finally consists of 255 observations. Tab. 2 gives the descriptive statistics for the 10 sensors and the odour concentration. Sensors W1C, W3C, and W5C have registered the lowest mean values (0.76, 0.80, and 0.84, respectively), while sensors W1W and W2W the highest mean values (9.80 and 6.65, respectively). This very different behaviour of sensors could be attributed to their impact on odour concentration.

DRAFT

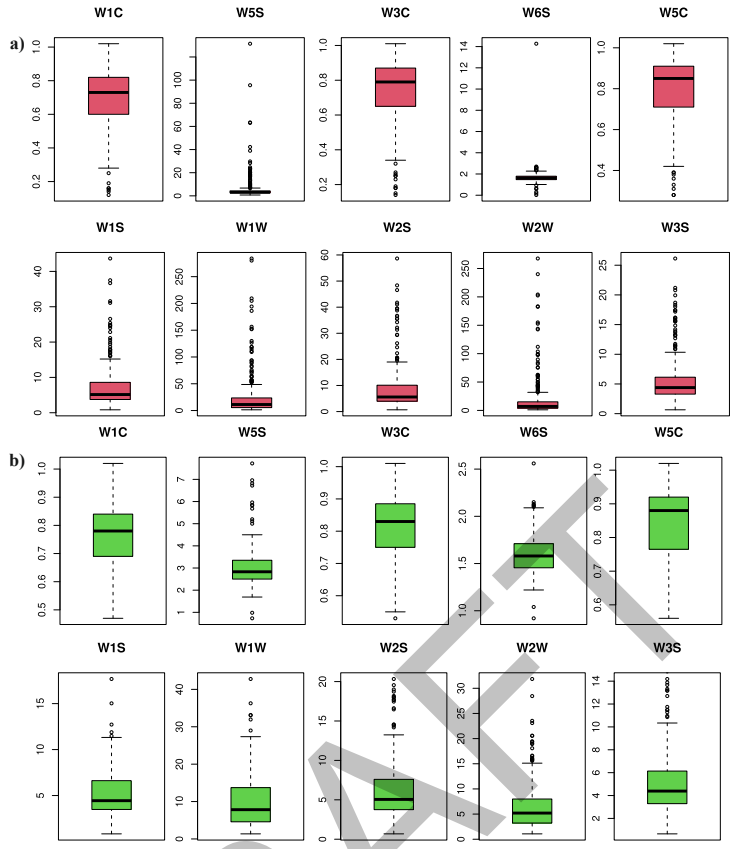



Figure 1: Box-plots of the distribution data related to the 10 sensors a) before removing the outliers, b) after removing the outliers.

DRAFT

Table 2: Descriptive statistics for the study variables

| Variable | Minimum | Mean | Median | Maximum | Std. Dev. |
|---|---|---|---|---|---|
| Odour concentration | 11 | 587 | 260 | 3153 | 707 |
| W1C | 0.47 | 0.76 | 0.78 | 1.02 | 0.10 |
| W5S | 0.73 | 3.03 | 2.83 | 7.72 | 0.92 |
| W3C | 0.53 | 0.80 | 0.83 | 1.01 | 0.10 |
| W6S | 0.92 | 1.59 | 1.58 | 2.56 | 0.20 |
| W5C | 0.56 | 0.84 | 0.88 | 1.02 | 0.10 |
| W1S | 0.83 | 5.30 | 4.44 | 17.68 | 2.66 |
| W1W | 1.36 | 9.80 | 7.83 | 42.78 | 6.92 |
| W2S | 0.68 | 6.38 | 5.05 | 20.33 | 3.91 |
| W2W | 1.07 | 6.65 | 5.20 | 31.85 | 4.87 |
| W3S | 0.65 | 4.51 | 4.01 | 14.19 | 2.13 |

## 2.1 MultiLayer Perceptron architecture

The MultiLayer Perceptron (MLP) is an ANN which is suitable to represent any smooth measurable functional relationship between the inputs (predictors) and the outputs (outcome). It is composed of at least three layers of neurons: the input layer that allocates the data in the network, the hidden layer(s) that processes the data and the output layer, namely the results extracted from specific inputs. The MLP could be represented as an Neural Simulation Language (3). Each hidden layer consists of nodes constructed by an activation function, which is used to transform summed input value of each neuron to its output value. In other words, the activation function is a function of input that the neuron receives, in order to convert the input signal on the node of ANN to an output signal.
The MLP architecture is characterized by its simple design and can be used for solving various classification and regression problems (2).
In the present research, the hyperbolic tangent activation function (3) has been adopted to develop the ANN-based model; this function has been chosen to accommodate the non linearity relationships between the input (sensors' data) and the output variable (odour concentrations).The information is processed within one single hidden layer (5) which enables the neural network to model non-linear relationships between parameters and provides a better generalization than a single-layer perceptron. In mathematical notation, it is given by:

$$m_t(x) = g(h_0 + \sum_{i=1}^{p} x_i h_{it}) \tag{1}$$

where $g(.)$ is the activation function, $x_i$ is the input variable with $i = 1, \dots p$, $h_{it}$ is the weight on $i - th$ variable at node $t$ and $h_0$ is the bias value between each variable and the corresponding hidden node $(t)$. This $m_t(x)$ value is simply the output of hidden node $t$. After choosing the number of hidden nodes in the hidden layer, the outcome value can be similarly defined as a linear combination of the nodes as follows (4):

$$f(x) = v_0 + \sum_{i=1}^{p} v_t m_t, \tag{2}$$

where $v_0$ is the bias value between the hidden node $t$ and the output, $v_t$ is the output of hidden node $t$, $m_t$ is the network weights from node $t$, and $f(x)$ corresponds to the estimated outcome values. In an ANN, the parameters are updated to minimize or reduce the sum of the squared residuals. It should be noted that there is no guarantee to reach the global optimum solution (4).

DRAFT

# 3. Results and discussion

As previously mentioned, the observed data have been processed to predict odour concentration and define a possible relationship between odour concentration and olfactometric measurements, by using the ANN approach. In this context, two different layer structures have been proposed to estimate the odour concentration from the sensor array. The first layer structure is based on 10 input, 1 hidden and 1 output. The second layer structure is based on 10 input, 2 hidden and 1 output. To evaluate the quality of the trained neural network as well as to compare the two layer structures the following measures are considered: mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE). Three different options of partitioning the dataset into training, validation, and hold-out data have been used; in particular, the first option is based on 70%, 20%, 10%, and the second option on 80%, 10%, 10%, of respectively training, validation, and hold-out data. At each epoch both both the traning and the validation data are randomly selected from the original data set, and finally the hould out set is used to evaluate the performance of the ANN.

Results in Tab. 3 show that the architecture 10/1/1 (in details, 10 input variables, i.e. the sensors, 1 hidden layer and 1 output variable, the odour concentration) with data partition based on 70% for training data, 20% for validation and 10% for hold out data, is the best fitted model, with $R^2$ equals to 61.6%

which indicates that a large proportion of the variation in odour concentrations has been explained by the study sensors. In this latter case, the MSE is equal to 0.016, the RMSE 0.126 and MAE 0.019 which indicates a better fit in predicting odour concentration. Computational aspects associated to the proposed ANN architecture have been performed by using statistical software package SPSS neural networks version 25.

Table 3: ANN results by using the sensor array composed by 10 metal oxide semiconductors.

| Data partition:70%; 20%; 10% | Architecture | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|---|
| | 10/1/1 | 0.016 | 0.126 | 0.019 | 0.616 |
| | 10/2/1 | 0.061 | 0.247 | 0.027 | 0.512 |
| Data partition:80%; 10%; 10% | Architecture | MSE | RMSE | MAE | $R^2$ |
| | 10/1/1 | 0.019 | 0.138 | 0.021 | 0.612 |
| | 10/2/1 | 0.035 | 0.187 | 0.026 | 0.541 |

## 4. Conclusions

In this paper, the ANN was used to define an apt model to estimate odour concentration levels by using the responses from a multi-parametric sensors system. The obtained model was based on all the available sensors and, thanks to its accuracy level, allows the analyst to obtain reliable estimation of odour concentrations from the data recorded by the sensors. This study also showed the applicability of neural network by using a small dataset. Further developments of the research will be focused on the possible identification of a few number of sensors that most influence the odour concentration as well as to combine neural networks and tree based methods to improve the performance.

**Acknowledgments** The authors are grateful to Prof. Sandra De Iaco and Prof. Monica Palma from University of Salento, for their precious comments and suggestions.

## References

[1] Choi, Y., Kim, K., Kim, S., Kim, D.: Identification of odour emission sources in urban areas using machine learning-based classification models. Atmospheric Environment: X **13** (2022)
[2] Ghanou, Y., Bencheikh, G. Architecture optimization and training for the multilayer perceptron using ant system. Int. J. (2016).
[3] Karlik, B., Olgac, A.V. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int. J. Artif. Intell. Expert Syst., 1, 111–122 (2011).
[4] Kuhn, M., Johnson, K. Applied Predictive Modeling (Vol. 26), New York: Springer (2013).
[5] Patterson, J., Gibson, A. Deep Learning. A Practitioner's Approach, 1st ed.; O Reilly Media: Sebastopol, CA, USA (2017).
[6] Rincon, C.A., De Guardia, A., Couvert, A., Le Roux, S., Soutrel, L., Daumoin, M., Benoist J.C. Chemical and odour characterization of gas emissions released during composting of solid wastes and digestates J. Environ. Manage., 233 (2019), 39–53, 10.1016/j.jenvman.2018.12.009 (2019).
[7] World Health Organization https ://www.who.int/airpollution/ambient/about /en/ (2019).

DRAFT

# Bayesian analysis of $PM_{10}$ concentration by spatio-temporal ARIMA and STS models

Michela Frigeri[a] and Ilenia Epifani[a]

[a]Department of Mathematics, Politecnico di Milano, Milano, Italy;
michela.frigeri@polimi.it, ilenia.epifani@polimi.it

**Abstract**

Po Valley is well known to be one of the most polluted area in Italy, due to its population density, shape and climate. In this work we focus on Emilia Romagna region and analyse the panel data of its daily $PM_{10}$ concentrations collected at 49 monitoring stations in 2018. We use different Bayesian spatio-temporal models. Specifically, we model the data time series following two popular techniques: structural time series (STS) and autoregressive integrated moving-average (ARIMA) process. Then in both cases we complement the model with some geographical and topographical covariates, a trigonometric seasonal component and a latent spatial Gaussian process. Based on the posterior inference made with the Stan software, the estimates of the effects of the station features on $PM_{10}$ concentrations are robust with respect to the time trend modeling choice, but the STS strategy performs better than ARIMA process in fitting $PM_{10}$ data.

***Keywords:*** ARIMA, Bayesian inference, $PM_{10}$, spatio-temporal models, structural time series

DRAFT

## 1. Introduction and data description

*Po Valley* is a well-known hotspot for the *atmospheric particulate matter* (PM) (or fine dust) pollution in Europe (6). Many factors might be responsible: a high population density, along with a high level of urban and industrial areas are the main contributors to the high level of emissions (from vehicle exhaust, high number of industries, residential heating), while its geographic shape and climate prevent an adequate dispersion of the emitted pollutants. PM is characterized by its micrometre size: $PM_{10}$ (less than 10 micrometre in diameter) to $PM_{2.5}$ (less than 2.5). Particulate matter is carefully monitored, because of its capability to reach the respiratory tract and hence have a significant effect on human health.

Given its relevance, in the environmental statistics literature there is a huge number of studies on the evolution and the prediction of $PM_{10}$ concentrations for different areas around the world. As an example, we cite the review (8) of the statistical models (basically multiple linear regression, Bayesian autoregressive time series and discrete Markov chain model with finite state space of the pollutant concentrations) and machine learning methods (multilayer perceptron, artificial neural network) applied to $PM_{10}$ concentrations prediction in Malaysia. In addition to the methods in (8), univariate and multivariate $PM_{10}$ data were also modeled as Structural Time Series (STS) or as Auto-Regressive Integrated Moving-Average (ARIMA) processes, with or without spatial components, with or without covariates, in a frequentist or Bayesian framework; see for instance (4) for a rigorous review of the STS applied to air pollution data, (9) for a frequentist univariate ARIMA modeling without spatial component, and (5) for a Bayesian multivariate autoregressive spatio-temporal model. Finally, we refer to (7) for a general overview on Bayesian modeling of spatio-temporal data.

In this work, we analyse the multivariate time series of $PM_{10}$ collected in Emilia Romagna in 2018, via fixed monitoring stations at 49 sites scattered across the region. As this data from multiple monitoring stations is intrinsically time-dependent and spatially-correlated between stations, we developed (and compared) two alternative Bayesian spatio-temporal models: in the first model the temporal process is of STS type and in the second is an ARIMA$(p, d, q)$ process. In both cases we include some geographical and topographical covariates, a trigonometric specification of the seasonal component and a latent spatial Gaussian process with exponentiated quadratic covariance. Instead, (2) and (3) proposed a spatio-temporal Bayesian model for $PM_{10}$ concentration in the Po Valley, where the temporal process is modelled via a harmonic regression model.

The dataset was gathered by the public environmental monitoring institution ARPA Emilia Romagna. Each of the 49 stations collects the concentration of many pollutants over time, such as nitrogen dioxide $NO_2$, benzene $C_6H_6$, ammonia $NH_3$ and $PM_{10}$, $PM_{2.5}$ particles. PM measurements are collected as daily averages and are always bounded from below by a *lower limit of detection*. Although $PM_{10}$ concentrations were available starting from 2014, our analysis focused on data collected in 2018. We were motivated by the intention of creating a general model, whose predictive capabilities would not be hindered by post-Covid anomalies, or by considerable amounts of missing or unreliable data. Moreover, as environmental pollution data are known to display annual periodicity related to seasons, one year is conceptually enough to collect observations for all the relevant climatic conditions, albeit preventing further considerations on annual and interannual periodicity. The dataset also provides exogenous information on the 49 monitoring stations that cover multiple altitudinal levels, and have been chosen to represent different degrees of anthropic pollution and human activities. For each station, we have information on its position (lat–long coordinates) and altitude (in meters above sea level), its *type* which summarizes the main emission source to which the given station is exposed and which is coded as traffic, industrial or background, the *area* in which the station can be located, which expresses the level of urbanization of the surroundings (e.g. urban, suburban, rural) and its *zoning* which translates the specific geographic cores of the region and can be East plain, West plain, Agglomerate, Apennines mountain chain.

A preliminary exploratory data analysis for $PM_{10}$ concentrations shows that rural areas tend to have a different behaviour over time and a lower overall $PM_{10}$ concentration level. Additionally, nearly all stations exhibit a gradual decline in $PM_{10}$ concentrations in the spring, followed by consistently low records in the summer, before measurements date back to the onset of the colder season. This type of time pattern is well described by a loose U-shape in the time series for almost all stations, and is in agreement with previous findings available in literature. The only six stations that do not have a visible U-shaped trend, but rather an inverse pattern, all belong to the Apennine zoning, and perhaps share some peculiar atmospheric characteristics. A more detailed exploratory analysis of the entire dataset can be found in (2).

DRAFT

## 2. Bayesian models

Bayesian STS and ARIMA models developed here for the log-transformed $PM_{10}$ concentration data collected in Emilia Romagna share the following additive hierarchical structure:

$$\begin{aligned} Y_{i,t} &= f_i(t) + g_i(t) + \boldsymbol{x}_i'\boldsymbol{\beta} + w(\boldsymbol{s}_i) + \epsilon_{i,t} \\ \epsilon_{i,t} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon) \end{aligned} \tag{1}$$

where $Y_{i,t}$ is the $\ln PM_{10}$ concentration measurement at station $i = 1, \dots, 49$ and day $t = 1, \dots, 365$, $f_i(t)$ models the station-specific temporal trend function, $g_i(t)$ is the trigonometric adjustment for data seasonality, $\boldsymbol{x}_i$ provides both continuous and categorical station-specific regressors, $\{w(\boldsymbol{s}_i)\}_i$ are the spatial residuals modeled through a proper Gaussian process such that the spatial dependence is a function of the stations' coordinates $\{\boldsymbol{s}_i\}_i$, and the pure error terms $\{\epsilon_{i,t}\}_{i,t}$.

**Likelihood and covariates** As regard to the ARIMA$(p, d, q)$ specification of the trend $f_i(t)$ of $\ln PM_{10}$, after a preliminary exploratory analysis of its sample autocorrelation and partial

autocorrelation functions separately for each station $i$, we chose $p = 2, d = 1, q = 1$ for all stations, and assumed different first and second order AR parameters for each station but a common parameter for the MA term. Hence we modelled $f_i(t)$ as follows:

$$f_i(t) = Y_{i,t-1} + \phi_{i,1}\Delta Y_{i,t-1} + \phi_{i2}\Delta Y_{i,t-2} + \theta\tau_{i,t-1} + \tau_{i,t} \quad i = 1, \ldots, 49 \quad t = 1, \ldots, 364 \tag{2}$$

where $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$ and

$$\{\tau_{i,t}\}_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\tau^2) \tag{3}$$

are the error terms.
As regard to the STS choice of $f_i(t)$, following (4) we assumed an $i$-specific stochastic Markov trend $\mu_{i,t}$ given by:

$$\begin{aligned} f_i(t) &= \mu_{i,t} \\ \mu_{i,t+1} &= \mu_{i,t} + \xi_{i,t} \\ \xi_{i,t} &\overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_i^2) \end{aligned} \tag{4}$$

for $t = 1, \ldots, 365$, where $\xi_{i,t}$ is the level disturbance for the *i-th* station at day $t$; note that $\epsilon_{i,t}$ and $\xi_{i,t}$ are all assumed to be mutually independent.
In order to catch the different patterns of the $PM_{10}$ annual trend, pointed out in the data description, we included the station-specific seasonality correction

$$g_i(t) = c_i \cos\Big(\frac{2\pi}{365}t\Big) \qquad i = 1, \ldots, 49 \tag{5}$$

As far as the regression component $\boldsymbol{x}_i'\boldsymbol{\beta}$, the regressors $\boldsymbol{x}_i$ of station $i$ consist of:

- the station *altitude* expressed in meters and standardized,
- the station *zoning* that provides the part of Emilia Romagna the station is in (East plain, West plain, Agglomerate, Apennines),
- the station *type* (traffic, industrial, background),
- the station *area* (rural, urban, suburban).

Finally, for the spatial residuals $\{w(\boldsymbol{s}_i)\}_i$ in (1) we assumed a Gaussian process centered in zero with *exponentiated quadratic covariance* $\Sigma_{\sigma^2,\,\rho}$. Hence we have

$$\begin{aligned} (w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_{49})) &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\sigma^2,\,\rho}) \\ \Sigma_{\sigma^2,\,\rho}(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \sigma^2 \exp\Big\{-\frac{1}{2\rho^2}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2\Big\} \end{aligned} \tag{6}$$

Further details can be found, for instance, in (7) and references therein.

**Prior** To complete the Bayesian models, we considered suitable marginal prior distributions for each parameter included in the spatial and temporal components of the likelihood, and we assumed independence between blocks of parameters.

We used a $\mathcal{N}(0, 1)$ prior for each component of the regression coefficients' vector $\boldsymbol{\beta}$ in (4) and for the cosine regression coefficients $c_i$'s in (5). In this context the standard Gaussian prior turns out to be sufficiently vague, since $\ln PM_{10}$ ranges between 0 and 5 and all the covariates have been standardized. To avoid infinite variances, the variances $\sigma_\epsilon^2$ of the pure errors $\epsilon_{i,t}$'s and $\sigma_w^2$ of $w(\boldsymbol{s}_i)$'s are all provided with $IG(3, 2)$ prior density. Finally, following (7), we fixed the spatial range parameter $\rho$ equal to three times the inverse of the maximum distance (in km) between the available monitoring stations, obtaining $\rho = 0.05$ in (6); this choice is made to avoid well–known issues that prevent spatial models to be completely identifiable.

DRAFT

As regard the prior for $f_i(t)$, in order to properly share information across stations, in case of the ARIMA(2,1,1) model in (2)-(3), the prior scheme is

$$\begin{aligned}
\phi_{i,j}|\mu_{\phi_j}, \sigma^2_{\phi_j} &\sim \mathcal{N}(\mu_{\phi_j}, \sigma^2_{\phi_j}) && \text{for } j = 1, 2 \\
\mu_{\phi_j} &\sim \mathcal{N}(0, 5) && \text{for } j = 1, 2 \\
\sigma^2_{\phi_j} &\sim IG(2.1, 1.1) && \text{for } j = 1, 2 \\
\theta &\sim \mathcal{N}(0, 1), \\
\sigma^2_\tau &\sim \text{Half-Cauchy}(0, 5)
\end{aligned}$$

Instead, in case of the STS model in (4), each level disturbance $\xi_{i,t}$ is modeled as $\xi_{i,t} = \sigma_i \gamma_t$ with

$$\begin{aligned}
\gamma_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \quad t = 1, \ldots, 365 \\
\sigma_i &\overset{\text{iid}}{\sim} \mathcal{U}_{(0,2)} \quad i = 1, \ldots, 49
\end{aligned}$$

As already discussed for the $\beta$ priors, also in this case the chosen hyperparameters provide sufficiently vague distributions, with respect to the observed values of the response variable.

## 3. Model selection, estimation and prediction

We used Stan (1) for computing the posterior distribution of all parameters involved in the two models. In both cases, we run two parallel MCMC chains, assuming 1,000 burn-in iterations and 1,000 sampling iterations. We compared the two models using Widely Applicable Information Criterion (WAIC) and Leave One Out Cross Validation (LOO-CV). In addition we also computed the coverage (CVG) as the percentage of true data falling inside the 95% posterior credible intervals; we expect this value to be close to 0.95 in well specified models. From Table 1 we see that the lowest values of both predictive model selection criteria are achieved by the STS model, which seems to provide better results also in terms of coverage. However, the posterior estimates of the effects of the station features on $PM_{10}$ concentrations are robust with respect to the competing time-trend models ARIMA and STS. In both models we can evaluate the presence of spatial correlation, supporting our choice to include a residual spatial term. Furthermore, analyzing the posterior 95% credible intervals of the regressors, only the altitude and traffic covariates appear to affect the concentration level of $PM_{10}$. In particular, as might be expected, higher altitudes are associated with a lower $PM_{10}$ concentration, while the proximity to traffic areas increases the level of pollution. Finally, the effects of the other geographical and topographical factors are completely obscured by the relevance of the random spatial process $\{w(\boldsymbol{s}_i)\}_i$. A further meaningful finding concerns the posterior predictions of $\ln PM_{10}$ under the ARIMA(2,1,1) time-trend (in Figure 1 in blue) and under the STS model (in Figure 2 in blue), both compared with the real observed $\ln PM_{10}$ (in pink).

Table 1: Predictive model selection criteria and coverage percentage for ARIMA and STS models.

| | ARIMA | STS |
|---|---|---|
| WAIC | 16313.1 | 10846.8 |
| LOO-CV | 16320.7 | 10849.7 |
| CVG | 0.9482 | 0.9514 |

Due to space constraints, we only show the case of *Ceno* monitoring station. We see that the posterior STS estimated trend follows the data very closely, with also a lower uncertainty than the estimates under the ARIMA model. However we still need to overcome a poor mixing performance for the STS model, that is due to an identification problem involving $f_i(t)$, $g_i(t)$ and caused by the specification of multiple station-specific parameters.

DRAFT

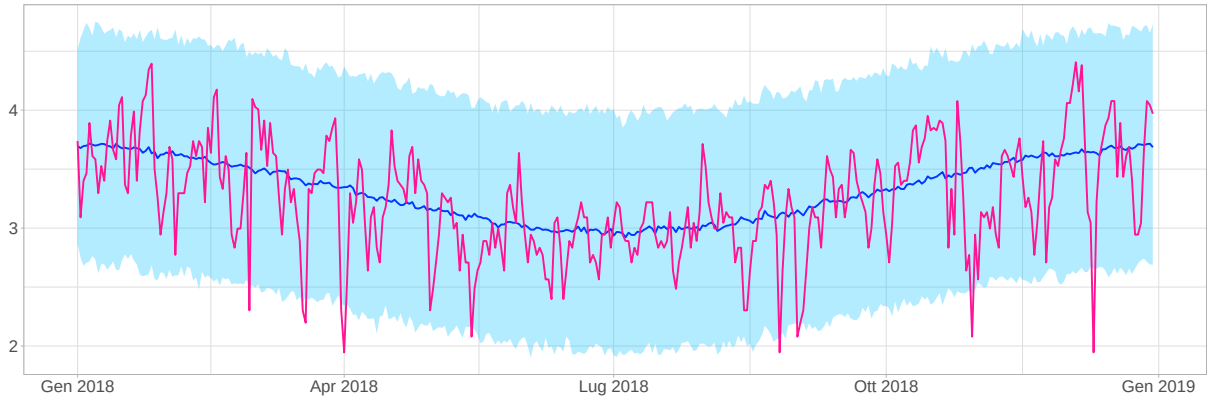



Figure 1: 95% credible intervals (light blue) and median (blue) of the posterior predictive distribution provided by ARIMA$(2, 1, 1)$ model for *Ceno* monitoring station. True recorded values are displayed in pink.

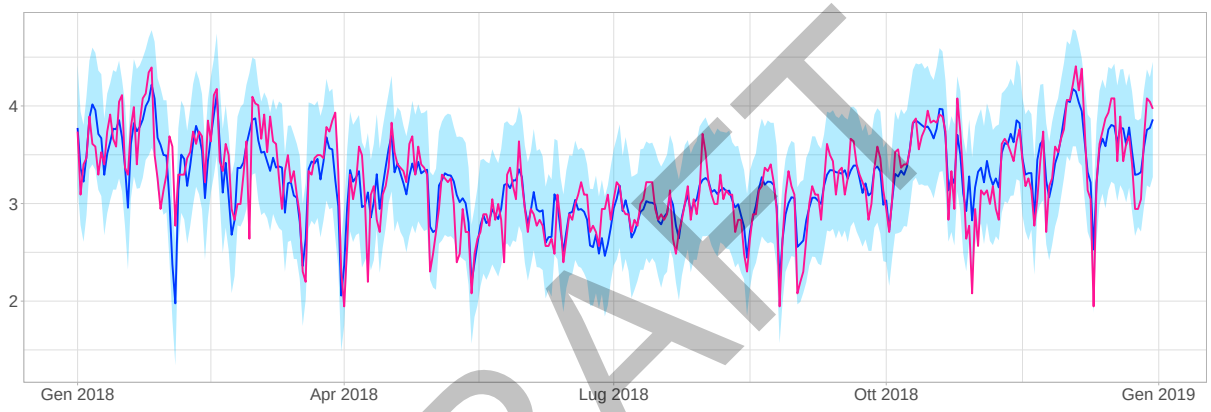



Figure 2: 95% credible intervals (light blue) and median (blue) of the posterior predictive distribution provided by STS model for *Ceno* monitoring station. True recorded values are displayed in pink.

DRAFT

**Acknowledgments** We would like to thank the students of the MSc course of Bayesian Statistics (MSc in Mathematical Engineering of Politecnico di Milano): Arrigoni F., Baracchi F., Cantalini C., Gjyli E., Ferrara S. and Ursino B. for the support in writing some code for posterior simulations.

## References

[1] Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017) doi: 10.18637/jss.v076.i01

[2] Frigeri, M.: Spatio-Temporal Models for Particulate Matter in the Po Valley. Thesis, M.Sc. in Mathematical Engineering, Politecnico di Milano (2022). http://hdl.handle.net/10589/195435

[3] Gianella, M., Guglielmi, A., Lonati, G.: A Bayesian spatio-temporal model of PM10 pollutant in the Po Valley. Book of Short Papers - SIS 2022, pp. 883–888. Pearson (2022) doi: 10.1109/ICEET.2009.468

[4] Lawson, A.R., Ghosh, B., Broderick, B.: Prediction of traffic-related nitrogen oxides concentrations using Structural Time-Series models. Atmos. Environ. **45**(27), 4719–4727 (2011) doi: 10.1016/j.atmosenv.2011.04.053

[5] Manga, E., Awang, N.: Bayesian autoregressive spatiotemporal model of PM10 concentrations across Peninsular Malaysia. Stoch. Env. Res. Risk Assess. **32** (2018) doi: 10.1007/s00477-018-1574-5
[6] Masiol, M., Squizzato, S., Formenton, G., Harrison, R.M., Agostinelli, C.: Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the Veneto region, NE Italy. Sci. Total Environ. **576**, 210–224 (2017) doi: 10.1016/j.scitotenv.2016.10.042
[7] Sahu, S.K.: Bayesian modeling of spatio-temporal data with R. Chapman and Hall/CRC (2022) https://doi.org/10.1201/9780429318443
[8] Shaziayani, W.N. e.a.: A Review of PM10 Concentrations Modelling in Malaysia. Proceedings of 2nd International Conference on Green Environmental Engineering and Technology, 2020, **616**. IOP Publishing (2020) doi: 10.1088/1755-1315/616/1/012008
[9] Wang, W., Guo, Y.: Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model. Proceedings of International Conference on Energy and Environment Technology, Guilin, China, 2009, pp. 7–10 (2009) doi: 10.1109/ICEET.2009.468

DRAFT

# Functional ANOVA to monitor yearly Adriatic sea temperature variations

Annalina Sarra[a], Adelia Evangelista[a], Tonio Di Battista[a], and Nicola Di Deo[b]

[a]University of Chieti-Pescara, Viale Pindaro, 42, 65127 Pescara, Italy;
annalina.sarra@unich.it,
adelia.evangelista@unich.it,tonio.dibattista@unich.it
[b]Regional Agency for the Environmental Protection of Abruzzo, V.le G. Marconi, 51, 65127 Pescara, Italy; n.dideo@artaabruzzo.it

**Abstract**

Temperature rises in marine habitats have been shown to have a variety of effects on biodiversity, such as algal blooms, altering biotic factors and disrupting life cycles. The aim of this study is to propose a Functional Analysis of Variance (FANOVA) to analyse the yearly sea temperature variations in the coastal zone of Abruzzo Region (Central Italy). Within this aim different FANOVA tests have been used to assess functional mean differences. FDA applied to marine temperature data found no evidence to reject similarity between temperature samples acquired at different depths, confirming the absence of temperature anomalies across the time period analysed.

***Keywords:*** Temperature sea rising, Adriatic sea, FDA, Functional Analysis of Variance, Functional mean differences

DRAFT

## 1. Introduction

Over the last years, there has been an increasing interest in acquiring a broad knowledge about temperature behaviour in the Mediterranean sea which is crucial for ensuring a sustainable development and conservation of its marine environments. It is well documented that temperature changes affect marine communities in a variety of ways, both directly and indirectly (9). On one hand, the variations in sea temperature directly impact the physiology, growth, reproduction, recruitment and behavior of poikilothermic organisms, like fishes. On the other hand, sea temperature rising has also indirect effects that may be mediated by biota interactions or by marine currents. For instance, the recent bleaching and subsequent massive mortality of corals in all marine environments is the visible response of the marine biota to prolonged anomalous increase in sea temperature. According to a new study conducted by international scholars, the Mediterranean experienced five consecutive years of marine heatwaves and mass mortality events between 2015 and 2019 (3). Moving along these lines of research, this paper analyses sea temperature series measured at seven geographical locations along the coast of Abruzzo (Central Italy). Specifically, we are aimed to investigate if there are yearly anomalies in the Adriatic sea temperature data. We ask the question: "does the shape of the overall temperature profile over the studied period depend on which year we focus on? Is there any footprints of temperature rise in the sea water off Abruzzo?". In analysing the dynamics of sea temperature series, provided by the Regional Agency for the Environmental Protection of Abruzzo (ARTA), and adressing the question, we followed a Functional

Analysis Data approach.
Functional Data Analysis (7) is a branch of statistics working with observations that come from a continuous process that is evaluated at discrete points. Accordingly, FDA tools have been developed to analyse intrinsically infinite-dimensional data, usually measured discretely. Two ideas make FDA unique: it takes into account the domain correlation structure of the data and leads to a global view of the problem through curve analysis instead of individual observations, enlarging the possibilities of research. Over the last years, FDA has expanded to a wider number of scientific fields, including, among others, the environment (1), health and medical research (2), industrial processes (6), econometrics (5), all involving continuous time monitoring process. In our study, the FDA enables the complete analysis of depth spectrum, along which the sea temperatures were recorded. To be specific, we contrast the similarities between sea temperature samples obtained at different levels of depth over the studied period (2011-2020) by means of Functional Analysis of Variance (FANOVA). The remaining part of this paper is organized as follows. Section 2. describes the study region and the data. Section 3. presents the methodology applied in the statistical analysis. Section 4. provides with the results of FANOVA and some concluding remarks.

## 2. Region of study and data

In our work,we focus our attention on the Adriatic Sea, that is the articulation of the Mediterranean Sea that separates the Apennines Peninsula from the Balkan Peninsula and the Apennines from the Dinaric Alps. The Adriatic Sea has a land area of 138,595 km$^2$, a length of 738 km, an average width of 159.3 km, and a depth of 173 m. Our study area is the Abruzzo coast, which is enclosed between the Tronto and Trigno rivers and has a variable morphology due to the geological structure of the immediate hinterland. We exploited data collected by the Regional Agency for the Environmental Protection (ARTA), that over the last years, according to the agreement with the Abruzzo Region, carried out the monitoring activities of the marine-coastal environment. The monitoring network of Abruzzo's marine-coastal waters consists of fourteen stations identified for sampling different environmental matrices, distributed along seven transects perpendicular to the coast and placed at 500 and 3000 m (Fig.1). The data set used for the statistical analysis consists of eleven years (2011-2021) of monthly measurements of temperature acquired at different depth (up to 11 m) in the water columns, sampled in the stations that were at 3000 m offshore.

DRAFT

## 3. Functional Anova (FANOVA)

Before applying functional methods to determine whether sea temperatures differ significantly in various years, we need to transform the discrete sea data into smooth functions $y(t)$, measured at $t = 1, \ldots, T$. In functional data analysis, t typically represents a real-valued time variable, but in the current problem it denotes depth at which sea temperature was recorded; $y(t)$ is the smoothed sea temperature value. The smoothed functions $y(t)$ are obtained as a linear combination of independent basis functions $\phi_k$, with coefficients, $c_k$, as:

$$y(t) = \sum_{k=1}^{K} c_k \phi_k(t). \tag{1}$$

The basis function can be chosen from a variety of options. In our case, cubic $\beta$-splines (piecewise polynomial of degree 3) is the most advisable choice as it combines fast computation capabilities and flexibility (10). To perform FANOVA, let us assume to have $M$ independent samples $Y_{gj}(t), j = 1, \ldots, ng, \quad t \in [a, b]$, extracted from $\mathcal{L}^2(l)$ processes $Y_g(t), g = 1, \ldots, M$, and their mean function is $E(Y_g(t)) = m_g(t)$ (11). FANOVA, like vector analysis, compares the distance between the mean levels of the factor variables. The goal of this comparison is to determine whether the set of functions under consideration is statistically distinguishable. When the functional sample is classified in several groups,

Figure 1: Localization of the sampling stations of the Regional Network.

DRAFT

such as $\{\mathcal{Y}_j\mathcal{G}_j\}_{i=1}^n$, $\in \mathcal{F} \times \mathcal{G} = 1, \ldots, G$, where $G$ is a discrete variable indicating the membership group, the null and alternative hypotheses can be specified as follows:

$$\begin{cases} H_0 : \bar{Y}_1(t) = \bar{Y}_2(t) = \ldots \bar{Y}_G(t) \\ H_1 : \exists \quad h, e \quad s.t. \bar{Y}_h(t) \neq \bar{Y}_e(t). \end{cases}$$

The model for $j-th$ observation in the $g-th$ group can be expressed as:

$$Y_{jg}(t) = \mu_t + \alpha_g(t) + \epsilon_{jg}(t) \tag{2}$$

where $Y_{jg}(t)$ is the functional value of temperature in the group $g$, $\alpha_g(t)$ is the effect of being part of a determined group and $\epsilon_{jg}(t)$ represents the unexplained variability for $j-th$ observation of group $g$. Adopting the matrix notation, the model in Equation 2 can be expressed as:

$$\boldsymbol{Y}(t) = \boldsymbol{Z}\gamma(t) + \boldsymbol{\epsilon}(t) \tag{3}$$

where $\boldsymbol{Y}(t)$ is a $N$-dimensional vector, $\gamma(t) = (\mu(t), \alpha_1(t), \ldots \alpha_M(t))^T$ a $(M+1)$ dimensional vector, $\boldsymbol{\epsilon}(t)$ a vector of $N$ residual functions and $\boldsymbol{Z}$ the design matrix with dimension $(N, M+1)$. It is worth noting that each row of the matrix $\boldsymbol{Z}$ corresponds to a single observation; the first column consists entirely of 1 to represent the overall mean and $M$ columns correspond to different groups, with value 1 if the observation belongs to the $g-th$ group and 0 otherwise. In order to assure the identifiability of the functional effects $\alpha_g(t)$, the sum to zero is introduced:

$$\sum_{g=2}^{M+1} \gamma(t) = 0 \quad \forall(t). \tag{4}$$

The parameter vector $\gamma(t)$ in Equation 3 can be estimated minimizing the standard least squares

$$LMSSE(\gamma) = \int [\boldsymbol{Y}(t) - \boldsymbol{Z}\gamma(t)]^T [\boldsymbol{Y}(t) - \boldsymbol{Z}\gamma(t)] d(t) \tag{5}$$

To compare the similarity of samples, a number of test statistics were proposed. For example, similar to the regular analysis of variance, one can compare between and within group variations:

$$Fn_t = \frac{SSR_n(t)/(M-1)}{SSE_n(t)/(n-Q)}$$

where

$$SSR_n(t) = \sum_{g=1}^{M} n_g(\bar{Y}_g(t) - \bar{Y}(t))^2$$

and

$$SSE_n(t) = \sum_{g=1}^{M}\sum_{j=1}^{n} g(\bar{Y}_g j(t) - \bar{Y}_g(t))^2$$

represents the variations between groups and within groups, respectively. As in classical ANOVA, a large $F_n(t)$ value indicates that the variance described by the model is greater than the variance not explained. The main difference between this method and traditional ANOVA (both univariate and multivariate) is that the value of $F_n(t)$ varies across the entire domain instead of being fixed. Note that, the classical significance level was designed to be used for a single hypothesis rather than a continuum of hypotheses. To overcome the possibility to have falsely claiming around the interval, a viable solution is to use the permutation test (7), which is the functional equivalent of the univariate F-test statistic. Additionally, in this work, we consider a permutation test, based on a representation of the base function, as described in (4), and Zhang and Liang's procedure to implement a global test, obtained via globalising the usual pointwise F-test (12).

## 4. Results and concluding remarks

A separate one-way FANOVA is run for each marine monitoring station to test the null hypothesis that there is no significant difference between the yearly mean group functions. All analyses were carried out in the R environment (8), using the R packages *fda* and *fdANOVA*. Table 1 summarizes the results of the permutation test based on basis representation (4) and the global point-wise F test (12). A permutation

Table 1: Results of FANOVA TESTS

| | FP[a] | | GPF[b] | |
|---|---|---|---|---|
| Monitoring stations | Test | p-value | Test | p-value |
| Alba Adriatica | 0.385 | 0.960 | 0.387 | 0.963 |
| Giulianova | 0.514 | 0.866 | 0.515 | 0.889 |
| Pineto | 0.494 | 0.894 | 0.492 | 0.914 |
| Pescara | 0.407 | 0.949 | 0.405 | 0.955 |
| Ortona | 0.433 | 0.928 | 0.434 | 0.942 |
| Vasto | 0.255 | 0.986 | 0.257 | 0.992 |
| San Salvo | 0.337 | 0.959 | 0.339 | 0.977 |

**Acronyms**: [a]FP-permutation test based on basis function representation (FP); [b]GPF-global point-wise F-test.

method is used in both procedures to approximate the null distributions. Furthermore, cubic $\beta$-splines and the Bayesian Information Criterion (BIC) are considered for both tests. The statistical analysis performed using the FDA approach finds no significant differences between yearly sea thermal variations,

DRAFT

corroborating the hypothesis of not rejecting the similarity in all years for sea temperature data. We also obtained the same evidence using Ramsay and Silverman's (7) functional F statistic (Fig.2 and 3). FDA proved to be an effective tool for easily monitoring sea temperature changes across the entire depth spectrum analysed, contrasting the results inferred from a simplistic visual inspection of the data.

## References


[1] Acal, C., Aguilera, A.M., Sarra, A. *et al*: Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic. Stoch. Environ. Res. Risk. Assess. **36**, 1083–1101 (2022) doi: 10.1007/s00477-021-02071-4

[2] Dombeck, D., Graziano, M., Tank, D.: Functional clustering of neurons in motor cortex determined by cellular resolution imaging in awake behaving mice. J. Neurosis. **29**, 13751–13760 (2009)

[3] Garrabou, J. *et al*: Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea. Glob. Chang. Biol. (2022) doi: 10.1111/gcb.16301

[4] Gorecki, T., Smaga, L.: Comparison of tests for the one-way anova problem for functional data. Comput. Stat. **30**(4),987–1010 (2015) doi: 10.1007/s00180-015-0555-0

[5] Müller, H.G., Sen, R., Stadtmüller, U.: Functional data analysis for volatility. J. Econometr. **65**, 233–245 (2011)

[6] Ordónez, C., Martínez, J., Saavedra, A., Mourelle, A.: Intercomparison Exercise for Gases Emitted by a Cement Industry in Spain: A Functional Data Approach. J. Air Waste Manag. Assoc. **61**, 135–141 (2011)

[7] Ramsay, J.O., Silverman, B.W.: Functional data analysis, 2nd edn. Springer-Verlag, New York (2005)

[8] R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria (2023) www.R-project.org/

[9] Southward, A.J., Hawkins, S.J., Burrows, M.T.: Seventy years' observations of changes in distribution and abundance of zooplankton and intertidal organisms in the western English Channel in relation to rising sea temperature. J. Therm. Biol. **20** 127–155 (1995)

[10] Wegman, J.E., Wright, W.I.: Splines in statistics. J. Am. Stat. Assoc. **78**, 351–365 (1983) doi: 10.1080/01621459.1983.10477977

[11] Zhang, J.-T.: Analysis of Variance for Functional Data. In: A Chapman & Hall Book (eds), Crc Monographs on Statistics & Applied Probability, p. 412. Taylor & Francis Group, Abingdon, UK (2013)

[12] Zhang, J. T., Liang, X.: One-way ANOVA for functional data via globalizing the pointwise F-test. Scand. J. Stat. **41**, 51–71 (2014) doi: 10.1111/sjos.12025


DRAFT

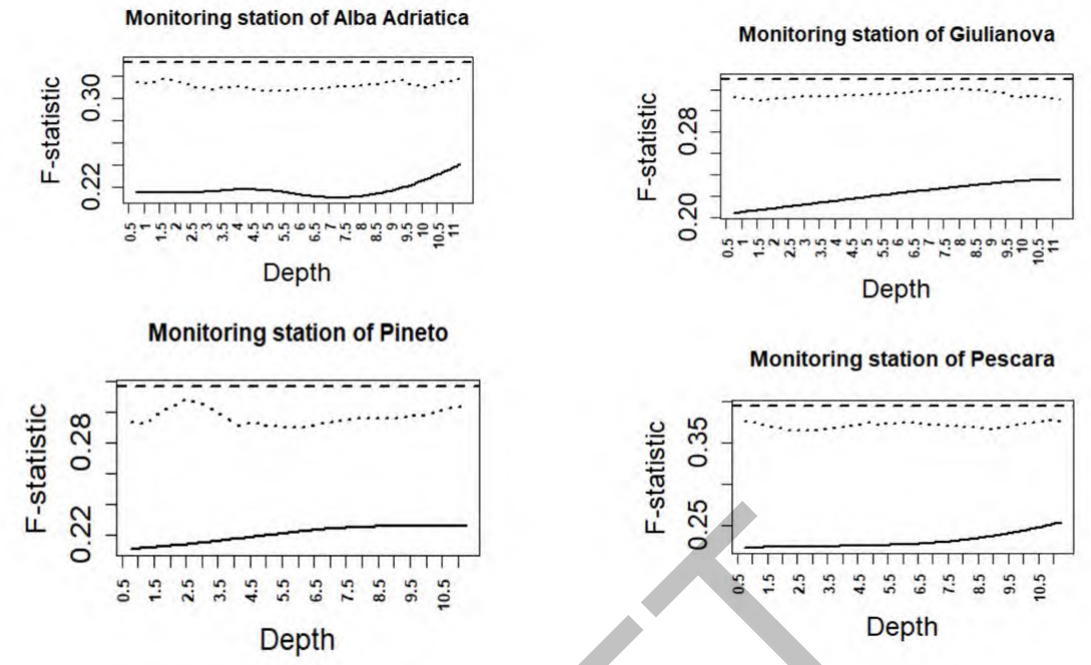



Figure 2: Functional F statistics of the monitoring stations of Alba Adriatica, Giulianova, Pineto and Pescara.

DRAFT

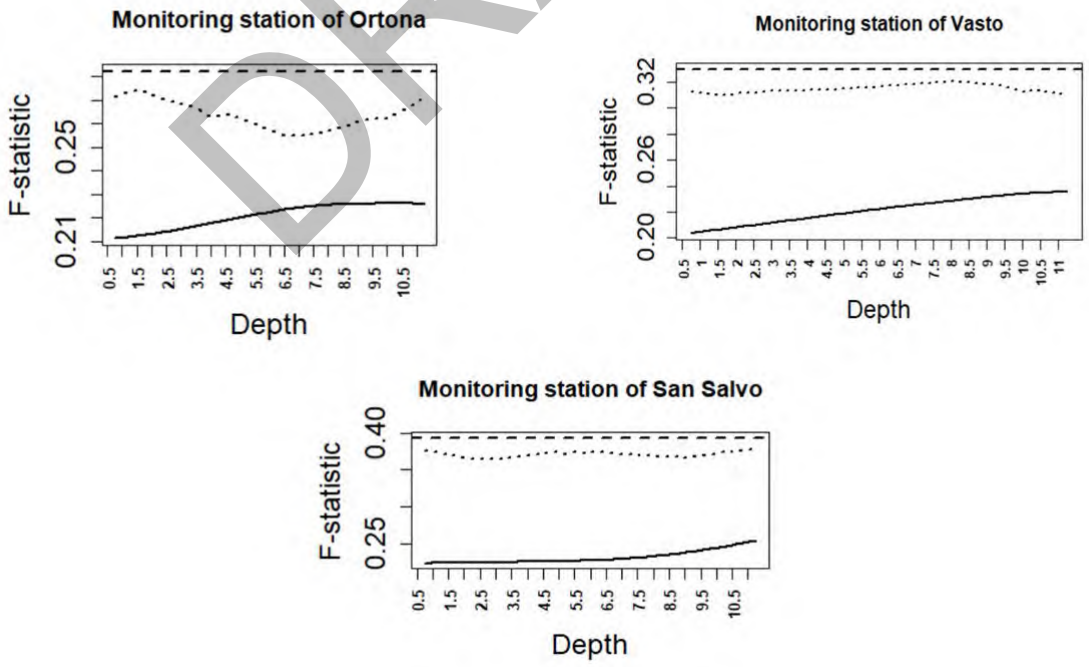



Figure 3: Functional F statistics of the monitoring stations of Ortona, Vasto and San Salvo.

# New perspectives in the measurement of biodiversity

Linda Altieri[a], Daniela Cocchi[a], and Massimo Ventrucci[a]

[a]Department of Statistical Sciences, University of Bologna; linda.altieri@unibo.it, daniela.cocchi@unibo.it, massimo.ventrucci@unibo.it

**Abstract**

Entropy is widely used in biodiversity studies, where data often present complex interactions. Difficulties arise in linking entropy to available covariates or data dependence structures, as all existing entropy estimators assume independence. We take a Bayesian model-based approach and focus on estimating the probabilities which compose an entropy index, accounting for data dependence. This way, the entropy estimate is not a single value, rather it becomes a curve or a two dimensional surface according to the data structure. We obtain an interpretable index of the latent biodiversity of a system.

***Keywords:*** Shannon's entropy, entropy estimation, diversity indices, CAR models

DRAFT

## 1. Introduction

Biological diversity, or biodiversity, is the variety of life forms and is commonly used to replace long-established terms such as species diversity and species richness, both indicating the number of species represented in an ecological community. If a lot of individuals are present but they all belong to the same species, there is no biodiversity; if there are few individuals, but they belong to different species, biodiversity is high, and related indices should reflect this. Both environmental variability and continuous colonization and extinction of communities create natural variability of biodiversity in space and time. Studies on biological diversity are a fundamental part of every ecological survey, as biological diversity plays a crucial role in the delivery of a range of ecosystem services such as natural harvests, carbon sequestration, pollination and soil formation (3). At the same time, biodiversity is threatened by climate changes, over-harvesting, pollution, habitat loss and invasive species, all of which can be attributed, either directly or indirectly, to human activities. In 2002, the UN Convention on Biological Diversity set a target "to significantly reduce the rate of biodiversity loss" (http://www.cbd.int/2010-target/); widespread concern about habitat and species loss led the United Nations to declare 2010 as its International Year of Biodiversity. This has stimulated the development of many different research studies, which have produced a large set of diversity indices. One of the most popular ones is Shannon's entropy (10), a successful measure in many fields, able to synthesize several concepts in a single number: entropy, information, heterogeneity, surprise, contagion. The flexibility of such index and its ability to describe any kind of data, including categorical variables, motivate its diffusion across applied fields such as geography, ecology, biology and landscape studies. In particular, the traditional formula of Shannon's entropy has been widely used in ecological studies to measure the biodiversity of a system. In descriptive studies entropy and biodiversity are the same thing, while in the field of estimation entropy is not a synonym of the *observed* biodiversity, rather it describes the *latent* biodiversity of the system. When entropy is low, the two things coincide: one (or few) of the species probabilities is predominant, i.e. no matter how many individuals are present, they tend to belong to one (or few) species. Conversely,

when entropy is high, species probabilities are more evenly distributed, therefore individuals in the area may belong to any of the considered species. When we have very few observations, like 1 o 0, it means that for that specific dataset and in that specific moment the *observed* biodiversity is low, but, *potentially*, new individuals may belong to any species: for example, if animals move from a neighboring site, or if a dependence between species and an environmental covariate is established. The idea should be not just to observe data, but to capture what may happen according to the underlying process behavior.

Typically, the amount of data for different species depends on environmental covariates, spatial location, temporal structures. The main drawback of Shannon's entropy in its traditional form is that it cannot account for such auxiliary information or data dependence. The inclusion of spatial information has been the target of intensive research (2; 5; 1), but there is no study on entropy measures for data with spatial correlation that takes the issue of estimation into account. When the goal is to estimate the biodiversity of a system based on data, the standard approach relies on the Maximum Likelihood (ML) estimator, which substitutes the unknown probability distribution of interest with relative frequencies and performs well when independence is an acceptable assumption. The most popular proposals consist of corrections of the ML estimator bias, e.g., the Miller-Madow correction, the Bayesian Nemenman, Shafee and Bialek estimator (7), or machine learning methods. Two main limits concern entropy estimation. Firstly, the literature only focuses on correcting or improving the performance of the ML estimator regarding the estimator bias. Secondly, independence among realizations is always assumed and no auxiliary information is considered. To our knowledge, no study faces the task of estimating entropy for data presenting dependence on available covariates, spatial/temporal association or other types of dependence. Recently, scientists have shown interest in how data structures are related to entropy; for example, ecological studies attempt at finding a relationship between quantities, such as species abundance and richness, and environmental factors or temporal/spatial effects, with a regression or mixed model approach. Such methods fail to detect any relevant association: they cannot account for absent species and for the possibility that species are not the same across observation sites, moreover, abundance, richness and entropy all depend on the occurrence probabilities of each species. Research would benefit from a focus between probabilities themselves and covariates or effects. Failing to consider data in neighbouring areas (or time points) produces a set of local entropy values, which is usually non-informative and difficult to interpret. Gelfand (4) reviewed recent literature on the distribution of species, and states "the ecology world recognized the need to incorporate spatial dependence in describing presence or abundance at a site. Joint species distribution models are replacing marginal models with a rapidly growing literature, but very little of it is spatial, despite the evident dependence within a site as well as the anticipated dependence across sites. In fact, this path offers an opportunity for future research". The mentioned issues are addressed by our proposal in the present paper.

The main aim of this paper is to present a new, model-based approach for entropy estimation. We take a new perspective that moves the focus from the index itself to its components. If the probability mass function (pmf) of the variable of interest can be properly estimated and any dependence on covariates or spatial/temporal factors can be assessed, such information can be used to enrich an entropy estimator. In the case of categorical variables, where probabilities can be described by a multinomial distribution, the crucial point is to estimate its parameters. A Bayesian model-based approach allows to derive such distribution and can be extended to account for dependence on covariates and/or correlation across realizations. It also exploits the information about the abundance of each category (species) and avoids reducing data to presence/absence with related issues raised by Gelfand (4). After obtaining a posterior distribution for all parameters, the posterior distribution of entropy is straightforward, since the entropy formula only depends on the probabilities. The entropy estimator can be, e.g., the distribution mean; credibility intervals and other syntheses may be obtained via standard tools of Bayesian inference. This approach can be used for any setting, including spatial and/or temporal data, with or without auxiliary information and independence assumptions. In the spatial context, coherently with standard procedures for variables linked to areal and point data, the estimation output is a spatial surface for the entropy over the area under study; for time series, the estimate can be represented by a curve. Curves and surfaces derived including dependence structures take into account the (spatial or temporal) neigbourhood of each observation point, smooth out any random variation in the data and allow to grasp the general behaviour of entropy. When wished, results can be further synthesized in a global entropy value

DRAFT

for the whole dataset, that is well-grounded as it is built out of consideration of data structures and not only relative frequencies. We underline the ability of our estimation approach to draw conclusion on the data at hand and on the relationship between entropy and any covariate or underlying structure.

When biodiversity studies involve an observation area, spatial data may be areal, geostatistical or point process data, and our approach works for any of these types. With point process data, the most common option is to discretize space with a very fine grid. The approximated model is known to converge to the true process (6) and results are very accurate in most applications. We show a case study with the gridding approach, which is usually a satisfying trade-off between complexity and accuracy of the results. A practical advantage of the gridding approach is that it may overcome the issues of spatial misalignment with the resolution of the available covariates. An extension to models in continuous space is possible in our approach, by exploiting the Stochastic Partial Differential Equation (SPDE) approach, available with INLA (Integrated Nested Laplace Approximation, (8)).

## 2. Model-based entropy estimation

Let $X$ be a categorical variable of interest with $I$ categories (species). Shannon's entropy of $X$ is:

$$H(X) = \sum_{i=1}^{I} p(x_i) \log \frac{1}{p(x_i)}, \tag{1}$$

where $p(x_i)$ is the probability of occurrence of the $i$-th category, while $\boldsymbol{p}_X = (p(x_1), \ldots, p(x_I))'$ is the probability mass function of $X$. Entropy ranges in $[0, \log I]$, where 0 is obtained when $p(x_{i^*}) = 1$ for some category $i^*$ and $p(x_i) = 0$ for $i \neq i^*$, and $\log I$ derives from $p(x_i) = 1/I$ for all $i$. Under the perspective of estimation, a stochastic process is assumed to generate the data according to an unknown probability function and, consequently, an unknown entropy. One realization of the process is observed and employed to estimate such entropy. We take a model-based approach for the estimation of the main components of an entropy index, i.e. $p(x_i)$ for $i = 1, \ldots, I$.

Conditional auto regressive (CAR) models (8) provide a way of including spatial or temporal correlation, by explaining a response via temporally or spatially structured random effects, and are usually defined for binary variables. When the number of species $I$ is greater than 2, an extension to the multinomial logit model arises naturally and modes becomes more complicated. For each observation point $u = 1, \ldots, n$ we have $n_{u1}, \ldots, n_{uI}$, the number of individuals of each species at site $u$, that may be equal or greater than 0, and we use them as a starting point to estimate $p_{u1}, \ldots, p_{uI}$, with $\sum_{i=1}^{I} p_{ui} = 1$; to ensure that all probabilities are proper, we model them as $p_{ui} = \exp\{g_{ui}\} / \sum_{i=1}^{I} \exp\{g_{ui}\}$, with:

$$g_{ui} = \boldsymbol{z}_{ui}^{'} \boldsymbol{\beta}_i + \phi_u \tag{2}$$

where $\boldsymbol{z}_{ui}^{'} \boldsymbol{\beta}_i$ is the vector of fixed effects associated to the $u$-th location and $i$-th species. The vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)' \sim MVN_n(\mathbf{0}, \boldsymbol{\Sigma})$ is a spatial effect with a structured covariance matrix $\boldsymbol{\Sigma} = [\tau(\boldsymbol{D} - \rho \boldsymbol{A})]^{-1}$, which depends on a precision parameter $\tau$ and a dependence parameter $\rho \in [-1, 1]$ quantifying the strength and type of the correlation between neighbouring units. The symbol $\boldsymbol{A}$ denotes the adjacency matrix reflecting the neighbourhood structure, and $\boldsymbol{D}$ is a diagonal matrix, where each element contains the row sums of $\boldsymbol{A}$. By tuning the structure of $\boldsymbol{A}$, different types of dependence in one or two dimensions may be included in the model. Fitting a multinomial model may be very complicated in practice: the presence of the denominator in (2) makes the likelihood problematic. We propose to exploit the Multinomial-Poisson transform, which turns the multinomial likelihood into a Poisson likelihood with extra parameters. It is established that the likelihood kernel of the transform is proportional to the multinomial likelihood kernel, so that the transfom returns the same estimates and asymptotic variances as the original distribution. Computational details on how to implement the multinomial-Poisson transform are given in (9). To model temporal random effects, we apply popular types of Intrinsic Gaussian Markov Random Field (IGMRF), namely the first and second-order Random Walk models RW1 and RW2 (8). To include spatial correlation, the simplest way of representing a neighbourhood system

DRAFT

is via $\boldsymbol{A} = \{a_{uu'}\}_{u,u'=1,\ldots,n}$, a square $n \times n$ matrix such that $a_{uu'} = 1$ when unit $u$ and unit $u'$ are neighbours, and $a_{uu'} = 0$ otherwise. The standard Intrinsic CAR model has been modified to include 12 nearest neighbours for grid data, i.e. two consequent pixels along each cardinal direction plus the four ones along the diagonals, and has recently become of standard use thanks to its implementation in INLA under the name of RW2d effect (8).

Once the probabilities are estimated for each category and observation, an estimate for the entropy is obtained in a deterministic way, using formula (1). Entropy is a function of the probabilities that returns values in $[0, \log I]$; therefore, the posterior of the estimated entropy $\widehat{H}$ is 1-dimensional. Usually, results are delivered using a point estimator, which in our case is:

$$\widehat{H}_u^{BMB}(X) = \sum_{i=1}^{I} \widehat{p}_{ui}^{BMB} \log \frac{1}{\widehat{p}_{ui}^{BMB}} \tag{3}$$

where BMB stands for Bayesian Model Based estimator, and $\widehat{p}_{ui}^{BMB}$ is the posterior mean derived by simulating from the posterior distribution; a credibility interval at any level may be obtained and associated to the estimator value to give an idea of the estimate uncertainty. By taking this approach to estimation, the entropy estimate $\widehat{H}_u^{BMB}(X)$ varies with $u$ when data present some type of dependence structure. By plotting $\widehat{H}_u^{BMB}(X)$, we obtain a curve for continuous covariates or temporal dependence, or a two-dimensional surface for georeferenced data.

Two main improvements must be highlighted with respect to a local (observation-specific) computation of any literature estimator. Firstly, with our approach, estimates can be computed even when very few observations of $X$ (even 1 or none) are available for each observation point/site. Secondly, when random effects are included, estimates take the neighbourhood into account (according to the specific dependence structure) and we obtain smooth curves/surfaces that catch the main behaviour of the data under study and remove the noise. Even when the model only includes covariates, the whole dataset is exploited to estimate the fixed effect coefficients and returns more reliable results. With this method, the resulting entropy curve/surface exploits the maximum amount of information for producing results. Moreover, by looking at the chosen model and at the values for the estimated parameters, we can infer the relationship between entropy and, say, environmental covariates, spatial coordinates and/or time effects.

DRAFT

## 3. A biodiversity case study

Our motivating dataset documents the presence of tree species over Barro Colorado Island, Panama. The considered dataset is a marked point pattern with $n = 5639$ trees over a rectangular window of

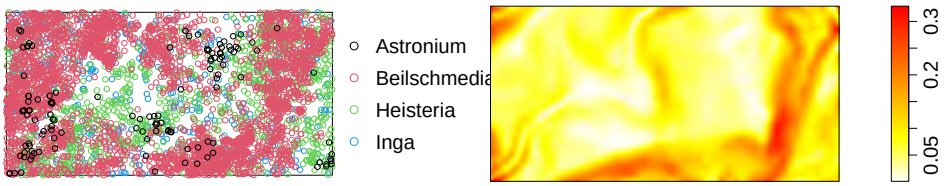



Figure 1: Rainforest tree data and covariate slope

$500000m^2$. The four species are $x_1 =$ *Inga sapindoides*, $x_2 =$ *Heisteria concinna*, $x_3 =$ *Beilschmiedia pendula*, $x_4 =$ *Astronium graveolens*, with $n_1 = 487$, $n_2 = 1141$, $n_3 = 3887$, $n_4 = 124$. Data is shown in Figure 1 together with an environmental covariate which measures the slope of the terrain. It is plausible that the slope somehow affects the growth of trees over the area as regards both richness and abundance, and thus influences the biodiversity of the system. In order to evaluate the biodiversity of the system, we partition the area into $50 \times 100$ cells of size $10 \times 10$ metres, each containing from 0 to 40 trees. Each cell is considered as one observation site $u$, and its spatial location is represented by the

centroid's coordinates. For each cell, we know the average values of the slope covariate and the counts of all species: our multinomial response variable is a table of $5000 \times 4$ counts.

We fitted the most general model together with all possible sub-models and selected the best one for entropy estimation using the Widely Applicable Information Criterion (WAIC) returned by INLA. The selected model includes covariate slope $z$ and a RW2d spatial effect applied to species 2, i.e. *Beilschmedia Pendula*, for model identifiability: $g_{ui} = \beta_{0i} + \beta_{1i} z_u + I(i = 2)\phi_u$. Based on such model we can derive, for each species, a smooth surface estimating its probability of occurrence over the observation area, and the final surface for the estimated entropy, which is displayed in the left panel of Figure 2. We can conclude that the biodiversity of the rainforest tree system depends on the soil slope, whose effect is significant on 3 of the 4 species. Indeed, *Astronium* trees tend to grow over flat areas, while *Beilschmiedia* and *Heisteria* trees tend to follow the pattern of the steepest areas (Figure 1). Moreover, entropy also shows an underlying spatial structure, whose effect particularly affects species *Beilschmiedia pendula*. By looking at Figure 2, one can detect the cold- and hot-spots as regards the latent biodiversity level. Note that the estimation procedure deals with areas with few or zero trees with no issues, thanks to the exploitation of a model that accounts for neighbouring sites.

All literature entropy estimators start from the estimated probability distribution based on relative frequencies, which is $\widehat{p}_X^{ML} = (0.087, 0.202, 0.689, 0.022)$. Data frequencies are the only needed information; covariates and spatial locations are discarded. In relative terms, the literature ML estimate corresponds to 63% of the maximum possible entropy ($\log 4$). Such value shows a low biodiversity level, which hints at an underlying structure in the data. Unfortunately, nothing more can be said with the literature estimators. In order to make a local comparison, we are forced to choose a rougher grid in order to have enough data per observation site to compute local relative frequencies. For all cells with 1 tree, the ML local entropy value degenerates to zero; for cells with no trees, no ML entropy value can be computed. The local ML estimate is in the right panel of Figure 2, presented in relative terms, i.e. ranging in $[0, 1]$ for comparability purposes. As can be seen, many cells are empty, and the remaining ones look like white noise and allow no conclusions on the behaviour of the biodiversity of the area. One can see that in some parts of the plot the two estimation approaches provide very similar results, for example the "cold spots"(light blue areas) are placed at similar locations, despite the noisy aspect of the ML estimate. A cold spot means low entropy, i.e. low biodiversity. The high entropy areas identified by our model are high *latent* biodiversity areas, according to the structure of the environmental covariate and of the spatial effect.

DRAFT

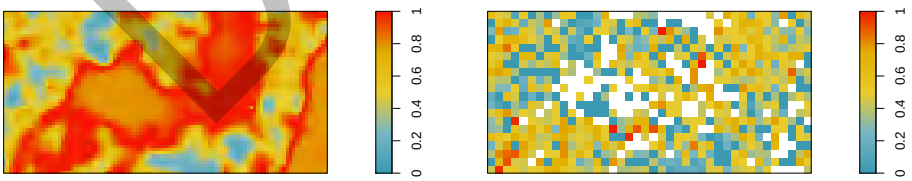



Figure 2: BMB entropy estimator (left); local ML entropy estimator (right)

## 4. Concluding remarks

The present work proposes a novel approach for entropy estimation suitable for ecological and environmental data, where independence is an unrealistic assumption. We highlight some key novel contributions to the field of biodiversity studies. First, our proposal relates entropy to covariates and/or data dependence structures, exploiting all the available information. Second, it allows to include information about local abundance and which species are present/absent at each observation point and at neighbouring points, by using multinomial data instead of presence/absence. Third, it uses INLA instead of MCMC (Markov Chain Monte Carlo) for estimation, overcoming many computational issues, and allows for extensions to continuous space if wished. In addition, we underline the ability of our estimation approach

to draw conclusions on the latent biodiversity of the system over a variety of situations, to check assumptions, select the most suitable model and evaluate the estimation uncertainty. We also show that it easily deals with empty, or nearly empty, observation points, thanks to the exploitation of a model structure. When wished, results can be further synthesized in a global entropy value for the whole dataset, that is well-grounded as it is built out of consideration of the data structures and not only of relative frequencies.

As for the rainforest tree species, we are able to find a relationship between biodiversity and the significant covariate describing the terrain slope. A smooth spatial effect is included in the model, so that our estimation result is an entropy surface that captures the biodiversity of the system, and where cold- and hot-spots in terms of tree species latent biodiversity can be easily identified. In comparison, we highlight the limit of producing a single number with no interpretation with the available methods, and the difficulties in obtaining a locally varying entropy surface.

## References


[1] Altieri, L., D. Cocchi, and G. Roli (2019). Advances in spatial entropy measures. *Stochastic Environmental Research and Risk Assessment 33*(4), 1223–1240.

[2] Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis 8*, 1–21.

[3] Frosini, B. V. (2004). *Descriptive measures of ecological diversity*. Paris, France: In Environmetrics. Edt J. Jureckova, A. H. El-Shaarawi in Encyclopedia of Life Support Systems (EOLSS), revised edn 2006.

[4] Gelfand, A. E. (2022). Spatial modeling for the distribution of species in plant communities. *Spatial Statistics 50*, 100582.

[5] Leibovici, D. G., C. Claramunt, D. LeGuyader, and D. Brosset (2014). Local and global spatio-temporal entropy indices based on distance ratios and co-occurrences distributions. *International Journal of Geographical Information Science 28*, 1061–1084.

[6] Møller, J. and R. P. Waagepetersen (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics 34*, 643–684.

[7] Paninski, L. (2003). Estimation of entropy and mutual information. *Journal of Neural Computation 15*, 1191–1253.

[8] Rue, H. and L. Held (2005). *Gaussian Markov random fields*. Boca Raton, Chapman and Hall.

[9] Serafini, F. (2019). *Multinomial logit models with INLA*. R-INLA tutorial.

[10] Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423 and 623–656.


DRAFT

# Feature Selection via anomaly detection autoencoders in radiogenomics studies

Alessia Mapelli[a,b], Michela Carlotta Massi[b], Nicola Rares Franco[a], Francesca Ieva[a,b], Catharine West[c], Petra Seibold[d], Jenny Chang-Claude[d], and the REQUITE and RADprecise Consortia

[a]MOX, Department of Mathematics, Politecnico di Milano, Milan 20133, Italy;
alessia.mapelli@polimi.it, nicolarares.franco@polimi.it

francesca.ieva@polimi.it

[b]HDS, Health Data Science Center, Human Technopole, Milan 20157, Italy;
michela.massi@fht.org

[c]The University of Manchester 555 Wilmslow Road, Manchester, M20 4G;
Catharine.West@manchester.ac.uk

[d]German Cancer Research Center (DKFZ) Im Neuenheimer Feld 280, 69120 Heidelberg;
p.seibold@dkfz-heidelberg.de, j.chang-claude@dkfz-heidelberg.de

DRAFT

**Abstract**

Genetic biomarkers are believed to be crucial in predicting the development of side effects of radiotherapy in cancer patients, and their inclusion in risk models may substantially improve personalized treatment planning. Feature selection is fundamental in genomic studies because of the high dimensionality of the data, but it is hindered by several complexities such as unbalanced classes, imputation and noise in genomic data collection, and the presence of high-order interactions among genes influencing the development of toxicities. In this study, we propose an interpretable feature selection of the most discriminant genetic variants via an ensemble of anomaly detection autoencoders designed to overcome these challenges. The model properties were studied in a simulation setting, and the method was applied to a case study.

***Keywords:*** Radiogenomics, late toxicity, anomaly detection autoencoder, denoising.

## 1. Introduction

Thanks to treatments, such as radiotherapy, the survival of patients diagnosed with cancer is increasing. However, approximately 5% of the patients receiving radiotherapy are particularly sensitive to irradiation and are likely to develop long-term side effects (2). These can occur years after radiotherapy, impairing their quality of life. Our work was developed within a large international study, namely RADprecise (2), aimed at personalizing radiotherapy treatment for cancer patients by improving prediction models for radiosensitivity (2). Radiosensitivity is a latent outcome, and it is only inferred through measurements of various types of Late Toxicities (LTs) quantified according to the Common Terminology Criteria for Adverse Events. Traditionally, physicians base treatment decisions on model-based risk

estimates known as Normal Tissue Complication Probabilities (NTCPs). Specifically, NTCPs model the risk of radiation-induced complications in terms of the radiation dose and partial volume irradiated. In recent years, the set of predictors has expanded to include clinical information and biomarkers, such as genetic variations, crucial for predicting LT development (1). A patient's genetic predisposition to a disease can be summarized in a Polygenic Risk Score (PRS). It is usually computed as the score associated with each patient by a predictive model that links the risk of developing LTs to the presence of associated genetic mutations in the patient's DNA. Its inclusion into wider risk prediction models may substantially improve personalized treatment planning.

In general, as with any other classification model, PRS models perform best when fed with highly influential features with discriminant properties for class separability. Moreover, feature selection (FS) is fundamental in genomic studies since variables are many and highly correlated and the curse of dimensionality plays an important role. Indeed, the present work focuses on the task of FS for genetic data. In the peculiar setting of genetic studies, proper FS is hindered by several endogenous and exogenous data complexities: high-dimensional genetic data are usually available for small samples; the study of rare phenotypical traits (such as LT) mostly determines unbalanced settings with very low case-control ratios, that may violate asymptotic assumptions of statistical inference. Additionally, several raw genetic features are not directly measured but estimated via imputation methods to achieve completeness in genetic information. Genetic variations are measured as categorical data representing the absence, or the type of mutation with DNA array-based approaches. However, this process can only be applied to a limited number of genetic variations. Imputation methods estimate genotype probabilities at variants not genotyped thanks to reference populations of more densely typed individuals and introduce genotypic uncertainty and noise in the analysis. Moreover, the latest radiogenic studies on late-toxicity radiotherapy revealed the biological relevance of gene-gene interactions in affecting polygenic susceptibility to common human diseases (1). This introduces another source of complexity: FS methods need to account for the potential predictive power of such interactions during selection. However, traditional FS techniques usually only consider the main effect of covariates when performing the selection, and become suboptimal when the high-order interactions are significant.

Some of the above-mentioned complexities were recently addressed in (3; 4), where the authors developed a Deep Sparse AutoEncoder Ensemble (DSAEE) method for unbalanced data. In brief, the DSAEE FS method exploits Deep Sparse AutoEncoders as weak learners. AutoEncoders are trained to learn the normal patterns in the majority class observations and tested on both the majority and minority class data. The reconstruction of the minority class is expected to be less efficient and the most discriminant features should present the highest reconstruction error difference in the two classes (4). The algorithm ranks feature importance based on the Reconstruction Error difference in the minority and majority class, within the test set, and selects those whose difference is above the predefined threshold. The FS method in (4) presents three major benefits: the ability to deal with heavy class imbalance, interaction-aware selection, and interpretability of the selection method. Notably, the genetic features selected by their DSAEE are subsequently included in an interaction-aware method for polygenic risk scoring (PRSi) (1). However, this effective algorithm does not account for possible noise in the genomic features, considering exclusively imputed continuous data. The main contribution of this study is the improvement of the DSAEE method to achieve robustness to imputation errors and enable an unbiased analysis in genomic studies.

DRAFT

## 2. Anomaly detection autoencoders for feature selection with imputed data

An AutoEncoder (AE) is a neural network trained to copy its input to its output. Mimicking the identity function, the AE learns an encoded version of the data compressing and aggregating information in input, in the best way for the network to reconstruct the original information from the latent representation. Let the matrix $\mathbf{X} \in R^{NxJ}$ be the set of N training vectors $\underline{x}_i$ characterized by J features. The network can be seen as constituted by two parts: an encoder and a decoder. The encoder maps each input

vector $\underline{x}_i$ into an encoded version of itself (i.e. a latent representation), usually in a low dimensional space of size H, with a function that can be represented as: $\underline{h}_i = f(\mathbf{W}\underline{x}_i + \underline{b})$. Here $f$ is the activation function, usually nonlinear, $\mathbf{W} \in R^{HxJ}$ is the weight matrix and $\underline{b}$ is a H-dimensional bias vector. The decoder maps back the latent representation vector to the original J-dimensional space with a function that can be represented as $\overline{\underline{x}}_i = g(\mathbf{W}'\underline{h}_i + \underline{b}')$. The parameters are defined analogously with $\mathbf{W}' \in R^{JxH}$ and $\underline{b}' \in R^J$. An AE can then be defined as a map $\phi(\underline{x}_i) : R^J \rightarrow R^J$

$$\phi(\underline{x}_i) = g(\mathbf{W}' f(\mathbf{W}\underline{x}_i + \underline{b}) + \underline{b}') \ , \tag{1}$$

and the parameters are optimized so that the reconstruction: $\overline{\underline{x}}_i = \phi(\underline{x}_i)$ is as close as possible, considering some loss $L(\underline{x}, \overline{\underline{x}})$, to $\underline{x}_i$. AEs typically do not provide exact reconstruction since $H <<< J$ but the latent representation is expected to be meaningful and a compact representation of the input (4). Better results can be achieved using constraints that force autoencoders to learn effective representations of such input in the latent space. In a Deep Sparse AutoEncoder (DSAE), the $L_1$ penalization is applied on $h(l)$, the function generating the latent representation, forcing the model to represent the input in the simplest way and incrementing generalization proprieties of the model.

$$L^S(\underline{x}_i, \phi(\underline{x}_i)) = L(\underline{x}_i, \phi(\underline{x}_i)) + \lambda |h(l)| , \tag{2}$$

The parameter $\lambda$ is usually optimized through grid search.

AEs are used for learning data representations, dimensionality reduction, and anomaly detection. An anomaly is a data point that is significantly different from the remaining data and arouses suspicion that it is generated by a different mechanism. One-class detection AEs are autoencoder-based anomaly-detection methods that exploit the reconstruction error as an anomaly score. In this case, an AE is trained exclusively on normal observations so that it reconstructs normal data very well while failing to do so with anomalous data that has not been encountered in training. Data points with high losses are considered to be anomalies. Our methodology, as the one presented in (3; 4), mimics AutoEncoders' usage in anomaly detection to perform FS.

We considered a binary supervised learning setup with an available set of N (input, target) pairs

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}_1}, Y_1), ..., (\underline{\tilde{x}_N}, Y_N)\} = (\tilde{\mathbf{X}}, Y)$$

where $Y_i$ is the endpoint that takes values in $\{0, 1\}$ and $\underline{\tilde{x}_i} \in R^J$ with $i = 1, ..., N$ is the input feature vector of imputed data or, in general, noisy data. Suppose that $\underline{x_i}$ true categorical feature vector is known for each sample present in the training set and that a fixed number of M categories is available for each feature. Therefore a second dataset is available with N (input, target) pairs

$$\mathbf{D} = \{(\underline{x_1}, Y_1), ..., (\underline{x_N}, Y_N)\} = (\mathbf{X}, Y)$$

where $\underline{x_i} \in \{1, ..., M\}^J$ with $i = 1, ..., N$ is the input feature vector of categorical data. In addition, suppose that imbalance in the classes is present, with a minority class $Y = 1$ (case class), whose numerosity will be referred to in the following as $O$, and a majority class $Y = 0$ (control class). Our full pipeline is illustrated in Figure 1.

The rationale of the denoising DSAEE follows in general the one presented in (3; 4) with the intention of including imputation noise in the reconstruction process. The idea behind the method is to exploit Denoising Autoencoder (DAE) to provide a customized denoising algorithm suited to our data. A DAE is a modification of the autoencoder to improve its generalization ability. Specifically, AEs are highly parametrized models and prompt to overfitting, especially in low sample settings, with the risk of generating a meaningless representation in the latent space. Denoising autoencoders solve this problem by corrupting the input data on purpose, adding noise or masking some of the input values, and forcing the network to reconstruct noise-free versions of their inputs.

In our methodology, we introduced a 'Denoising' version of the Deep Sparse Autoencoders to restore imputed noisy input data to their categorical correspondence under the hypothesis that, as autoencoders

DRAFT

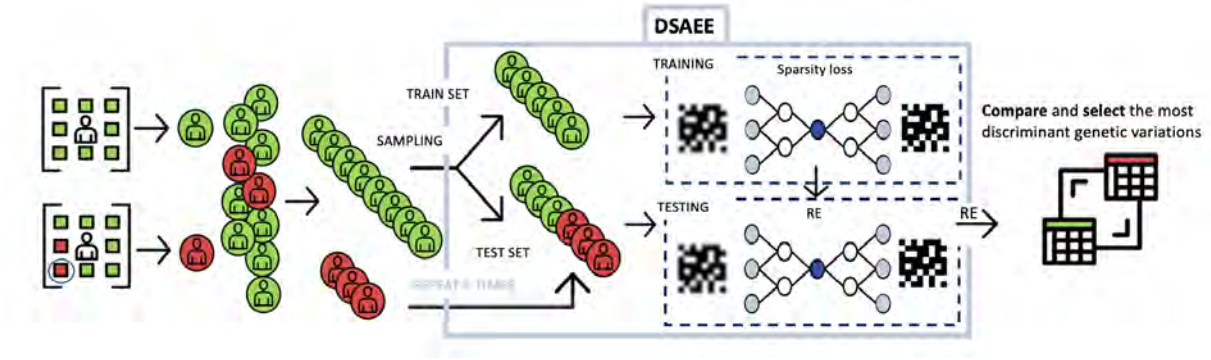



Figure 1: **Illustration of the developed methodlogy.** The method is based on an ensemble of Deep Sparse Autoencoders. In each ensemble iteration, the test set is constructed including all the minority class patients with a random sample of the same size from the majority class. The remaining observations of the majority class are included in the training set. The autoencoders are trained to optimally reconstruct the majority class and tested on a mixed population. Each reconstructs the continuous input into a categorical output so that the possible error due to imputed data is accounted for in the comparison. The most discriminant features are selected as those that present reconstruction error differences in the two classes.

DRAFT

are able to reconstruct corrupted input error, distinguishing between imposed noise and the underlying signal, similarly it is possible to force them to distinguish the true genomic signal from the random imputation noise.

At each iteration, once the train and test sets are defined, the network's weights and reconstruction map are optimized to have the best possible categorical reconstruction of the majority class noisy continuous features exploiting the loss in Equation (4).

$$\overline{\phi(\tilde{\underline{x}})} = argmin_{\phi} L(\underline{x}, \phi(\tilde{\underline{x}}))$$

$$L(x_j, \phi(\tilde{x}_j)) = -\sum_{k=0}^{M} (x_{jk} * log(\phi(\tilde{x}_{jk}))) \quad for \;\; j \in \{1, ..., J\} \tag{3}$$

$$L^S(\underline{x}, \phi(\tilde{\underline{x}})) = \sum_{j=1}^{J} L(x_j, \phi(\tilde{x}_j)) + \lambda |h(l)| \tag{4}$$

where $\tilde{\underline{x}} \in \tilde{\mathbf{X}}_{\mathbf{train}}$ and $\underline{x} \in \mathbf{X}_{\mathbf{train}}$. Once the network has been trained, the reconstruction error is evaluated on each sample of the test set as in (3). The test set REs from each of the B ensemble repetitions are scored in two matrices $\mathbf{Q_{maj}}, \mathbf{Q_{min}} \in R^{B*OxJ}$ based on each observation belonging to the majority or minority class.

For each feature j, the $B * O$ observations in $\mathbf{Q_{min}}$ and $\mathbf{Q_{maj}}$ are considered as a sample extracted from each feature distribution $re_j|$ minority sample $\sim f_j^{min}$ and $re_j|$ majority sample $\sim f_j^{maj}$. Performing a Smirnov test[1] we can compare the samples for each feature and select those with a p-value lower than the Bonferroni corrected threshold of 0.05.

## 3. Simualtion studies

Through a simulation study, we verified the capability of the denoising DSAEE to improve imputation error handling with respect to the DSAEE presented in (4), which reconstructs continuous noisy inputs to themselves. In the following, the latter will be referred to as continuous DSAE to better distinguish the two methodologies. Simulated data were built to reproduce the peculiar characteristics of genomic data: 1000 samples with 100 binary covariates, representing the variants, with a 10% relative

[1]The Smirnov test is a non-parametric two-sample test, used to determine if two independent random samples appear to follow the same distribution.

frequency, are considered. The outcome was defined with a minority class of 100 samples. Complex gene-gene interactions are represented by sets of interacting features containing 20 variants with a co-occurrence frequency with the phenotype of 70%. The noisy dataset was generated from the original categorical dataset by adding random exponential noise. The same simple AE architecture was implemented in the continuous and denoising DSAEE. The in-training convergence of both methods was analyzed through accuracy and loss: the MSE in the continuous DSAEE, where the continuous inputs are reconstructed as they are, and the binary-cross-entropy in the denoising DSAEE, where the categorical reconstruction is compared to the true discrete data. The in-train metrics are shown in Figure 2. The training was performed by excluding a validation set to mimic the performance of both algorithms on unseen data. The metrics of the validation sets are also presented in Figure 2.

We can observe that the denoising DSAEE maintains both the validation and training set losses low

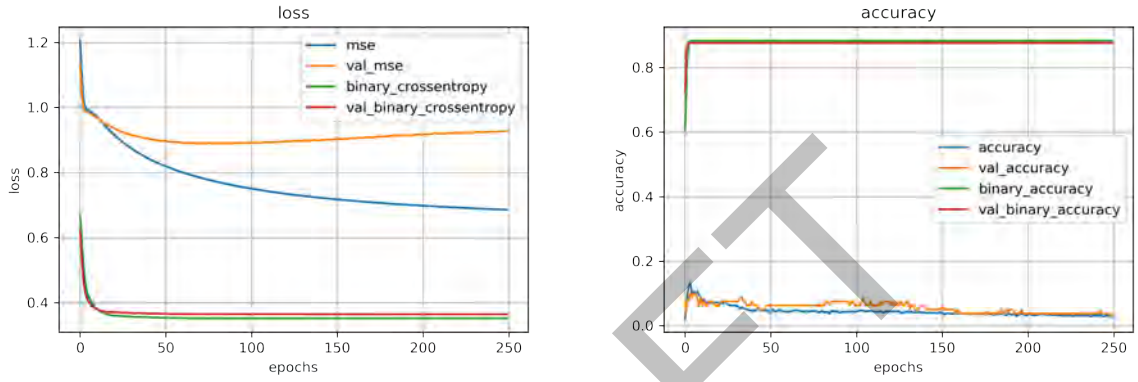



Figure 2: **In-train metrics of the compared methodologies.** The loss was evaluated via MSE in the continuous DSAEE and via cross-entropy in the denoising DSAEE.

and close with respect to the continuous DSAEE, showing a better reconstruction performance on seen and unseen data and better generalization ability. The accuracy in data representation is fundamental to better distinguish the classes of the binary outcome, and consequently to perform a better feature selection. Moreover, both the loss and accuracy plots reveal a faster and smoother convergence in denoising the DSAEE. Reducing the computational effort in training has great advantages in terms of the total computational time, enabling a higher number of repetitions to be performed and higher performance in ensemble learning algorithms.

DRAFT

## 4. Case study application in radiogenomics

As mentioned in the introduction, the selection and discovery of genomic variants predictive of late toxicities can inform downstream models such as PRSs and NTCPs. Therefore, in this section, we briefly present the case study application of the proposed algorithm on the RADPrecise Breast Cancer Cohort.

The considered sample includes 459 patients with a documented follow-up visit two years after the initial cancer treatment. Seven late toxicity endpoints, referred to as $\{y_1, ...., y_7\}$, are considered: six have an incidence below 10%, while $y_1$ occurs for approximately 30% of the subjects. The pool of genetic features to select from included 122 variants previously identified in the literature as correlated to radio-induced LTs in breast cancer patients. Both the developed denoising algorithm and the one presented in (4) were applied in this case study. The resulting selected variants were exploited to construct seven interaction-aware PRS for breast cancer late toxicities. The PRSs are computed following the PRSi algorithm presented in (1). In brief, starting from the selected genetic variants a data mining algorithm selects the most informative genetic interaction linked to each toxicity. The PRS is then built by weighting the contribution of each interaction term accordingly to the weights obtained when fitting a Logistic Regression model (LR). This being an unsupervised setting it is hard to comment on the precision of the results. However, the metrics of the logistic regression exploited in the definition of the PRSs reveal bet-

Table 1: AUC of the logistic regression in PRSi definition

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ |
|---|---|---|---|---|---|---|---|
| Continuous DSAEE | 0.55 | 0.54 | 0.54 | 0.62 | 0.36 | 0.64 | 0.5 |
| Denoising DSAEE | 0.56 | 0.58 | 0.71 | 0.70 | 0.66 | 0.54 | 0.65 |

ter predictive abilities in the denoising DSAEE with respect to the continuous DSAEE, probably induced by a more effective initial selection. Results are reported in Table 1.

## 5. Conclusion

The innovation of this study is the development of a methodology capable of performing feature selection in the peculiar setting of genomic studies, overcoming the challenges of feature imputation (i.e., noise), class imbalance, and the need to account for predictive high-order interactions among features. The proposed method builds upon the original work in (4) and, based on our studies, we can say that it succeeds in the improvement of the representation of noisy data and of the selection of the most discriminant features.

The importance of this model is its clinical applicability. In addition to selection, the model can be exploited for the discovery and validation of influential features that determine the phenotype of interest, and for the interpretation of biologically relevant variants. The developed method, which improves the definition of genetic predisposition to general toxicities, may substantially improve personalized treatment planning. Some of the limitations of the developed model are the need for a ground-truth definition of noisy input data and the difficulty in scaling the input features owing to the high computational cost. Further developments can be introduced into this model. Variational autoencoders have already been proposed for anomaly detection, and their theoretical background makes them a more principled and objective method than classical autoencoders. However, variation autoencoders require a large training dataset, which hinders their applicability in this study.

DRAFT

**Acknowledgments** This research was made possible thanks to the ERA PerMed Cofund program, grant agreement No. ERAPERMED2018-44, RADprecise-Personalized radiotherapy: incorporating cellular response to irradiation in personalized treatment planning to minimize radiation toxicity.

## References

[1] Franco, N. R., Massi, M. C., Ieva, F. et al.: Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. Radiotherapy and Oncology, 159:241-248 (June 2021) doi: 10.1016/j.radonc.2021.03.024

[2] Krebsforshungszentrum, D.: RADprecise. In: Reasearch. Deutsches Krebsforschungszentrum (n.d.)
https://www.dkfz.de/en/epidemiologie-krebserkrankungen/units/genepi/ge_pr13_RADprecise.html

[3] Massi, M. C., Gasperoni, F., Ieva, F. et al.: A Deep Learning Approach Validates Genetic Risk Factors for Late Toxicity After Prostate Cancer Radiotherapy in a REQUITE Multi-National Cohort. Frontiers in Oncology, 10:541281 (Oct. 2020) doi: 10.3389/fonc.2020.541281.

[4] Massi, M. C., Gasperoni, F., Ieva, F. et al.: Feature selection for imbalanced data with deep sparse autoencoders ensemble. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(3):376-395 ( June 2022) doi: 10.1002/sam.11567.

# Further considerations on the Spectral Information Criterion

Luca Martino[a]

[a]Dip. di Economia e Impresa, Università di Catania.

**Abstract**

In this work, we extend and analyze some aspects of the spectral information criterion (SIC) considering sub-linear and super-linear model penalizations. We provide numerical results applying the method in a variable selection problem with a real dataset.

***Keywords:*** Model selection, Information criteria, AIC, BIC, Spectral Information Criterion.

DRAFT

## 1. Introduction

Model selection can be considered one of the fundamental tasks of scientific inquiry. Indeed, the majority of the problems in statistical inference can be interpreted in some way as a statistical modeling problems [3]. The information criteria are one of the most important tools for model selection in nested models, that have introduced in the literature [9]. Several information criteria have been proposed, such as a well-known examples are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) [7; 8]. Here we consider an approach, the spectral information criterion (SIC), which includes several information criteria as special cases [4].

More specifically, in this work we consider an observation model which connects a vector of $k$ parameters $\boldsymbol{\theta}_k = [\theta_1, ..., \theta_k]^\top$, to some observed vector of data $\mathbf{y}$. Namely, we know and can evaluate a likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_k)$. Our goal is make inference about $\boldsymbol{\theta}_k$ given the set of observed data $\mathbf{y}$. In many applications, the interest is also learning the dimension $k$ of $\boldsymbol{\theta}_k$, with $k \leq K$ (this is the case of the so-called *nested models*). Note that $k$ can represent the order of a polynomial function or the number of feature in a regression problem, or the number of clusters in a clustering problem, etc. We remark that the maximum number of components/variables/clusters (depending on the specific application) is denoted as $K < \infty$.[1] In this work, we focus on the task of selecting the optimal number of components $k^* \leq K$. We also refer to $k^*$ as a possible "elbow" of the problem. In several parts of the work, we refer specifically to the nomenclature and notation of a variable selection problem, without loss of generality just clarity in the exposition.

Furthermore, recalling that $k$ is the dimension of the parameter vector parameters $\boldsymbol{\theta}_k$, we consider a

[1]We assume $K$ finite, just for the sake of simplicity. However all the results can be extended $K \to \infty$.

function $V(k)$ represents a generic non-increasing function[2] with a finite value at $k = 0$, i.e., $V(0) < \infty$ (hence $V(k)$ takes always finite values). This fitting term can be obtained directly from the likelihood function, we can have

$$V(k) = -2\log(\ell_{\max}) \quad \text{where} \quad \ell_{\max} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k). \tag{1}$$

However, other choices (in some cases equivalent) can be considered. For instance, the root mean square error (MSE) or the mean absolute error (MAE) in a regression problem, i.e., $V(k) = \text{MSE}(k)$, as a function of an integer $k$, where $k$ can represent the order of a polynomial or the number of variables involved in the regression. We could also set $V(k) = 1 - \text{Accuracy}(k)$ in a classification problem using the first $k$ most important features, or $V(k)$ can present the $k$-th eigenvalue of the covariance matrix of the data in a principal component analysis (PCA), where the eigenvalues are ordered in decreasing order etc. Note that, as an example, $k = 0$ corresponds to a constant model in a regression problem, when the case of "no variables" are used (in a variable selection example), i.e., $V(0) = \text{var}(\mathbf{y})$ which is the variance of the data. Finally, without of lost of generality, we can always assume

$$\min_k V(k) = V(K) = 0, \tag{2}$$

since we can always set $V'(k) = V(k) - \min_k V(k) = V(k) - V(K)$, where we have used $\min_k V(k) = V(K)$ since $V(k)$ is a non-increasing function. Below, we recall quickly the SIC procedure and its more relevant features.

## 2. Spectral information criterion (SIC) method

In this section, we recall briefly the spectral information criterion (SIC) method [4]. Considering a linear penalization, we introduce the cost function that we desire to minimize,

$$C(k, \lambda) = V(k) + \lambda k, \quad k = 0, ..., K, \quad \lambda \in [0, \lambda_{\max}], \tag{3}$$

where $V(k)$ is a generic fitting term (we assume $V(k) \geq 0$ without loss of generality; see below), $\lambda k$ is a penalization term of model complexity, where $\lambda$ is a constant and $k$ represents the dimension of the model. We study all the possible values of $\lambda \in [0, \lambda_{\max}]$ where $\lambda_{\max}$ is defined as

$$\lambda_{\max} = \{\min \lambda : \quad \arg\min_k C(k, \lambda) = 0\}, \tag{4}$$

and can be analytically obtained as

$$\lambda_{\max} = \max_k \left[ \frac{V(0) - V(k)}{k} \right], \quad \text{for } k = 1, ..., K. \tag{5}$$

Since above we consider $k = 1, 2, ..., K$, we can perform an exhaustive search and, considering Eq. (11), then obtain $\lambda_{\max}$. The expression above can be easily understood: given $\lambda'_k = \frac{V(0)-V(k)}{k} > 0$, we have $C(k, \lambda'_k) = V(k) + \lambda'_k k = V(0)$, that can be obtained just replacing $\lambda'_k$ inside Eq. (3). Since $\lambda_{\max} \geq \lambda'_k$ by definition (and $V(k) \geq 0$, $k \geq 0$ are both positive values), we have that $C(k, \lambda_{\max}) \geq C(k, \lambda'_k) = V(0)$. Therefore, the minimum value will be $\min C(k, \lambda_{\max}) = V(0)$ since we are sure that is reached, at least once, at $k^* = 0$.

The underlying idea in SIC is similar to the idea of to "integrating out" $\lambda$ as usually done in Bayesian analysis, i.e., removing the dependence of $\lambda$ in our problem. Namely, unlike the other IC schemes in the literature, we avoid picking a specific value of $\lambda$. For the sake of simplicity, let assume in this section that $V(k)$ is a decreasing function, with $V(0) < \infty$. With this assumption, it can be proved that $C(k, \lambda)$ has a unique minimum. See a graphical example in Figure 1(a). Now, we study the function

[2] This condition can be relaxed.

DRAFT

$k^*(\lambda) : [0, \lambda_{\texttt{max}}] \subset \mathbb{R} \to \{0, 1, 2..., K\}$, defined as

$$k^*(\lambda) = \arg\min_k C(k, \lambda), \tag{6}$$

which takes real values in the interval $[0, \lambda_{\texttt{max}}]$ and convert them into discrete values within the set $\{0, 1, 2..., K\}$. It is a non-increasing, piecewise constant function where $k^*(0) = K$ and $k^*(\lambda) = 0$ for $\lambda \geq \lambda_{\texttt{max}}$, i.e.,

$$\begin{cases} k^*(0) = K, \\ k^*(\lambda_{\texttt{max}}) = 0, \end{cases} \tag{7}$$

as shown in Figure 1(b). A very important consideration is that some values of $k \in \{0, 1, 2..., K\}$ could not have a corresponding $\lambda$ associated. Moreover, a complete example of the piecewise constant function $k^*(\lambda)$ is given in Figure 1(b). Several values of $\lambda$ can be associated with the same minimum $k^*$, or some value $k'$ could not have any $\lambda$ associated. Generally, to each $k$, we can associate an interval of lambda values, $\mathcal{S}_k \subset [0, \lambda_{\texttt{max}}]$. Observe that $\mathcal{S}_0 =$ by definition since we consider $\lambda \in [0, \lambda_{\texttt{max}}]$, so that $|\mathcal{S}_0| = 0$. These intervals, for $k = 1, ..., K$, form a partition of $[0, \lambda_{\texttt{max}}]$, i.e., $\mathcal{S}_1 \cup \mathcal{S}_2 ... \cup \mathcal{S}_K = [0, \lambda_{\texttt{max}}]$, and $\mathcal{S}_k \cap \mathcal{S}_j = 0$, for all $k \neq j$. As stated above, some value $k' \neq 0$ could be never a minimum, so that $|\mathcal{S}_{k'}| = 0$.

We can use the information provided by the measures $|\mathcal{S}_k|$, defining the weights $\bar{w}_k \propto |\mathcal{S}_k|$, i.e.,

$$\bar{w}_k = \frac{|\mathcal{S}_k|}{\sum_{j=0}^{K} |\mathcal{S}_j|} = \frac{|\mathcal{S}_k|}{\sum_{j=1}^{K} |\mathcal{S}_j|}, \quad k = 0, ..., K, \tag{8}$$

where we have used $|\mathcal{S}_0| = 0$, since $\lambda \leq \lambda_{\texttt{max}}$. Note that $\bar{w}_k$, for $k = 1, ..., K$, defines a probability mass function (pmf), $\sum_{k=1}^{K} \bar{w}_k = 1$. The main part of the SIC method is to compute (approximately) the probabilities $\bar{w}_k$. This approximation can be obtained with a quasi-Monte Carlo strategy (i.e., with a simple grid) or with a standard Monte Carlo approach using a number $M$ of samples (see Table 1). The algorithm in Table 1 is generally fast even with choices of $M$ such as $M = 10^6$, $M = 10^7$, or greater.

Table 1: Computation of the weights in the SIC method by Monte Carlo.

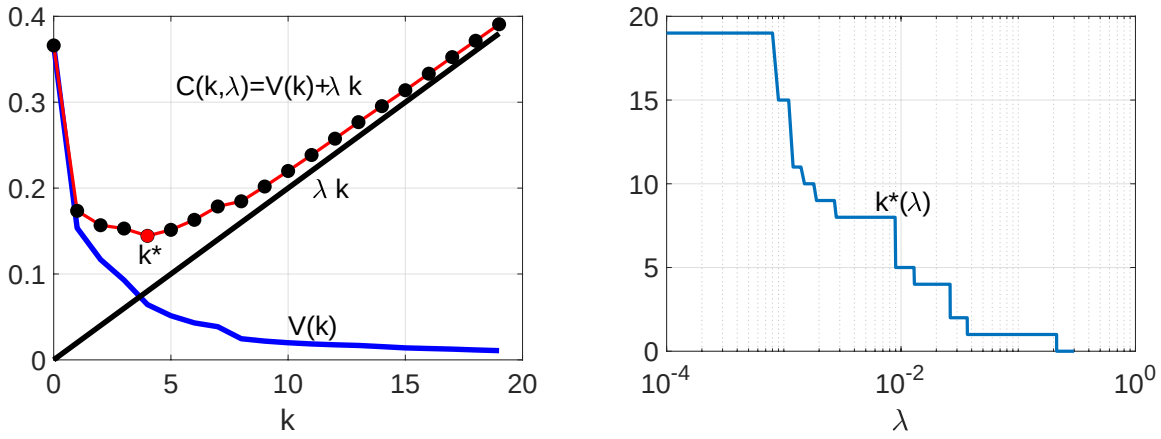
- For $i = 1, ..., M$ :
  1. Draw $\lambda_i \sim \mathcal{U}([0, \lambda_{\texttt{max}}])$.
  2. Compute $k_i^* = \arg\min_k C(k, \lambda_i) = \arg\min_k [V(k) + \lambda_i k]$.
- Return the number of occurrences of the event $\{k_i^* = j\}$ for $j = 1, ..., K$, or equivalently return the weights

$$\bar{w}_j = \frac{\#\{k_i^* = j\}}{M}, \qquad j = 1, ..., K. \tag{9}$$

**The output and main features of SIC.** We can define the set $\mathcal{E}$ of indices $k$ such that the corresponding weight is non-zero, $\bar{w}_k > 0$,

$$\mathcal{E} = \{\text{all } k : \ \bar{w}_k > 0\} = \{k^{(1)}, k^{(2)}, ..., k^{(J)}\}.$$

DRAFT




(a) (b)

Figure 1: **(a)** An example of function $V(k)$, a penalization term $\lambda k$, and the corresponding cost function $C(k, \lambda)$ (shown with dots). **(b)** Example of piecewise constant function $k^*(\lambda)$ in log-domain.

DRAFT

They can be interpreted as a possible "elbow" of the curve, i.e., any possible selected model is represented by the index $k^{(j)}$. We have denoted $J = |\mathcal{E}|$. Note that $J \leq K$ and, in some cases, $J << K$. In [4], some strategies for picking a unique index inside $\mathcal{E}$ have been discussed, i.e., choosing a unique specific model. However, in this work we focus on studying the changes in the set $\mathcal{E}$ when we change the type of model penalization.

**Remark.** Note that the set $\mathcal{E}$, by definition, contains all the IC solutions in the literature which use a cost function of type in (3), such as AIC and BIC, for instance.

**Choosing Uniformly in $[0, \lambda_{\max}]$.** Moreover, the values of $\lambda$'s are choosing uniformly in $[0, \lambda_{\max}]$. One could consider a different procedure, similarly to use a generic (non-uniform) prior over $\lambda$ in a complete Bayesian approach.

**Linear penalization.** In Eq. (3), we employ a *linear* penalization of the complexity $\lambda \cdot k$, since this linear term appears in different theoretical derivations in the literature [1; 7; 8]. Moreover, it appears not just in several IC formulations but also in other more general approaches, e.g., involving marginal likelihood with uniform priors [2, App. A and B] and alternative geometric solutions [5]. Then, assuming a linear penalty for the complexity seems to have strong theoretical support from different points of view.

However, in this work, we study the behavior of SIC considering different penalizations as power of the number of components $k$, i.e., $k^r$.

## 3. Extended SIC (ESIC): changing the model penalization

In this section, we consider the following cost function,

$$C(k, \lambda, r) = V(k) + \lambda k^r, \quad k = 0, ..., K, \quad \lambda \in [0, \lambda_{\max}^{(r)}], \quad r \in (0, R]. \tag{10}$$

where, for each $r \in (0, R]$, we have

$$\lambda_{\max}^{(r)} = \max_k \left[ \frac{V(0) - V(k)}{k^r} \right], \quad \text{for } k = 1, ..., K, \tag{11}$$

that can be computed analytically. Furthermore, for each $r$, we obtain a probability mass $\bar{w}_k^{(r)}$, with $k = 0, ..., K$. For simplicity, we consider a value of $R$ great enough such that it provides, for any $\lambda \leq \lambda_{\max}^{(R)}$, all the weight mass concentrated at $k = 1$, i.e., $\bar{w}_1^{(R)} \approx 1$. Fixing a big $M$ in Table 1, The idea is to analyze the outputs of SIC for different values of $r \in (0, R]$, studying the set of possible models (indices denoting the "elbows", number of parameters in each model)

$$\mathcal{E}_r = \{\text{all } k : \bar{w}_k^{(r)} > 0\} = \{k_r^{(1)}, k_r^{(2)}, ..., k_r^{(J_r)}\},$$

where $J_r = |\mathcal{E}_r|$. Note that, for $0 < r < 1$, we have sub-linear penalizations whereas, for $r > 1$, we have super-linear penalizations. This analysis can show that which of the points contained in $\mathcal{E}_r$ obtained by SIC have really a geometric relevance with respect to the non-increasing curve $V(k)$.

## 4. Numerical results

We test ESIC in a variable selection problem with real data (within a regression framework). Generally, we have a dataset of $N$ pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$, where each input vector $\mathbf{x}_n = [x_{n,1}, ..., x_{n,K}]$ is formed by $K$ variables, and the outputs $y_n$'s are scalar values. We consider the case that $K \leq N$ and assume a linear observation model,

$$y_n = \theta_0 + \theta_1 x_{n,1} + \theta_2 x_{n,2} + ...\theta_K x_{n,K} + \epsilon_n, \tag{12}$$

where $\epsilon_n$ is a Gaussian noise with zero mean and variance $\sigma_\epsilon^2$, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. Here, we consider a real dataset studied in [6], there are $K = 122$ features and $N = 1214$ number of data. Moreover, the dataset in [6] has two outputs: "arousal" and "valence". Here, we study the "arousal" output.
In this experiment, we set $V(k) = -2\log(\ell_{\max})$ with $\ell_{\max} = \max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}_k)$ with $k \leq K$ (we also use the trick $V'(k) = V(k) - \min_k V(k) = V(k) - V(K)$ to have $V(K) = 0$), after ranking the 122 variables (see [6]). The likelihood $p(\mathbf{y}|\boldsymbol{\theta}_k)$ is induced by the Eq. (12). Hence, in this experiment, we can compare again with other IC measures in the literature. See Table 2. We employ $M = 10^6$ samples for ESIC (for, $r > 2$ we use $M = 5 \cdot 10^6$ to stabilize numerically the results). Note that, in all the cases, $J_r = |\mathcal{E}_r| << 122$, i.e., the number of suggested models are quite less the total number of possible models. The maximum number of models is $J_{1.905} = 26 << 122$ obtained for a super-linear penalization $r = 1.905$. The information criteria proposed in the literature are the first in appearing and also show a certain resistance in disappearing, as $r$ grows.

DRAFT

## 5. Conclusions

We have tested an extended version of the spectral information criterion (SIC), which contains as special cases several information criteria given in the literature. The ESIC procedure is able to extract geometric information from the non-increasing curve $V(k)$. The results show that the "elbow points" representing other information criteria in the literature appear easily even with sub-linear model penalizations. This can be seen as a confirmation that these points are particularly relevant "elbows". Moreover, the method introduced in this work could be employed to select the order $r$ (e.g., sub-linear if $r < 1$, or super-linear if $r > 1$) of the model penalization.

## References

[1] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195, 1979.

[2] F. Llorente, L. Martino, E. Curbelo, J. Lopez-Santiago, and D. Delgado. On the safe use of prior densities for bayesian model selection. *WIREs Computational Statistics*, page e1595, 2022.

Table 2: Results of the numerical simulations ($M = 10^6$ samples for each $r$). The results obtained applying other information criteria proposed in the literature are highlighted in red: AIC=44 [8], HQIC=41 [1], BIC=15 [7], AED=11 [5].

| **Value of** $r$ | $\mathcal{E}_r$ | $J_r = \lvert\mathcal{E}_r\rvert$ |
|---|---|---|
| $r = 21$ | $\{1\}$ | 1 |
| $r = 15$ | $\{1, 2\}$ | 2 |
| $r = 10$ | $\{1, 2, 3\}$ | 3 |
| $r = 7$ | $\{1, 2, 3, 5, 6, 7\}$ | 6 |
| $r = 6$ | $\{1, 2, 3, 5, 6, 7, 8, 9\}$ | 8 |
| $r = 5$ | $\{1, 2, 3, 5, 6, 7, 8, 10, 11, 12\}$ | 10 |
| $r = 4$ | $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17\}$ | 15 |
| $r = 2.995$ | $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 20, 25, 36, 37\}$ | 19 |
| $r = 2.945$ | $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 20, 24, 25, 28, 37\}$ | 20 |
| $r = 2.915$ | $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 20, 24, 25, 28, 40, 41\}$ | 21 |
| $r = 2.835$ | $\{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 19, 20, 24, 25, 28, 41, 70, 75\}$ | 22 |
| $r = 2.590$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 70, 96\}$ | 23 |
| $r = 2.335$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 16, 17, 19, 20, 24, 25, 28, 40, 41, 44, 70, 96, 122\}$ | 24 |
| $r = 2.010$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 25 |
| $r = 1.905$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 69, 70, 71, 96, 122\}$ | 26 |
| $r = 1.710$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 15, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 25 |
| $r = 1.600$ | $\{1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 24 |
| $r = 1.570$ | $\{1, 2, 3, 5, 6, 7, 9, 11, 12, 16, 17, 19, 20, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 23 |
| $r = 1.445$ | $\{1, 2, 3, 5, 6, 7, 9, 11, 12, 16, 17, 19, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 22 |
| $r = 1.340$ | $\{1, 2, 3, 5, 6, 7, 9, 11, 16, 17, 19, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 21 |
| $r = 1.145$ | $\{1, 2, 3, 5, 6, 7, 9, 11, 16, 17, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 20 |
| $r = 1$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 19 |
| $r = 0.690$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 28, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 19 |
| $r = 0.650$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 46, 70, 71, 96, 122\}$ | 18 |
| $r = 0.620$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 46, 70, 96, 122\}$ | 17 |
| $r = 0.605$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 70, 96, 122\}$ | 16 |
| $r = 0.430$ | $\{1, 3, 5, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 96, 122\}$ | 15 |
| $r = 0.420$ | $\{1, 3, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 96, 122\}$ | 14 |
| $r = 0.415$ | $\{3, 6, 7, 9, 11, 16, 17, 25, 40, 41, 44, 96, 122\}$ | 13 |
| $r = 0.355$ | $\{3, 6, 7, 9, 11, 16, 17, 25, 41, 44, 96, 122\}$ | 12 |
| $r = 0.345$ | $\{3, 7, 9, 11, 16, 17, 25, 41, 44, 96, 122\}$ | 11 |
| $r = 0.340$ | $\{3, 7, 9, 11, 16, 17, 41, 44, 96, 122\}$ | 10 |
| $r = 0.265$ | $\{7, 9, 11, 16, 17, 41, 44, 96, 122\}$ | 9 |
| $r = 0.130$ | $\{7, 9, 11, 17, 41, 44, 96, 122\}$ | 8 |
| $r = 0.105$ | $\{9, 11, 17, 41, 44, 96, 122\}$ | 7 |
| $r = 0.090$ | $\{11, 17, 41, 44, 96, 122\}$ | 6 |
| $r = 0.050$ | $\{17, 41, 44, 96, 122\}$ | 5 |
| $r = 0.035$ | $\{41, 44, 96, 122\}$ | 4 |
| $r = 0.025$ | $\{96, 122\}$ | 2 |
| $r = 0.01$ | $\{122\}$ | 1 |

DRAFT

[3] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM Review (SIREV)*, 65(1):3–58, 2023 - arXiv:2005.08334.

[4] L. Martino, R. S. Millan-Castillo, and E. Morgado. Spectral information criterion for automatic elbow detection. *preprint - viXra:2209.0123*, pages 1–20, 2022.

[5] E. Morgado, L. Martino, and R. San Millán-Castillo. Universal and automatic elbow detection for learning the effective number of components in model selection problems. *preprint - viXra:2209.0132*, pages 1–12, 2022.

[6] R. San Millán-Castillo, L. Martino, E. Morgado, and F. Llorente. An exhaustive variable selection

study for linear models of soundscape emotions: Rankings and Gibbs analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2460–2474, 2022.

[7] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[8] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, 64:583–616, 2002.

[9] P. Stoica and Y. Selén. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, pages 36–47, 2004.

DRAFT

# How to increase the power of the test in sparse contingency tables: a simulation study

Federica Nicolussi[a] and Manuela Cazzaro[b]

[a]Politecnico di Milano; federica.nicolussi@polimi.it
[b]Università di Milano-Bicocca; manuela.cazzaro@unimib.it

**Abstract**

When analyzing categorical or ordinal data one often comes across sparse contingency tables. One of the biggest problems with this situation is due to the low power of tests for independencies between variables. In this paper, we propose a procedure, based on context-specific independencies, to increase the power of tests. The idea is to focus on sub-tables where the number of null cells is relatively low. In addition, including these kinds of independencies in the Union-Intersection procedure can provide comforting results. These results are shown in a simulative study of different scenarios.

***Keywords:*** Context-specific independencies, sub-tables, union-intersection principle

DRAFT

## 1. Introduction

The study of relationships between categorical or ordinal variables is often reduced to frequency analysis in contingency tables (see (1)). However, as the number of variables or the number of allowable categories for each variable increases, the corresponding contingency tables become sparse (full of null cells). The main problem when dealing with sparse tables is the low power of the classical statistics and an inaccurate type I error. Many works in the literature have dealt with this topic. For instance, in (7), a study on sparsity, it is showed that Fisher's exact test and the asymptotic $X^2$ Pearson's test give contradictory results for high levels of sparseness. In (12), it is studied the goodness of fit of the $X^2$ test, likelihood ratio test $G^2$, and Cressie-Read statistics. In (6) it is showed that the Gaussian approximation of the likelihood ratio statistic $G^2$ is more accurate than the $\chi^2$ approximation in sparse contingency tables. Further, the problematic likelihood ratio test's asymptotic properties in sparse tables are studied in (4) and (5).

In the case of sparseness, often, at the exploratory stage, many categories are merged because they are poorly tested in the observed sample. However, this procedure can distort the results obtained by inflating the frequencies of some categories. Moreover, even in the best situations, categorically aggregating leads to a loss of informativeness. In this paper, we propose to take advantage of the study of conditional independencies defined in sub-tables that identify a dense portion of the contingency table related to all the variables under consideration. An example of these relationships is context-specific independencies (CSI) (see (8) and (9)) that study the relationship between two groups of variables in conditional tables. Generally speaking, context-specific independencies are conditional independencies holding for particular values of the variables in the conditioning set. Next, we broaden this concept by considering portions of variables for which independence applies. Subsection 2.1 is devoted to defining these independence relations and the parametric model needed to determine them. Subsection 2.2, on the other hand, explains how the Union-Intersection procedure (see (10) and (11)) can be applied to this type of independence to

extend the previously identified relationships, where possible, over the entire contingency table. Section 3 is devoted to the study of simulations to support the theory presented. Conclusions are reserved for Section 4.

## 2. Methodology

In the following subsections, we define a new type of independencies defined on contingency sub-tables. We also see how their use can be employed to have greater power in the likelihood ratio test than the standard test on the whole contingency table.

### 2.1 Context-specific independencies and their extensions

Let us consider a vector of random variables $X_V = (X_j)_{j \in V}$ where each variable $X_j$ takes value $i_j$ in a set of finite categories $\mathcal{I}_j = (1, \ldots, i_j, \ldots, I_j)$. Let $|\cdot|$ be the cardinality of a set. The contingency table of the $|V|$ variables is defined by $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$ where each cell is defined as $\boldsymbol{i}_V = (i_j, j \in V)$. The strictly positive probability associated with any cell $\boldsymbol{i}_V$ is denoted with $\pi(\boldsymbol{i}_V)$. The vector of the probability of the whole contingency table is represented by $\boldsymbol{\pi}$, obtained by stacking each $\pi(\boldsymbol{i}_V)$ in the lexicographical order. Similarly, by considering a subset of variables $X_{\mathcal{M}}$ with $\mathcal{M} \subseteq V$ which generates the marginal $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}} = \times_{j \in \mathcal{M}} \mathcal{I}_j$, the marginal probability of the generic cell $\boldsymbol{i}_{\mathcal{M}}$ is $\pi(\boldsymbol{i}_{\mathcal{M}})$, obtained by summing with respect to the variables in $X_{V \setminus \mathcal{M}}$. The whole set of these marginal probabilities defined on the $\mathcal{M}$-contingency table $\mathcal{I}_{\mathcal{M}}$, is represented by the vector $\boldsymbol{\pi}_{\mathcal{M}}$.
Given three incompatible subsets of variables $X_A$, $X_B$ and $X_C$, a CSI is a independence statement like

$$X_A \perp\!\!\!\perp X_B | (X_C = \boldsymbol{i}'_C), \qquad \boldsymbol{i}'_C \in \mathcal{K}_C, \tag{1}$$

where $\boldsymbol{i}'_C$ is the vector of certain level(s) of the variable(s) in $X_C$, such that $X_j = i'_j$ for all $j \in C$, and it takes value in the list of levels $\mathcal{K}_C \subseteq \mathcal{I}_C$ for which the independence in formula (1) holds. Here, the table obtained as a cartesian product of $\mathcal{K}_C$ and $\mathcal{I}_{AB}$ is the sub-table where the independence is defined. The following formula is a generalization of the previous CSI where also the first two arguments of the independence statement are constrained to a subtable. Hereafter, we refer to this relationship as sub-CSI:

$$(X_A = \boldsymbol{i}'_A) \perp\!\!\!\perp (X_B = \boldsymbol{i}'_B) | (X_C = \boldsymbol{i}'_C), \qquad (\boldsymbol{i}'_A, \boldsymbol{i}'_B, \boldsymbol{i}'_C) \in \mathcal{K}_A \times \mathcal{K}_B \times \mathcal{K}_C \tag{2}$$

or in short $\boldsymbol{i}'_A \perp\!\!\!\perp \boldsymbol{i}'_B | \boldsymbol{i}'_C$, with $(\boldsymbol{i}'_A, \boldsymbol{i}'_B, \boldsymbol{i}'_C) \in \mathcal{K}_A \times \mathcal{K}_B \times \mathcal{K}_C$. Trivially, by replacing $\mathcal{K}_A$ with $\mathcal{I}_A$ and $\mathcal{K}_B$ with $\mathcal{I}_B$ in formula (2), we easily obtain the CSI in formula (1).
Although the class of sub-CSIs may seem difficult to interpret and of little use, a first advantage lies in the fact that the definition of these statements corresponds to linear constraints on log-linear parameters. For a more comprehensive treatment of the phenomenon, below we take advantage of marginal models, see e.g. (2) which impose constraints on marginal distributions of the tables in order to test different independence hypotheses. More specifically, we focus on hierarchical multinomial marginal (HMM) models, see (8) and (9) for interesting applications. In HMM models, the elements of $\boldsymbol{\eta}$ are the parameters based on different types of logits and defined on marginal distributions. The whole parametrization can be expressed in matrix form as

$$\boldsymbol{\eta} = C \log(M \boldsymbol{\pi}) \tag{3}$$

where $C$ is a contrasts matrix and $M$ is a 1's and 0's matrix which elements provide a suitable sum of probabilities. In general, the vector of parameters associated with the variables in $X_{\mathcal{L}}$ and defined in the marginal table $\mathcal{I}_{\mathcal{M}}$, $\boldsymbol{\eta}^{\mathcal{M}}_{\mathcal{L}} = \left\{\eta^{\mathcal{M}}_{\mathcal{L}}(\boldsymbol{i}_{\mathcal{L}})\right\}_{\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{I}_{\mathcal{L}} - \mathbf{1})}$ where $\mathbf{1}$ represents the first (reference) cell used for the *baseline* codification of the parameters. The above parameters are contrasts of logarithms of sums of probabilities.

**Theorem 1.** *Let us consider a set of variables $X_V$, with probability distribution $\mathcal{P}$ parametrized through the parameters in formula (3), where the baseline criterion is used. Then, the probability distribution*

DRAFT

*$\mathcal{P}$ obeys the sub-CSI in formula (2) if and only if the following constraints on the HMM parameters are satisfied:*

$$\sum_{c \subseteq C} \eta_{\mathcal{L}}^{\mathcal{M}}(\boldsymbol{i}_{\mathcal{L}}) = 0 \qquad \textit{where } \mathcal{L} = a \cup b \cup c \textit{ and } \boldsymbol{i}_{\mathcal{L}} \in \mathcal{K}_a \times \mathcal{K}_b \times \mathcal{K}_c, \tag{4}$$

*where $a \cap A \neq \emptyset$ and $b \cap B \neq \emptyset$, $\mathcal{K}_a \subsetneq (\mathcal{I}_a - \mathbf{1}_a)$ and $\mathcal{K}_b \subsetneq (\mathcal{I}_b - \mathbf{1}_b)$.*

The following example shows how to apply formula (4).

**Example 1** In order to test $\boldsymbol{i}'_A \perp\!\!\!\perp \boldsymbol{i}'_B | \boldsymbol{i}'_C$, we need to impose $\eta_{AB}^{ABC}(i'_A, i'_B) + \eta_{ABC}^{ABC}(i'_A, i'_B, i'_C) = 0$. If the constrain is satisfied for $i'_A i'_B \in \mathcal{I}_{ab} - \mathbf{1}$ (all the categories of the variables $A$ and $B$ except the reference category) then the global conditional hypothesis $H_0 : A \perp\!\!\!\perp B | C$ is satisfied.

Before proceeding, two clarifications should be made. First, sub-CSI in formulation (2) differs from CSI in formulation (1) if $\mathcal{K}_{\mathcal{A}}$ and $\mathcal{K}_B$ are proper subsets of $\mathcal{I}_A - \mathbf{1}_A$ and $\mathcal{I}_B - \mathbf{1}_B$. In fact, by construction, the parameter sets associated with the reference cell are worth zero. So if we impose the constraints in formula (4) for all cells except the reference cell, we automatically have that it is also satisfied for the reference cell. The second clarification always concerns the reference categories of the variables. Since the value of the parameters is zero at these categories, it is difficult to discriminate whether the independence statement (2) also involves the reference cell. However, the choice of the reference cell, which by default is defined as the first cell of the contingency table, can be defined as desired without changing the truthfulness of the constraints.

## 2.2 Union-Intersection principle

The Union-Intersection (UI) principle has been proposed by (10) and (11). In nutshell, the UI test states that we can express a (global) null hypothesis $H_0$ as the intersection of $k$ several component hypotheses $H_{0_i}$ and a (global) alternative as the union of the $k$ component alternatives $\overline{H}_{0_i}$:

$$H_0 : \bigcap_{i=1}^{k} H_{0_i} \quad H_1 : \bigcup_{i=1}^{k} \overline{H}_{0_i}. \tag{5}$$

We reject the global null hypothesis if any of the tests on the component hypotheses lead to rejection, and we retain the global null hypothesis if none of the component tests leads to rejection.
Note that to test any $H_{0_i}$ it is possible to choose a different level $\alpha_i$. When $H_0$ holds it is desirable that the rejection probability of the UI test is not greater than a fixed level $\alpha^*$. This is ensured if the levels $\alpha_i$ of the component tests are such that $\sum_{i=1}^{k} \alpha_i = \alpha^*$. Bonferroni's correction is a popular choice to achieve this: $\alpha_i = \frac{\alpha^*}{k}$. Generally, the rejection probability of the UI test is not less than the rejection probability of any component test. Thus, when $H_0$ does not hold, the rejection probability provides the power of the test and, in this case, the global test's power is greater than or equal to that of the component test with the highest power.
In our context, the Union-Intersection principle may offer advantages when the component hypotheses are defined on lower dimensional sub-tables less affected by sparsity than the whole table on which $H_0$ is defined. Reasoning in terms of sub-CSIs, we can divide hypothesis testing for conditional independence as a set of hypotheses about sub-CSIs, such that the unified support of the variables in all individual tests covers the entire support. For greater clarity, consider the following example.

**Example 2** Let us suppose to have a contingency table involving three variables, $X_1$, $X_2$ and $X_3$. Let the suitable $(\mathcal{K}_1 \times \mathcal{K}_2 \times \mathcal{K}_3)$ be dense, while the remaining sub-tables, defined by $(\bar{\mathcal{K}}_1 \times \mathcal{I}_2 \times \mathcal{I}_3)$, $(\mathcal{K}_1 \times \bar{\mathcal{K}}_2 \times \mathcal{I}_3)$, and $(\mathcal{K}_1 \times \mathcal{K}_2 \times \bar{\mathcal{K}}_3)$ are more or less sparse. Here the symbol $\bar{\mathcal{K}}$ denotes the complementary set of $\mathcal{K}$. We can think of $H_0$ as the conditional independence statement $X_1 \perp\!\!\!\perp X_2 | X_3$. The component hypotheses $H_{0_i}$ can be seen as sub-CSI independence statements $\boldsymbol{i}'_1 \perp\!\!\!\perp \boldsymbol{i}'_2 | \boldsymbol{i}'_3$, where the list of admissible values for $\boldsymbol{i}_{\mathcal{L}} = (\boldsymbol{i}'_1, \boldsymbol{i}'_2, \boldsymbol{i}'_3)$ discriminates between the hypothesis. In detail, we have

DRAFT

that $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \mathcal{K}_2 \times \mathcal{K}_3)$ for $H_{0_1}$, $\boldsymbol{i}_{\mathcal{L}} \in (\bar{\mathcal{K}}_1 \times \mathcal{I}_2 \times \mathcal{I}_3)$ for $H_{0_2}$, $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \bar{\mathcal{K}}_2 \times \mathcal{I}_3)$ for $H_{0_3}$ and $\boldsymbol{i}_{\mathcal{L}} \in (\mathcal{K}_1 \times \mathcal{K}_2 \times \bar{\mathcal{K}}_3)$ for $H_{0_4}$.

The truthfulness of the union of all these hypotheses implies the truthfulness of the global one. The idea is to split the global hypothesis into several sub-CSIs. If we retain all of them then the global conditional hypothesis is satisfied. One way to proceed could be as follows. An exploratory survey is carried out in the whole contingency table to see where the table is most sparse. This would identify blocks in the global table: the dense block, the sparse block, and the middle ground tables. One of the component hypotheses $H_{0_i}$ is constructed in the densest sub-table. If in general some of the $H_{0_i}$ on the sparse sub-tables leads to the rejection of the global hypothesis but the $H_{0_i}$ hypothesis on the dense sub-table is instead in favour of $H_0$, this provides interesting information on the sub-CSI between the variables involved in the contingency table.

## 3. Simulation study of the power

In this section, we want to show some preliminary results obtained through simulations as evidence to support the proposed methodology. We considered a simple case where we have 3 variables: $X_1 \in (1, \ldots, 5)$, $X_2 \in (1, \ldots, 5)$ and $X_3 \in (1, \ldots, 5)$. We want to investigate the global hypothesis $H_0 : X_1 \perp\!\!\!\perp X_2 | X_3$ for $\boldsymbol{X} = (X_1, X_2, X_3)$ taking values in $\mathcal{I}_{123} = (1, \ldots, 5)^3$, against the alternative.
In order to perform the Union-Intersection procedure, we divided the contingency table into 4 sub-tables.

- $\mathcal{K}^I = (1, 2) \times (1, 2) \times (1, 2)$, 8 cells;
- $\mathcal{K}^{II} = (3, 4, 5) \times (1, 2, 3, 4, 5) \times (1, 2, 3, 4, 5)$, 75 cells;
- $\mathcal{K}^{III} = (1, 2) \times (3, 4, 5) \times (1, 2, 3, 4, 5)$, 30 cells;
- $\mathcal{K}^{IV} = (1, 2) \times (1, 2) \times (3, 4, 5)$, 12 cells.

Then, we define 4 component hypotheses $H_{0_i}$ to test $H_0$ in the Union-Intersection procedure as sub-CSI. The null component hypotheses are $H_{0_1}$: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^I$, $H_{0_2}$: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{II}$, $H_{0_3}$: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{III}$ and $H_{0_4}$: $i_1 \perp\!\!\!\perp i_2 | i_3$ for $\boldsymbol{i} \in \mathcal{K}^{IV}$. Since the degree of freedom of the $\chi^2$ for the distribution of the likelihood test is not accurate when the sparseness occurs, we simulated $m = 10000$ samples and we evaluated the MC distribution of the $G^2$ statistics of the likelihood ratio test as follows.
First, we build a probability distribution $P_0$ where the independence in $H_0$ holds. We use that distribution to simulate $m = 10000$ frequency tables under $H_0$. In particular, we impose that the sub-table $A$ must be dense with 50 observations on 8 cells; the remaining three sub-tables are sparse with $57 \sim 75/4 * 3$ observations for $B$, $20 \sim 30/3 * 2$ observations for $C$ and $6 \sim 12/2$ observations for $D$.
Then, we evaluate the statistic $G^2$ of the LR test in the full table and in each sub-tables of each sample, obtaining the MC distribution under $H_0$ for $\boldsymbol{G}^2_{\mathcal{I}_{123}}$, $\boldsymbol{G}^2_{\mathcal{K}^I}$, $\boldsymbol{G}^2_{\mathcal{K}^{II}}$, $\boldsymbol{G}^2_{\mathcal{K}^{III}}$, and $\boldsymbol{G}^2_{\mathcal{K}^{IV}}$, respectively.
On the previous distributions, we evaluated the simulated critical values for the tests as the quantile of order $1 - \alpha = 0.95$ ($X^2_{H_0,\alpha}$), for the global hypothesis and as the quantile of order $1 - \alpha_1$, $1 - \alpha_2$, $1 - \alpha_3$, and $1 - \alpha_4$ for the component hypothesis $X^2_{H_{0_1},\alpha_1}$, $X^2_{H_{0_2},\alpha_2}$, $X^2_{H_{0_3},\alpha_3}$, and $X^2_{H_{0_4},\alpha_4}$. The only constraint on the significance level of the component hypotheses is that their sum must be equal to $\alpha$. The analysis is led for different values of significant levels in order to investigate how it is possible to gain more power.
Further, we proceed in the same way with the generation of $m = 10000$ samples from an alternative distribution, where the global hypothesis does not hold.
In order to cover all possible scenarios we build different probability distributions under $H_1$:

- $P_1$ where only the sub-CSI $H_{0_1}$ holds (the dense table);
- $P_2$ where the sub-CSI $H_{0_1}$ does not hold but the others (in the sparse tables) hold.
- $P_3$ where all the sub-CSIs do not hold.

In any of the above scenarios, we evaluated the rejection rate as an estimator of the test power. The results were collected in the following subsection.
All the analyses were carried out with the software R and the package `hmmm` (3) for the estimation of parameters.

DRAFT

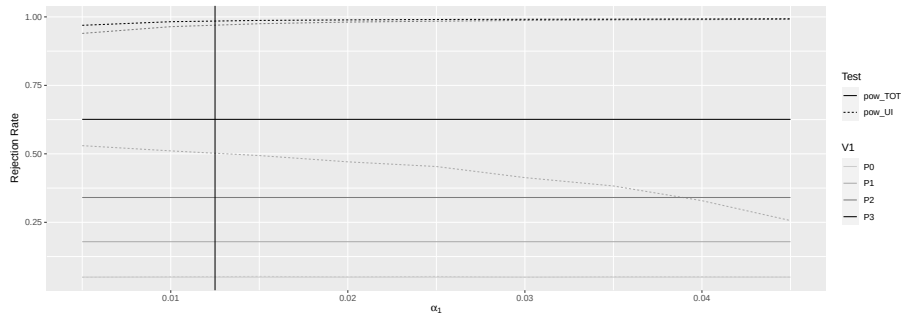



Figure 1: Simulated rejection rate distribution (dashed line) related to the UI procedure by increasing the size $\alpha_1$ of test $H_{0_1}$ using the alternative distribution $P_3$. In solid lines are the corresponding (constant) simulated rejection rates for the global test.

## 3.1 Results description

Table 1 reports the rejection rate in all the scenarios detailed above and also for the case where the alternative distribution satisfies $H_0$, in order to provide information on the significant level of the Union-Intersection test. It can be seen from the results in Table 1 that the test based on the UI procedure is always more powerful than the test based on the global hypothesis. In particular, the power of the test through the UI procedure gains a lot of power when the null hypothesis does not apply in the dense sub-table ($P_2$ and $P_3$). However, due to the segmentation of the entire contingency table, even in the case of $P_1$ the UI procedure outperforms the test conducted on the entire contingency table. Looking at the row referring to the $P_3$ distribution, where in each sub-table independence does not hold, it is evident how the power increases in the dense situation.

Table 1: Rejection rates for all possible scenarios evaluated on the distributions $P_1$, $P_2$, $P_3$ and $P_0$, respectively. The shaded cells show the simulated levels of the tests and non-shaded cells show the simulated power.

| $n$ | $m$ | $r(H_{0_1},\alpha_1)$ | $r(H_{0_2},\alpha_2)$ | $r(H_{0_3},\alpha_3)$ | $r(H_{0_4},\alpha_4)$ | $r(H_0,\boldsymbol{\alpha})_{T_{UI}}$ | $r(H_0,\alpha^*)_{T_0}$ |
|---|---|---|---|---|---|---|---|
| $P_1 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$ | | | | | | $\rightarrow \boldsymbol{H_0}$ ***rejected*** | |
| 133 | 10000 | 0.0133 | 0.3691 | 0.1400 | 0.0657 | 0.4996 | 0.1792 |
| $P_2 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$ | | | | | | $\rightarrow \boldsymbol{H_0}$ ***rejected*** | |
| 133 | 10000 | 0.9698 | 0.0114 | 0.0112 | 0.0164 | 0.9706 | 0.3406 |
| $P_3 : i_1 \not\perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$ and $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$ $\rightarrow \boldsymbol{H_0}$ ***rejected*** | | | | | | | |
| 133 | 10000 | 0.9690 | 0.3689 | 0.1400 | 0.0657 | 0.9844 | 0.6259 |
| $P_0 : i_1 \perp\!\!\!\perp i_2 \vert i_3$ for $\boldsymbol{i} \in \mathcal{K}^{I}$ and $\boldsymbol{i} \in \mathcal{K}^{II}$ and $\boldsymbol{i} \in \mathcal{K}^{III}$ and $\boldsymbol{i} \in \mathcal{K}^{IV}$ | | | | | | $\rightarrow \boldsymbol{H_0}$ ***not rejected*** | |
| 133 | 10000 | 0.0125 | 0.0125 | 0.0125 | 0.0151 | 0.0516 | 0.0500 |

$n$: number of observations; $m$: number of simulated elements in the MC distributions; $r(H,\alpha)$: rejection rate of the component tests $H_{0_1}$, $H_{0_2}$, $H_{0_3}$, $H_{0_4}$, and $H_0$, with test size equal to $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.0125$, $\boldsymbol{\alpha}^T = (0.0125, 0.0125, 0.0125, 0.0125)$, and $\alpha^* = 0.05$.

Further, Figure 1 shows how the power of the UI test changes by varying the significant levels of the component hypothesis and by varying the different degree of connection between the two conditional distribution of $X_1$ and $X_2$ given $X_3$. Note that the scenario reported in Table 1 corresponds to the vertical line in $\alpha = 0.125$. Looking at the solid lines, Figure 1 reports the rejection rates of the different tests. The red line belongs to the global test, the other lines represent the (greater) power of the other tests. The greater power is related to the test evaluated on the $P_3$ distribution where none of the sub-CSIs holds. Concerning the dashed lines, it is worthwhile to note that the power of the tests evaluated on $P_2$, $P_3$ distributions is increasing and always increases the power of the overall test. In contrast, this trend is decreasing with respect to the $P_1$ distribution, where independence does not hold in the densest table and holds in all others. Moreover, for the different levels of $\alpha_1$, the red dashed line traces the corresponding value of the classical test.

DRAFT

## 4. Conclusion and further research

This study proposes a method to increase the power of tests performed in sparse contingency tables. The idea is based on the logic of the Union Intersection procedure to decompose the null hypothesis $H_0$. The original proposal is to consider a set of context-specific hypotheses. This type of hypothesis focuses on sub-spaces of variables. By identifying the least sparse ones, a discrete increase in power can be achieved. The work is still preliminary, but early results from simulations give promising results in terms of power. Clearly, it is necessary to extend the simulations already carried out for different levels of $n$, $m$ and different scenarios among $\alpha_i$, for $i = 1, \ldots, k$.

## References


[1] Agresti, A.: Categorical Data Analysis - Third Edition. Wiley, Hoboken (2013)

[2] Bergsma, W. P.,& Rudas, T.: Marginal models for categorical data. Ann. Stat. **30(1)**, 140-159 (2002).

[3] Colombi, R., Giordano, S., Cazzaro, M.: hmmm: An R Package for Hierarchical Multinomial Marginal Models. J. Stat. Softw. **59**, 1–25 (2014)

[4] Dale, J.R.: Asymptotic normality of goodness-of-fit statistics for sparse product multinomials. J. R. Stat. Soc. Series B Stat. Methodol. **48**, 48–59 (1986)

[5] Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models. Ann. Stat. **40**, 996–1023 (2012)

[6] Koehler, K.J.: Goodness-of-fit tests for log-linear models in sparse contingency tables. J. Am. Stat. Assoc. **81**, 483–493 (1986)

[7] Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in r×c contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)

[8] Nicolussi, F., Cazzaro, M.: Context-specific independencies in hierarchical multinomial marginal models. Stat. Methods Appt. **29**, 767–786 (2020)

[9] Nicolussi, F., Cazzaro, M.: Context-Specific Independencies in Stratified Chain Regression Graphical Models. Bernoulli **27**, 2091–2116 (2021)

[10] Roy, S.N.: On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Stat. **24**, 220–238 (1953)

[11] Roy, S.N., Mitra, S.K.: An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis. Biometrika. **43**, 361–376 (1956)

[12] Rudas, T.: A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. J. Stat. Comput. Simul. **24**, 107–120 (1986)


DRAFT

# Latent event history models for quasi-reaction systems

Framba Matteo[a], Vinciotti Veronica[a], and Wit Ernst[b]

[a]University of Trento; matteo.framba@unitn.it, veronica.vinciotti@unitn.it

[b]Università della Svizzera Italiana; ernst.jan.camiel.wit@usi.ch

**Abstract**

Various processes can be modelled as quasi-reaction systems of stochastic differential equations, such as cell differentiation and disease spreading. Often the underlying interactions, such as reactions between particles or contacts between people, are unobserved, and only the state variables, such as cell type counts and number of infected individuals, are available. Statistical inference of the parameters driving these systems have been developed from concentration data over time. Whereas observing the continuous time process at a time scale as fine as possible should in theory help with parameter estimation, existing Local Linear Approximation (LLA) methods do not work well in these scenarios, due to numerical instability caused by small changes of the system at successive time points.

In this manuscript we propose a method that reconstructs the underlying unobserved interactions from the observed count data. Motivated by this, we first formalize the latent event history model underlying the observed count process. We then propose a penalized Expectation-Maximization algorithm for parameter estimation. The complete log-likelihood has the form of a LLA likelihood with a penalty term corresponding to the latent Poisson event history model. The larger this term the stronger the information about the events driving the count process. A simulation study shows the performance of the proposed method and highlights the settings where it is particularly advantageous compared to the existing LLA approaches.

DRAFT

***Keywords:*** Diffusion process, Euler-Maruyama, Penalized inference, Latent Variable Models

## 1. Introduction

The measurement of many complex temporal phenomena can be described by means of stochastic differential equations (SDEs) (2), as these are able to capture both measurement uncertainty as well as the intrinsic stochasticity in the process. Their dynamics depend critically on parameters, which are often unknown. As such, various techniques have been developed to perform parameter estimation using observational data from these processes (6). Since the Markovian transition density of the underlying SDE has rarely an explicit form (5), Local Linear Approximation (LLA) methods provide an explicit but approximate formulation of the likelihood function under various assumptions (9). In stochastic kinetic systems of interacting particles, however, parameter estimation using these methods worsens if the inter-observations times are too close, as the strong correlation of the observations from one time point to the next causes computational instability. On the other hand, the state of the system at each time point can be seen as a function of an underlying counting process of the occurred events, which can be approximated by an independent non-homogeneous Poisson process (10).

Motivated by this, we propose a penalized Expectation-Maximization (EM) algorithm for parameter estimation that combines the observed counts of the state over time with a latent event history model describing the occurrence of events. A penalty term has the effect of adjusting the weight of the observed and latent components of the likelihood according to the time distance of the observations. The structure of the paper is as follows: in Section 2, we formalize the statistical modelling of quasi-reaction systems, while in Section 3, we describe the proposed penalized EM-algorithm for parameter estimation. A simulation study is shown in Section 4.

## 2. Probability models of dynamic processes

In this section we describe the two stochastic models that describe the interactions between the particles, and the state of the particles, respectively. They provide two dual versions of same process.

### 2.1 Latent history event models

Consider a closed system in which $p$ substrates interact, with concentration levels governed by a system of reactions. Let $\mathcal{J}$ be the set of all reactions. For each time unit, consider the following multivariate counting process

$$N_j(t) = \#\{\text{Reactions of type } j \text{ occurred in time interval } [0,t]\}.$$

We assume that $N_j(0) = 0$ and that the process is adapted to the stochastic basis of a $\sigma-$algebra $\{\mathcal{F}_t\}_{t\geq 0}$ which is complete and right continuous. $N_j(t)$ is therefore a local sub martingale. Due to the Doob-Meyer decomposition, there exists a predictable increasing process $\Lambda_j(t)$ such that $\Lambda_j(0) = 0$ and $N_j(t) - \Lambda_j(t)$ is an $\mathcal{F}_t$-local martingale. Assuming that reactions cannot occur simultaneously, i.e.,

$$\Pr\big([N_j(t) - N_j(s)] \geq 1 | \mathcal{F}_s\big) = 1 - e^{-[\Lambda_j(t) - \Lambda_j(s)]},$$

one can uniquely identify the reaction that occurred at a certain instant in time. We call this reaction $r_j(t)$ and the pair $e_j(t) := (t, r_j(t))$ an *event*. If $\Lambda_j(t)$ is differentiable, it is possible to define a new predictable continuous process, $\lambda_j(t) : \mathbb{R}^p \to [0,\infty)$ such that $\frac{d}{dt}\Lambda_j(t) = \lambda_j(t)$, which is the instantaneous hazard rate of the $j$th reaction. Roughly speaking, $\mu_j(t) = \lambda_j(t)dt$ is the probability that the $j$-reaction occurs in the interval $[t, t+dt)$. As there is no change in the substrates concentration until a new reaction takes place, the hazard function is constant for each time interval. For all $s,t \in \mathbb{R}^+$, the waiting time for the $j$-reaction to occur in $[s,t)$ is distributed as an exponential random variable with cumulative density function

$$\Pr\big([N_j(t) - N_j(s)] = 1 | \mathcal{F}_s\big) = 1 - e^{-\lambda_j(t)(t-s)}.$$

From this, in a framework with discrete time steps, we conclude that the distribution of the number of events in an interval $[t_{i-1}, t_i]$ is governed by the following non-homogeneous Poisson process:

$$N_j(t_i) - N_j(t_{i-1}) | \mathcal{F}_{t_{i-1}} \sim \text{Po}(\lambda_j(t_i)(t_i - t_{i-1})), \tag{1}$$

where we assume that the hazard function remains constant on the interval $[t_{i-1}, t_i]$.

### 2.2 Quasi-reaction models and local linear approximation algorithm

Let $(Y_1, \ldots, Y_p)$ define the state of a reaction system. With some abuse of notation, we consider a set $\mathcal{J}$ of reactions, in which the $j$-th reaction is described as:

$$k_{1j}Y_1 + \ldots + k_{pj}Y_p \xrightarrow{\theta_j} s_{1j}Y_1 + \ldots + s_{pj}Y_p, \tag{2}$$

where $k_{ij}, s_{ij} \in \mathbb{Z}$ are the stoichiometric coefficients, $\theta_j \in \mathbb{R}^+$ the reaction rate, for $i = 1, \ldots, p$, and $j = 1, \ldots, |\mathcal{J}| = r$. The net change of $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$ is defined by the stoichiometric matrix $V$ with $(i,j)$ element $v_{ij} = s_{ij} - k_{ij}$.

DRAFT

Considering now observations over time, and assuming that $\boldsymbol{Y}_t = (Y_{1t}, \ldots, Y_{pt}) \in \mathbb{N}_0^p$ is a continuous time counting Markov process, the hazard function described before can now be related to model (2) by

$$\lambda_j(t) = \lambda_j(\boldsymbol{Y}_t; \theta_j) = \theta_j \prod_{i=1}^{p} \binom{Y_{it-1}}{k_{ij}}, \tag{3}$$

where $\binom{Y_{it-1}}{k_{ij}} = 0, \forall Y_{it-1} < k_{ij}$. As the distribution of $\boldsymbol{Y}_t$ satisfies the stochastic differential "chemical master equations" (8), which are tractable only in few cases, discrete-event simulation algorithms, in particular the Gillespie algorithm (4), are typically used to generate samples from this distribution. Alternatively, an Euler-Maruyama approximation can be used, based on the first two moments of the jump process, given by

$$\frac{d}{dt}\mathbb{E}[\mathbf{Y}_t|t] = V\boldsymbol{\lambda}(\boldsymbol{Y_t}; \theta), \quad \frac{d}{dt}\text{Var}[\mathbf{Y}_t|t] = V\text{Diag}(\boldsymbol{\lambda}(\boldsymbol{Y_t}; \theta))V^T,$$

respectively (7), where $\boldsymbol{\lambda}(\boldsymbol{Y_t}; \theta)$ is the vector of hazard rates defined in (3). Assuming a Gaussian distribution for the jumps and constant rates on discretized time intervals, a simulation algorithm as well as a parameter estimation method can be devised based on the assumption that

$$\left(\boldsymbol{Y}_{t_i} - \boldsymbol{Y}_{t_{i-1}}\right)\Big|\boldsymbol{Y}_{t_{i-1}} \sim \mathcal{N}\Big(V\boldsymbol{\lambda}(\boldsymbol{Y}_{t_{i-1}}; \boldsymbol{\theta})(t_i - t_{i-1}), V\text{Diag}(\boldsymbol{\lambda}(\boldsymbol{Y}_{t_{i-1}}; \boldsymbol{\theta}))V^T(t_i - t_{i-1})\Big), \tag{4}$$

on the interval $[t_{i-1}, t_i]$.

## 2.3 Exogenous drivers of the reaction system

An intrinsic feature of stochastic reaction systems are their emergent behaviour, i.e., the occurrence of a particular reaction changes the state, which in turn changes the rates of the reactions. However, besides endogenous drivers, many dynamic systems are also subject to external forces. For example, the cell differentiation may depend on the particular characteristics of the patient. For this reason, we extend the traditional modelling frameworks with the inclusion of external covariates, which can affect the transition probabilities of the Markov process (1). In particular, we will model the direct effect of covariates on the reaction rates $\theta_j$ by

$$\theta_j = \exp(\mathbf{x}\boldsymbol{\beta_j}),$$

with $\mathbf{x} \in \mathbb{R}^q$ a vector of $q$ covariates, which are subject specific but do not vary over time. The aim of this manuscript is to estimate the parameters $\boldsymbol{\beta} \in \mathbb{R}^{qr}$ describing all the reaction rates.

DRAFT

# 3. Penalised EM for parameter estimation

Assume that $N + 1$ observations are collected per subject, at $N$ not necessarily equispaced time intervals. Using the modelling framework of Section 2, the complete *penalized* log-likelihood of the latent reaction events $e$ and the observed states $y$ for one subject is given by

$$l_{e,y}(\boldsymbol{\beta}; \alpha, \gamma) = l_y(\boldsymbol{\beta}) + \gamma l_{e|y}(\boldsymbol{\beta}) - \alpha\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{5}$$

where

$$l_{e|y}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j\in\mathcal{J}}\left\{-\lambda_j(\boldsymbol{y}_{t_i}; \mathbf{x}, \boldsymbol{\beta})(t_i - t_{i-1}) + e_{ij}\log\Big(\lambda_j(\boldsymbol{y}_{t_i}; \mathbf{x}, \boldsymbol{\beta})(t_i - t_{i-1})\Big)\right\}$$

is the contribution of the Poisson likelihood from the latent model (1) and

$$l_y(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{N}\Big\{\log(|g_{t_i}(\boldsymbol{\beta})|) + \big[\Delta y_i - f_{t_i}(\boldsymbol{\beta})\big]^T g_{t_i}(\boldsymbol{\beta})^{-1}\big[\Delta y_i - f_{t_i}(\boldsymbol{\beta})\big]\Big\} \tag{6}$$

is the contribution from the Gaussian likelihood from model (4), with $f_t(\boldsymbol{\lambda}(\mathbf{y}_t; \mathbf{x}, \boldsymbol{\beta}))$ and $g_t(\boldsymbol{\lambda}(\mathbf{y}_t; \mathbf{x}, \boldsymbol{\beta}))$ denoting the mean and variance of the process. The full log-likelihood from multiple processes (e.g. several patients) will simply be the sum of all the log-likelihoods.

Notice how some penalty terms are introduced to weigh the different components of the likelihood. In particular, increasing $\gamma \in \mathbb{R}^+$ puts more weight on the latent event information, while $\gamma = 0$ corresponds to the likelihood used by the LLA existing method. In general, we would expect that the finer the time scale, the smaller the value of $\gamma$ that should be set. An additional tuning parameter $\alpha \in \mathbb{R}^+$ induces a ridge penalty term which should improve the estimation accuracy in the presence of strongly correlated responses across time, as it is often the case at finer time scales.

The complete log-likelihood in (5) depends both on the observed concentration data and on the hidden reactions. Its maximum can be found efficiently via an EM algorithm (3). The procedure iterates the following two steps:

- **E-step**: calculate the conditional expected value of the complete log-likelihood based on the current estimate $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$. This value will be denoted by $\mu^*_{ij}(t) = \lambda^*_{ij}(t)dt$ and is obtained by evaluating Equation (3) using $\boldsymbol{\beta}^*$.
- **M-step**: Fix $\mu^*_{ij}$ from the E-step and find the optimal $\boldsymbol{\beta}$ by maximising the Q-function:

$$\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} l_y(\boldsymbol{\beta}) + \gamma \sum_{i=1}^{N} \sum_{j=1}^{r} \Big[ -\boldsymbol{\mu}_{ij}(\boldsymbol{\beta}) + \boldsymbol{\mu}^*_{ij} \log\left(\boldsymbol{\mu}_{ij}(\boldsymbol{\beta})\right) \Big] - \alpha \boldsymbol{\beta}^T \boldsymbol{\beta}.$$

For the optimisation, we make use of the Broyden-Fletcher-Goldfarb-Shanno algorithm, which is an approximation of the Newton's method. The iterative process continues until no significant change is obtained between successive estimations of the parameters. The initial parameter values are sampled from a normal distribution with zero mean and unitary variance and the first estimation is conducted assuming no covariance structure.

DRAFT

## 4. Simulation Study

Inspired by (7), we set up the simulation study in the context of cell dynamics. Here, the substrates correspond to cell types and the interactions between particles to chemical reactions. Figure 1 (left) shows the specific process that we consider, involving 2 cell duplication (DP), 2 cell death (DT), and 2 cell differentiation (DF) reactions. The rates $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^T$ depend on 2 covariates, which vary across 4 subjects by drawing their values from a Uniform[0,1] distribution. Together with an intercept, we set the parameters $\boldsymbol{\beta} \in \mathbb{R}^{18}$ to

$$\begin{aligned} \boldsymbol{\beta}_{DP} &= (1.70, 1.02, 1.17; 1.20, 1.03, 1.24), \quad \boldsymbol{\beta}_{DT} = (-1.20, 0.03, 0.13; -2.40, 0.65, 1.80) \\ \boldsymbol{\beta}_{DF} &= (-1.89, 1.07, 0.34; -4.32, 1.26, 2.10) \end{aligned}$$

We set the initial concentration to $\mathbf{y}_0 = (200, 100, 200)$. From this, a Gillespie algorithm is used to generate the state of the process over time. Since we want to evaluate how the method performs when data are collected at a finer versus a coarser time scale, we consider the following two scenarios:

- **Scenario A**: For each subject, the process is simulated for 100 event times, but only every other observation is retained, for a total of $N$=49 time intervals. The LLA approach is expected to lead to inaccurate parametric estimates here due to the strong correlation of the variables from one time point to the next.

- **Scenario B**: For each subject, the process is simulated for 1500 event times, but only every other 30th observation is retained, for a total of $N$=49 time intervals. Observations are here much more distant in time and reconstruction of the latent event history process is more challenging.

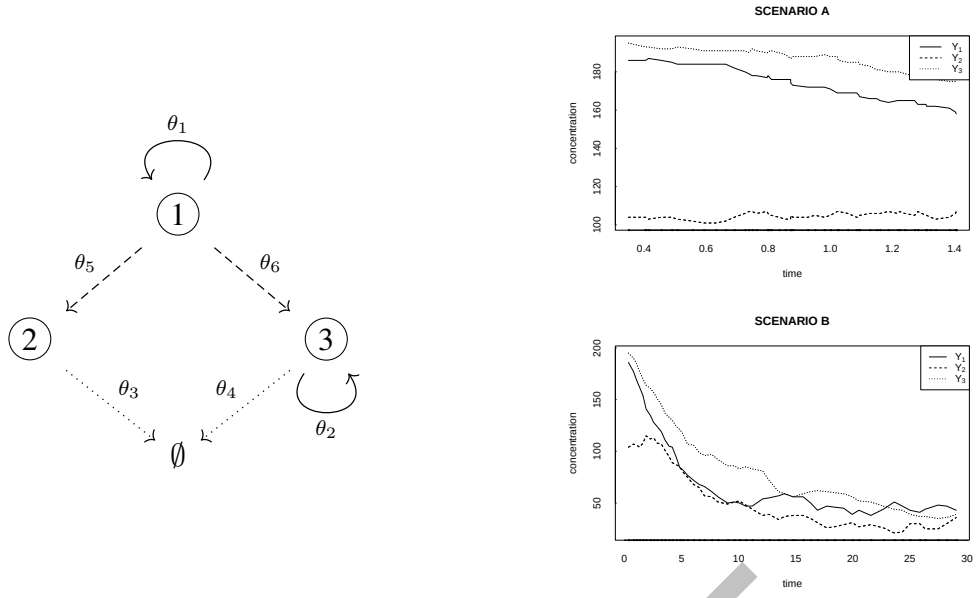



***Figure 1:*** *Left: the stochastic cell differentiation process used in the simulation. Each of the 3 substrates is represented by a node, whereas birth, death and differentiation events are denoted by full, dotted and dashed edges, respectively. Right: the trajectories of the process for one subject, generated via a Gillespie algorithm in scenario A (top) and B (bottom).*

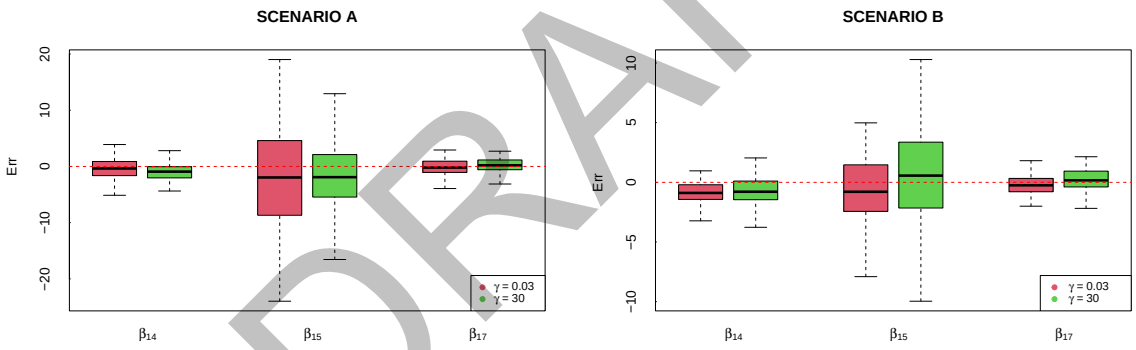



***Figure 2:*** *Distribution of the errors for scenario A (left) and scenario B (right) and for a selection of the $\boldsymbol{\beta}$ parameters. The accuracy of the parameters is shown when using a small (red) and a large (green) value of $\gamma$, respectively.*

We repeat each scenario 100 times and report the average accuracy of the estimators by calculating $\sum_{i=1}^{100} \frac{\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i}{\boldsymbol{\beta}_i}$. For parameter estimation, the tolerance in the EM algorithm is set to $0.01$. The two plots in Figure 2 show the distribution of the accuracy in the two scenarios, respectively, when using a small (red) and a large (green) value of $\gamma$, respectively. Moreover, we set the ridge tuning parameter in scenario A to $\alpha = 2 \times 10^{-5}$ and in scenario B to $\alpha = 2 \times 10^{-4}$. The results show how a high value of the tuning parameter $\gamma$ leads to better estimates of $\boldsymbol{\beta}$ in the first scenario, i.e., the Poisson component of the likelihood is particularly beneficial when observations are close in time, while a low value of $\gamma$, i.e., an estimation procedure close to the existing LLA methods, works well in the second scenario.

DRAFT

## 5. Conclusions

In this paper, we have proposed an innovative algorithm for the estimation of stochastic rate constants governing kinetic diffusion processes. Noting that LLA methods perform poorly when the system is observed at fine time intervals, due to numerical instability caused by the strong correlations in the observations from one time point to the next, our proposed method focusses on reconstructing the underlying process of hidden events. Using an Expectation-Maximation algorithm, we have shown, by means of simulation studies, how the new approach leads to more accurate estimates precisely in these settings.

## References


[1] Bartolucci, F., A. Farcomeni, and F. Pennoni. "Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates." Test 23 (2014): 433-465.

[2] Craigmile, P. et al. "Statistical inference for stochastic differential equations." Wiley Interdisciplinary Reviews: Computational Statistics (2022): e1585.

[3] Dempster, A., N. Laird, and D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." JRSS B 39.1 (1977): 1-22.

[4] Gillespie, D. "Exact stochastic simulation of coupled chemical reactions." The Journal of Physical Chemistry 81.25 (1977): 2340-2361.

[5] Golightly, A. and D. Wilkinson. "Bayesian inference for stochastic kinetic models using a diffusion approximation." Biometrics 61.3 (2005): 781-788.

[6] Kitano, H. "Foundations of systems biology." The MIT Press Cambridge, Massachusetts London, England, 2001.

[7] Pellin, D. et al. "Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking." Applied Network Science 4.1 (2019): 1-26.

[8] Schnakenberg, Jürgen. "Network theory of microscopic and macroscopic behavior of master equation systems." Reviews of Modern physics 48.4 (1976): 571.

[9] Shoji, Isao, and Tohru Ozaki. "Estimation for nonlinear stochastic differential equations by a local linearization method." Stochastic Analysis and Applications 16.4 (1998): 733-752.

[10] Wilkinson, Darren J. Stochastic modelling for systems biology. Chapman and Hall/CRC, 2018.


DRAFT

# Quantile-based graphical models for continuous and discrete variables

Luca Merlo[a], Marco Geraci[b], and Lea Petrella[b]

[a]Department of Human Sciences, European University of Rome, Italy; luca.merlo@unier.it
[b]MEMOTEF Department, Sapienza University of Rome, Italy; marco.geraci@uniroma1.it, lea.petrella@uniroma1.it

**Abstract**

In this paper we develop a mixed graphical model for identifying conditional independence relations between continuous and discrete variables in a quantile framework using Parzen's definition of mid-quantile. To recover the graph structure and induce sparsity, we consider the neighborhood selection approach in which conditional mid-quantiles of each variable in the network are modeled as a sparse function of all others. Building on previous work, we propose a two-step estimation procedure where, in the first step, conditional mid-probabilities are obtained and, in the second step, the model parameters are estimated by solving an implicit equation with a LASSO penalty. The empirical application investigates the relationship between depression and inflammation on a sample of individuals from the National Health and Nutrition Examination Survey 2017-2020.

***Keywords:*** LASSO, mixed random variables, mid-CDF, neighborhood selection, NHANES

DRAFT

## 1. Introduction

Graphical models have become a popular and effective framework for the statistical analysis of complex dependence relations among variables. Within this literature, Gaussian Graphical Models (GGMs, 11) have received considerable attention as they provide a model for the pair-wise conditional correlation structure of the variables of interest. Under the assumption of normality, the underlying conditional dependence structure is completely characterized by the inverse of the covariance matrix of the corresponding GGM. However, GGMs suffer from two limitations. First, they rely on the assumption of normally distributed data. Despite its simplicity and mathematical tractability, this assumption is hardly met in actual applications, and deviation from normality makes it harder to characterize conditional dependence structures. To overcome this issue, semi-parametric Gaussian copula models (15; 22) or power transformations of the data may be considered. Alternatively, one may forgo the normal distribution and consider more robust alternatives such as the multivariate t-distribution (7). Such proposals, however, have mainly relied on the use of symmetric distributions or on, more in general, location-shift models. In contrast, a quantile-based approach (1; 5) allows one to infer the conditional dependence structure without having to introduce assumptions on the form of the distributions.

Another, drawback of the GGMs is that they are confined to the modeling of continuous variables only. In many applications of practical relevance, however, the dataset of interest consists of mixed variables (categorical, counts, and continuous). Unfortunately, the literature regarding graphical models in which the variables are of different types is fairly limited. In parallel efforts, (23) and (3) introduced the class of Mixed Graphical Models (MGMs), which specify the conditional distribution of each variable

(continuous and discrete) given the rest as a member of the exponential family of distributions. Subsequently, in a related line of research, (14) and (4) proposed a generalization of the conditional Gaussian model of (12) for mixed data.

The aim of this paper is to introduce a quantile-based graphical model for mixed variables that tackles conditional dependency structures, without making assumptions on the functional form of the distributions. We start from the work of (8) who developed a quantile regression method for discrete responses by extending Parzen's definition of marginal mid-quantiles (19). Intuitively, mid-quantiles can be viewed as fractional order statistics and have been extensively studied by (16). In this context, using mid-quantiles comes with desirable advantages as opposed to existing approaches, based on either jittering or latent constructs. Most importantly, they offer a unifying theory for quantile estimation with discrete or continuous variables, and are well-behaved asymptotically.

In our approach, to identify conditional independence relations and induce sparsity in the network, we model the conditional mid-quantiles of each variable as a sparse function of all others and fit separate regularized regressions using the neighborhood selection method of (17). For each variable, the parameters are estimated via a two-step procedure where conditional mid-probabilities are first obtained semi-parametrically and then regression coefficients are estimated by solving a LASSO-penalized implicit equation. The proposed method allows us to embed in a common graphical framework both continuous (possibly, e.g., heavy-tailed, skewed, multimodals) and discrete (e.g., binary, ordinal, count) variables, thus offering a much richer class of conditional distribution estimates than the conditional mean.

The relevance of this methodology is shown using observations from adult participants of the National Health and Nutrition Examination Survey (NHANES) 2017-2020 to investigate the association between C-Reactive Protein (CRP) and depression symptoms.

The rest of this paper is organized as follows. Sect. 2. formally describes the proposed model while the estimation procedure is discussed in Sect. 3. Finally, the empirical application is presented in Sect. 4.

DRAFT

## 2. Methods

In this section we illustrate the proposed mid-quantile mixed graphical model. In order to introduce our methods, we extend the mid-quantile regression of (8) to the graphical modeling framework with both continuous and discrete variables. Subsequently, using the neighborhood selection approach of (17), we show how to estimate a sparse mixed graphical model characterizing conditional independence relations among variables via node-wise penalized mid-quantile regressions.

Let $\mathbf{Y} = (X_1, \dots, X_{p_1}, Z_1, \dots, Z_{p_2})'$ denote a $p$-dimensional random vector, where $X_1, \dots, X_{p_1}$ are $p_1$ continuous variables and $Z_1, \dots, Z_{p_2}$ are $p_2$ discrete variables. Also, let $\mathcal{G} = (V, E)$ denote an undirected graph where $V = \{1, \dots, p\}$ is the set of nodes such that each component of the random variable $\mathbf{Y}$ corresponds to a node in $V$, and $E \subseteq V \times V$ represents the set of undirected edges.

Following (8), we first introduce the conditional mid-cumulative distribution function (mid-CDF, 19; 20) of $Y_j$ given all other variables as

$$G_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j}) = F_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j}) - 0.5 m_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j}), \tag{1}$$

where $\mathbf{Y}_{\neg j}$ denotes all variables except $Y_j$, $F_{Y_j|\mathbf{Y}_{\neg j}}(\cdot \mid \cdot)$ is the conditional CDF of $Y_j$ and $m_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j}) = \Pr(Y_j = y_j \mid \mathbf{Y}_{\neg j} = \mathbf{y}_{\neg j})$. The definition of conditional mid-CDF in eq. (1) applies to both continuous and discrete variables. Indeed, if $Y_j$ is discrete, $G_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j})$ is a step function while it reduces to $F_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j})$ if $Y_j$ is continuous since $\Pr(Y_j = y_j \mid \mathbf{Y}_{\neg j} = \mathbf{y}_{\neg j}) = 0$.

Let $\mathcal{S}_{Y_j}$ be the set of $s$ distinct values in the population that the random variable $Y_j$ can take on. Then, the conditional mid-quantile function (mid-QF) of $Y_j$, $H_{Y_j|\mathbf{Y}_{\neg j}}(\tau)$, is defined as the piecewise linear function connecting the values $G^{-1}_{Y_j|\mathbf{Y}_{\neg j}}(\pi_{jh} \mid \mathbf{y}_{\neg j})$, where $\pi_{jh} = G_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j})$, $h = 1, \dots, s$, for a given quantile level $\tau \in (0, 1)$. We model the $\tau$-th conditional mid-quantile of $Y_j$ given all the other variables using the following mid-quantile regression model:

$$H_{g_j(Y_j)|\mathbf{Y}_{\neg j}}(\tau) = \beta_j^0(\tau) + \mathbf{y}'_{\neg j}\boldsymbol{\beta}_j(\tau), \quad j = 1, \dots, p, \tag{2}$$

where $g_j(\cdot)$ is a known monotone and differentiable "link" function, and $\boldsymbol{\beta}_j(\tau) = (\beta_j^1(\tau), \ldots, \beta_j^{p-1}(\tau))'$ is a vector of $p-1$ unknown regression coefficients, with $\beta_j^0(\tau)$ being an intercept term, for a given $\tau$.

To study conditional independence relations between the components of $\mathbf{Y}$ through the graph $\mathcal{G}$, we establish a result that allows us to make inference on the edge structure $E$ using mid-quantile regressions. Following (1) and (5), the next proposition characterizes the relationship between the conditional mid-quantile function in eq. (2) and the conditional independence between any pair of variables in $\mathbf{Y}$ given the rest.

**Proposition 1.** *Suppose that the conditional mid-QF of a random variable $Y_j$, for some $j = 1, \ldots, p$, is defined by the mid-quantile regression model in eq. (2). Then, $Y_j$ is conditionally independent from $Y_k$, with $k = 1, \ldots, p$ and $k \neq j$, given all of the other variables if and only if $\beta_j^k(\tau) = 0$ for all $\tau \in (0, 1)$.*

The proof of Proposition 1 follows from the relationship between the conditional mid-quantile and CDF of each node given the others. Most importantly, from Proposition 1 follows that the zero elements of the vector $\beta_j^k(\tau)$ for all $\tau \in (0, 1)$ identify conditional independence relations between the components of $\mathbf{Y}$. Hence, the edge set $E$ of the graph $\mathcal{G}$ is completely determined by the non-zero components in the regression vector $\boldsymbol{\beta}_j(\tau)$, that is, $(j, k) \in E$ if and only if $\beta_j^k(\tau) \neq 0$. Based on this result, we can build a mixed quantile graphical model to characterize conditional independence relationships between the elements of $\mathbf{Y}$ by inferring the sparsity pattern of $\boldsymbol{\beta}_j(\tau)$.

We exploit the neighborhood selection approach of (17) by running separate mid-quantile regressions of each component in $\mathbf{Y}$ on all the others. Specifically, let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)$ be a grid of $L$ ordered quantile levels with $\tau_l \in (0, 1)$, $l = 1, \ldots, L$. Large values of $L$ allow us to investigate conditional independence more accurately, but they also increase the computational cost of estimating the model. To infer the graph structure, we consider the linear model in eq. (2) for the conditional mid-QF, $H_{g_j(Y_j)|\mathbf{Y}_{\neg j}}(\tau_l)$, over all variables $j = 1, \ldots, p$ and levels $l = 1, \ldots, L$. Consequently, the corresponding edge set $E$ of conditional dependencies is defined as

$$E = \Big\{(j,k) : \max_{l=1,\ldots,L} \{\max\{\mid \beta_j^k(\tau_l) \mid, \mid \beta_k^j(\tau_l) \mid\}\} > 0, \quad \text{for} \quad 1 \leq j \neq k \leq p\Big\}. \tag{3}$$

In the next section, we describe a procedure to estimate the proposed graphical model $\mathcal{G}$ and induce sparsity in the regression coefficients.

DRAFT

## 3. Estimation

Consider a sample $\mathbf{Y}_i, i = 1, \ldots, n$, with corresponding observations $\mathbf{y}_i$. For each variable $Y_j$, $j = 1, \ldots, p$ and level $\tau_l$, $l = 1, \ldots, L$, estimation of the model in eq. (2), and in turn, of the set $E$ in eq. (3), proceeds in two steps.

Let $z_{jh}$, $h = 1, \ldots, k$, be the $h$-th distinct observation of $Y_j$ that occurs in the sample, with $z_{jh} < z_{jh+1}$ for all $h = 1, \ldots, k-1$. In the first step we estimate the mid-CDF in eq. (1), $\widehat{G}_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j})$, where $\widehat{F}_{Y_j|\mathbf{Y}_{\neg j}}$ is obtained by fitting $h$ separate logistic regressions, one for each value of $z_{jh}$, $h = 1, \ldots, k$. In the second step, we define $\widehat{G}^c_{Y_j|\mathbf{Y}_{\neg j}}(y_j \mid \mathbf{y}_{\neg j})$ as the function interpolating the points $(z_{jh}, \widehat{G}_{Y_j|\mathbf{Y}_{\neg j}}(z_{jh} \mid \mathbf{y}_{\neg j}))$, where the ordinates have been obtained in the first step. The goal now is to estimate $(\beta^0(\tau_l), \boldsymbol{\beta}_j(\tau_l))$ in eq. (2) by solving the implicit equation $\tau_l = \widehat{G}^c_{Y_j|\mathbf{Y}_{\neg j}}(\eta(\tau_l) \mid \mathbf{y}_{\neg j})$, where $\eta(\tau_l) = g_j^{-1}\{\beta^0(\tau_l) + \mathbf{y}'_{\neg j}\boldsymbol{\beta}_j(\tau_l)\}$, and, at the same time, capture the most relevant interconnections between the variables, which motivates us to use a sparse estimator that automatically shrink the elements of $\boldsymbol{\beta}_j(\tau_l)$. Following (8), we thus obtain an estimate of $\boldsymbol{\beta}_j(\tau_l)$, denoted $\widehat{\boldsymbol{\beta}}_j(\tau_l)$, by minimizing the following objective function

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \Big(\tau_l - \widehat{G}^c_{Y_j|\mathbf{Y}_{\neg j}}(\eta_i \mid \mathbf{y}_{\neg j})\Big)^2 + \lambda \mid\mid \mathrm{diag}(\mathbf{w})\boldsymbol{\beta}_j(\tau_l) \mid\mid_1, \tag{4}$$

where we consider the linear interpolation function

$$\widehat{G}^{c}_{Y_j|\mathbf{Y}_{\neg j}}(\eta_i \mid \mathbf{y}_{\neg j}) = b_{h_i}(\eta_i - z_{jh_i}) + \widehat{\pi}_{jh_i}, \quad z_{jh_i} \leq \eta_i \leq z_{jh_i+1}, \tag{5}$$

with $b_{h_i} = \frac{\widehat{\pi}_{jh_i+1} - \widehat{\pi}_{jh_i}}{z_{jh_i+1} - z_{jh_i}}$ and $\widehat{\pi}_{jh_i} = \widehat{G}_{Y_j|\mathbf{Y}_{\neg j}}(z_{jh_i} \mid \mathbf{y}_{\neg j})$. The penalization in eq. (4) is given by a Lasso-type penalty on $\boldsymbol{\beta}_j(\tau_l)$ where we allow a different weight for each coefficient by using the vector $\mathbf{w}$ to avoid that variables of different types are on different scales, and where $\lambda \geq 0$ is the overall tuning parameter of the model. The parameter $\lambda$ controls the strength of the penalization and determines the sparsity of the graph: a higher (lower) value is responsible for a lower (higher) number of edges; when $\lambda = 0$, $\widehat{\boldsymbol{\beta}}_j(\tau_l)$ reduces to the closed-form estimator in (8, see eq. 2.9). Finally, to infer the graph structure we solve the minimization problem in eq. (4) for all $Y_j$, $j = 1, \ldots, p$ and $\tau_l$, $l = 1, \ldots, L$, and estimate the edge set $E$ as follows:

$$\widehat{E} = \Big\{(j,k) : \max_{l=1,\ldots,L} \{\max\{|\, \widehat{\beta^k_j}(\tau_l)\,|, |\, \widehat{\beta^j_k}(\tau_l)\,|\}\} > 0, \quad \text{for} \quad 1 \leq j \neq k \leq p\Big\}. \tag{6}$$

To select the optimal value of the penalty parameter $\lambda$, we adopt the following BIC-type criteria (5):

$$\text{BIC}(\lambda) = \sum_{l=1}^{L}\sum_{j=1}^{p}\left[\ln\left(\sum_{i=1}^{n}\rho_\tau(y_{ij} - \beta^0_j(\tau_l) - \mathbf{y}'_{i\neg j}\boldsymbol{\beta}_j(\tau_l))\right) + \frac{\ln n \ln(p-1)}{2n}\nu_{jl}\right], \tag{7}$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile loss function (9), with $I(\cdot)$ being the indicator function, and $\nu_{jl}$ is the number of estimated non-zero components in $\widehat{\boldsymbol{\beta}}_j(\tau_l)$ for node $j$ at quantile level $\tau_l$. Specifically, we fit the model for a grid of candidate values of $\lambda$ and then select the optimal tuning parameter as that corresponding to the lowest BIC value in eq. (7).

DRAFT

## 4. Application

To evaluate the performance of the proposed methods, we illustrate an application to depression symptoms and inflammatory proteins from the NHANES 2017-2020. There is mounting evidence that inflammatory proteins adversely affect functional ability, quality of life, and well-being of individuals (18; 13). Among these proteins, C-Reactive Protein (CRP) is arguably the most extensively studied inflammatory index in depression research. CRP is a protein synthesized by the liver during the acute phase of an inflammatory/infectious process in response to stimulation from other pro-inflammatory proteins (e.g., elevated cytokine levels (6)). Research suggests that the presence, size, and direction of the association between CRP level and depression vary. One possible explanation for this might be the heterogeneity in the population due to, e.g., age and race/ethnicity. Another explanation may be that CRP is also associated with numerous factors (confounders), such as socio-demographic variables and the overall health status of the individual. As potential reasons for these inconsistencies, (21) also pointed out differences in population settings (e.g., inpatient, outpatient, or community), depression assessment (e.g., sum-scores or diagnoses), or adjustment for important chronic conditions.

Graphical models represent the ideal tool to help disentangling the intricate dependencies between CRP, depression symptoms and individual characteristics. Following (8), before carrying out the analysis we remove the effect of NHANES oversampling and then restrict the dataset to individuals aged 20-70 years. The final sample size for analysis is $n = 3690$, composed of about 59.4% of white persons and 50.1% females. In this network, we include the concentration of CRP (mg/L), nine depression symptoms measured via the Patient Health Questionnaire-9 (PHQ-9, 10) scored on a 4-point Likert scale and 17 socio-demographic, clinical, and lifestyle variables collected among the survey participants, resulting in a total of $p = 17$ nodes. Specifically, the PHQ-9 is a nine-item self-report questionnaire that was administered to assess the frequency of nine major depression criteria listed in the Diagnostic and Statistical Manual of Mental Disorders (2). The questionnaire is a well-established, validated tool that evaluates how often individuals had been bothered by any of the nine items in the previous 2 weeks, on a scale ranging from 0 ("not at all") to 3 ("nearly every day").

We fit the proposed model using a grid of $L = 7$ quantile levels, $\boldsymbol{\tau} = (1/8, \ldots, 7/8)$, across an equispaced sequence of 100 values of the tuning parameter $\lambda$ on the log scale from $0.001$ to $5$. Prior to fitting, continuous variables have been centered around zero and divided by their standard deviation. For continuous and count variables, we take $g(\cdot)$ to be, respectively, the identity and the logarithmic function, while we use the logistic mid-quantile model for binary nodes. Finally, the edge set $\widehat{E}$ is estimated as described in eq. (6). To reduce model uncertainty and improve reliability of the inferred interactions, we adopt a model averaging approach. Specifically, 500 bootstrap datasets are created by resampling from the original one. Then, we fit the proposed model on the 500 bootstrap re-samples and estimate the edge structure for each bootstrap dataset. Eventually, in the final network we retain only those edges that are present in at least 85% (18) of the learned graphs.

Fig. 1 provides a graphical representation of the estimated network, where the width of the edges is proportional to the absolute value of the strength of the interaction and the edge colors specify the sign of the corresponding interaction (green = positive, red = negative, grey = undefined). The colors of the nodes map to the three different domains, Inflammation Marker, Depression Criteria, and Covariates.

Results indicate that CRP is associated with greater changes in appetite, presenting a non-zero edge in 90% of bootstrap re-samples. CRP also shows noteworthy connections with other variables: it is positively associated with BMI and gender but negatively with recent smoking and alcohol consumption. Finally, smoking and gender are proximal to several symptom criteria including fatigue, appetite problems, psychomotor changes and thoughts of death, suggesting that there may be gender differences underlying these relationships.

DRAFT

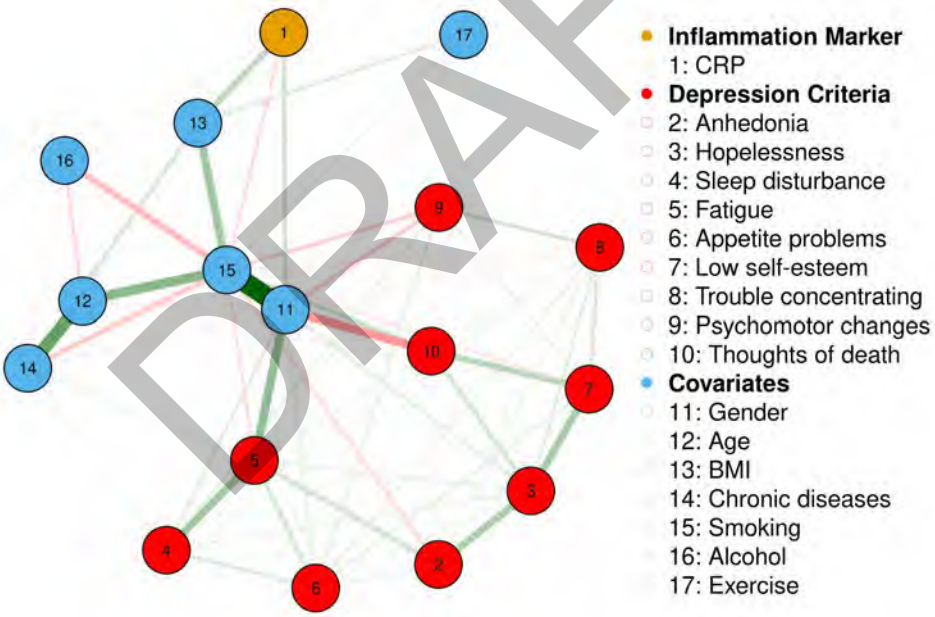



Figure 1: Estimated graph structure. Edges represent interactions that was found to be non-zero in at least 85% of the 500 bootstrap re-samples. Green edges in the networks depict positive associations, red edges represent negative associations, and grey ones signify interactions wherein no sign is defined. Thicker edges depict stronger associations.

## References


[1] Ali, A., Kolter, J.Z., Tibshirani, R.J.: The multiple quantile graphical model. Adv. Neural. Inf. Process. Syst. **29** (2016)

[2] American Psychiatric Association, D., Association, A.P., et al.: Diagnostic and statistical manual of mental disorders: DSM-5, vol. 5. American Psychiatric Association Washington, DC (2013)

[3] Chen, S., Witten, D.M., Shojaie, A.: Selection and estimation for mixed graphical models. Biometrika **102**(1), 47–64 (2015)

[4] Cheng, J., Li, T., Levina, E., Zhu, J.: High-dimensional mixed graphical models. J. Comput. Graph. Stat. **26**(2), 367–378 (2017)

[5] Chun, H., Lee, M.H., Fleet, J.C., Oh, J.H.: Graphical models via joint quantile regression with component selection. J. Multivar. Anal. **152**, 162–171 (2016)

[6] Du Clos, T.W.: Function of C-reactive protein. Ann. Med. **32**(4), 274–278 (2000)

[7] Finegold, M., Drton, M.: Robust graphical modeling of gene networks using classical and alternative t-distributions. Ann. Appl. Stat. pp. 1057–1080 (2011)

[8] Geraci, M., Farcomeni, A.: Mid-quantile regression for discrete responses. Stat. Methods Med. Res. **31**(5), 821–838 (2022)

[9] Koenker, R., Bassett, G.: Regression Quantiles. Econometrica **46**(1), 33–50 (1978)

[10] Kroenke, K., Spitzer, R.L., Williams, J.B.: The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)

[11] Lauritzen, S.L.: Graphical models, vol. 17. Clarendon Press (1996)

[12] Lauritzen, S.L., Andersen, A.H., Edwards, D., Jöreskog, K.G., Johansen, S.: Mixed graphical association models. Scand. J. Stat. pp. 273–306 (1989)

[13] Lee, C., Min, S.H., Niitsu, K.: C-reactive protein and specific depression symptoms among older adults: An exploratory investigation of multi-plane networks using cross-sectional data from NHANES (2017–2020). Biol. Res. Nurs. **25**(1), 14–23 (2023)

[14] Lee, J.D., Hastie, T.J.: Learning the structure of mixed graphical models. J. Comput. Graph. Stat. **24**(1), 230–253 (2015)

[15] Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High-dimensional semiparametric Gaussian copula graphical models. Ann. Stat. **40**(4), 2293–2326 (2012)

[16] Ma, Y., Genton, M.G., Parzen, E.: Asymptotic properties of sample quantiles of discrete distributions. Ann. Inst. Stat. Math. **63**(2), 227–243 (2011)

[17] Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. Ann. Stat. **34**(3), 1436–1462 (2006)

[18] Moriarity, D.P., Horn, S.R., Kautz, M.M., Haslbeck, J.M., Alloy, L.B.: How handling extreme C-reactive protein (CRP) values and regularization influences CRP and depression criteria associations in network analyses. Brain Behav. Immun. **91**, 393–403 (2021)

[19] Parzen, E.: Change PP plot and continous sample quantile function. Optim. **22**(12), 3287–3304 (1993)

[20] Parzen, E.: Quantile probability and statistical data modeling. Stat. Sci. pp. 652–662 (2004)

[21] Smith, K.J., Au, B., Ollis, L., Schmitz, N.: The association between C-reactive protein, Interleukin-6 and depression among older adults in the community: a systematic review and meta-analysis. Exp. Gerontol. **102**, 109–132 (2018)

[22] Xue, L., Zou, H.: Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Ann. Stat. **40**(5), 2541–2571 (2012)

[23] Yang, E., Baker, Y., Ravikumar, P., Allen, G., Liu, Z.: Mixed graphical models via exponential families. In: Artif. Intell. and Stat., pp. 1042–1050. PMLR (2014)


DRAFT

# The logratio Student t distribution

G.S. Monti[a] and G. Mateu-Figueras[b]

[a]University of Milan-Bicocca, Department of Economics, Management and Statistics;
gianna.monti@unimib.it
[b]University of Girona, Department of Computer Science, Applied Mathematics and Statistics;
gloria.mateu@udg.edu

**Abstract**

The main parametric families of distributions for random compositions include the Dirichlet distribution and some appropriate generalization of the Dirichlet family (7; 4), or the logratio normal family (6). In this contribution we consider as a valid alternative the logratio Student t distribution on the simplex, an extension to the additive logistic Student t distribution (2) derived from the multivariate t distribution in real space. We give its representation within the algebraic-geometric structure of the simplex compatible with the Aitchison measure. We illustrate the use of this model by a real data example, showing its usefulness in the modelling of compositional data.

***Keywords:*** Aitchison measure, logratio normal family, coordinate representations

DRAFT

## 1. The model

Katz and King (2) proposed an Additive logistic Student t distribution to model the composition of multiparty electoral data. Then the multivariate t distribution is used to model the additive logratio transformation (alr) vector and, finally, the inverse logistic transformation is used to return to the unit simplex $\mathcal{S}^D$. The alr is a non-isometric transformation which maps a composition in the $D$-part simplex into $\mathbb{R}^{D-1}$, considering the last part as common denominator of the others.

The additive logistic multivariate t distribution includes the additive logistic normal as a limiting special case. Here we define a multivariate t distribution on $\mathcal{S}^D$, that can be called logratio t distribution, in analogy with the logratio normal. The model, instead of using the alr transformed vector, is obtained using orthonormal log-ratio coordinates of a random composition and defining a density with respect to the Aitchison measure $\lambda_a$, which is compatible with the inner vector space structure of the simplex defined in Pawlowsky-Glahn (8), and could be considered a natural measure for the simplex. The Euclidean vector space structure of the simplex, the principle of working on coordinates (5) and the Aitchison measure allows us to define such isometric logistic Student t distribution.

A random composition $\mathbf{X} \in \mathcal{S}^D$ has a logratio Student t distribution with location $\boldsymbol{\mu} \in \mathbb{R}^{D-1}$, scatter matrix $\boldsymbol{\Upsilon}$ and $\nu$ degrees of freedom, if it has density function

$$f_a(\mathbf{x}) = \frac{dP_a}{d\lambda_a}(\mathbf{x}) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\boldsymbol{\Upsilon}|^{1/2}} \left(1 + \frac{1}{\nu}(h(\mathbf{x}) - \boldsymbol{\mu})'\boldsymbol{\Upsilon}^{-1}(h(\mathbf{x}) - \boldsymbol{\mu})\right)^{-\left(\frac{\nu+d}{2}\right)},$$

where $d = D-1$, $h(\cdot)$ stands for the orthonormal coordinates and $P_a$ is the logratio Student t probability measure. This distribution will be denoted by $\mathbf{X} \sim t_{\mathcal{S}}^D(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \nu)$.

Using the inverse of the Jacobian, $|d\lambda_a/d\lambda| = (\sqrt{D}x_1x_2\cdots x_D)^{-1}$, we can change the measure by the Radon-Nikodym chain rule, and express this density function with respect to the Lebesgue measure $\lambda$ in $\mathbb{R}^{D-1}$. The resulting expression is

$$f(\mathbf{x}) = \frac{dP}{d\lambda}(\mathbf{x}) = \frac{\Gamma(\frac{\nu+d}{2})(\sqrt{D}x_1x_2\cdots x_D)^{-1}}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\boldsymbol{\Upsilon}|^{1/2}}\left(1+\frac{1}{\nu}(h(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Upsilon}^{-1}(h(\mathbf{x})-\boldsymbol{\mu})\right)^{-\left(\frac{\nu+d}{2}\right)}. \quad (1)$$

The mode and the expected value of $\mathbf{X} \sim t_{\mathcal{S}}^D(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \nu)$, if $\nu > 1$, else is undefined, with respect to the measure $\lambda_a$ coincide and are

$$\text{mode}_a(\mathbf{X}) = \text{E}_a(\mathbf{X}) = h^{-1}(\boldsymbol{\mu}),$$

independently of the orthonormal coordinates $h(\mathbf{x})$.

For a composition $\mathbf{X} \sim t_{\mathcal{S}}^D(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \nu)$ there is no closed form for $\text{E}(\mathbf{X})$ with respect to the Lebesgue measure $\lambda$. Since the integral expression exists, it is not reducible to any simple form and it is necessary to use numerical integration to obtain it. Given $\mathbf{X} \sim t_{\mathcal{S}}^D(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \nu)$, when $\nu > 2$, then $\text{Mvar}(\mathbf{X}) = \frac{\nu}{\nu-2}\text{trace}\,(\boldsymbol{\Upsilon})$.

The parameter estimation is based on the maximum likelihood estimation (MLE) and the algorithm is obtained from the expectation-maximization (EM) method (3).

## 2. An example

The data set used is taken from Katz and King (1) related to 1979 British House of Commons electoral data. The three parties vote proportions, namely Conservative, Labour and Liberal from the British electoral system were considered. For this example the data were filtered from outliers and the missing values were removed.

After taking a suitable ilr transformation, the maximum likelihood estimates for the Normal Distribution in the Simplex are:

$$\hat{\boldsymbol{\mu}} = (-0.161, -0.268), \quad \hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.181 & -0.359 \\ -0.359 & 0.959 \end{pmatrix}.$$

While the maximum likelihood estimates for the Student t Distribution in the Simplex, where the degrees of freedoms $\nu$ are obtained in an iterative estimation with the rest of the parameters via the EM algorithm, are:

$$\hat{\boldsymbol{\mu}} = (-0.158, -0.278), \quad \hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.131 & -0.261 \\ -0.261 & 0.695 \end{pmatrix}, \quad \hat{\nu} = 6.364\,.$$

Figure 1 shows the sample in a ternary diagram and the isodensity curves at levels 50% and 95% of the fitted Normal Distribution in the Simplex (Left) and Student t Distributions in the Simplex (Right).

A different fit using the logratio t distribution is evident from the two ternary diagrams, in particular the outer region is slightly wider for the t than the normal as expected.

Figure 2 shows contours and sample plotted in the ilr coordinate space. The better fit of the multivariate t distribution is confirmed by the Akaike information criterion which is equal to 1145.367 and 1128.209 for the normal and the Student t model respectively (the lower, the better).

## References


[1] Katz, J. and G. King (2007). Replication data for: A Statistical Model of Multiparty Electoral Data. https://doi.org/10.7910/DVN/NDS9AT, Harvard Dataverse, V4.

[2] Katz, J. N. and G. King (1999). A statistical model for multiparty electoral data. *Am Polit Sci Rev 93*(1), 15–32.


DRAFT

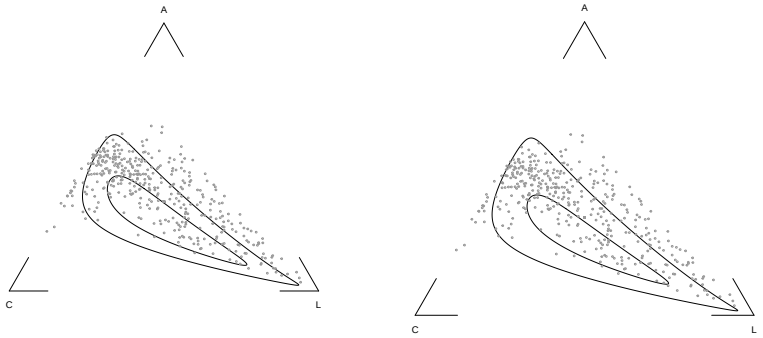



Figure 1: Ternary diagrams for 1979 British House of Commons electoral data (C=Conservative, L= Labour, A= Liberal/Alliance), and 50% and 95% confidence regions based on the Normal (Left) and on the Student t Distributions (Right).

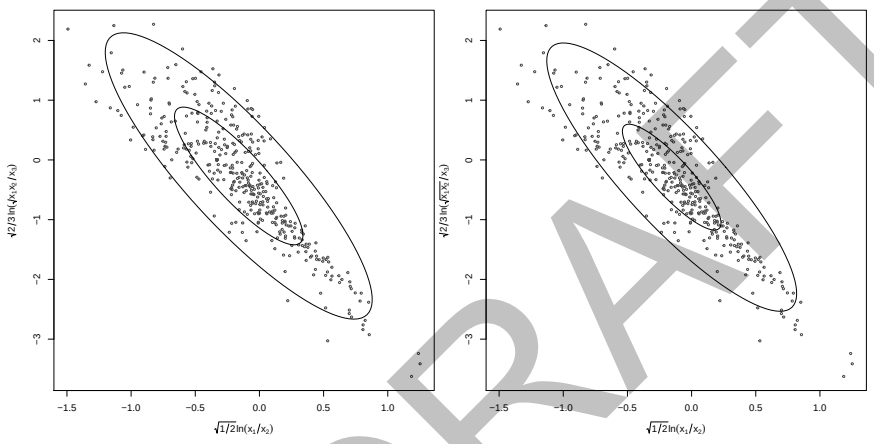


Figure 2: Plot of the ilr coordinates for 1979 British House of Commons electoral data ($x_1$=Conservative, $x_2$=Labour, $x_3$=Liberal/Alliance), with the corresponding fitted densities: the Normal in the Simplex (Left) and the Student t Distribtions in the Simplex (Right) and 50% and 95% confidence regions.

DRAFT

[3] Liu, C. and D. B. Rubin (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Stat Sin 5*(1), 19–39.

[4] Mateu-Figueras, G., G. S. Monti, and J. Egozcue (2021). Distributions on the simplex revisited. In P. Filzmoser, K. Hron, J. A. Martin-Fernandez, and J. Palarea-Albaladejo (Eds.), *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*, Chapter 3, pp. 61–82. Springer International Publishing.

[5] Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd.

[6] Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2013). The normal distribution in some constrained sample spaces. *SORT 37*, 29–56.

[7] Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). Notes on the scaled dirichlet distribution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis*, Chapter 10, pp. 128–138. John Wiley & Sons, Ltd.

[8] Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and

J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. University of Girona, Girona (Spain).

DRAFT

# A decomposition of the changes in tourism demand in Tuscany over the 2019-2021 period

Mauro Mussini[a]

[a] Department of Economics, Management and Statistics, Piazza dell'Ateneo Nuovo, 1 – 20126, Milan
mauro.mussini1@unimib.it

**Abstract**

The paper examines the changes in tourism demand in the provinces of Tuscany in the 2019-2021 period, when the tourism sector was severely hit by the COVID-19 pandemic. The change in the number of overnight stays is broken down into components by using the index decomposition analysis, which separates the contributions of changes in tourist arrivals, tourist-mix and length-of-stay. The decomposition results show that a fall in tourist arrivals was the main driver in reducing the number of overnight stays whereas the length-of-stay component partially mitigated the decline in overnight stays in all provinces, except Florence. A relatively small tourist-mix component suggests that the changes in the relative composition of tourism demand had a minor role in the change in overnight stays.

**Keywords:** decomposition, tourism demand, tourist flows.

DRAFT

## 1. Introduction

The tourism sector in Italy was severely affected by the outbreak of the COVID-19 pandemic, with a tourism demand that started to recover in 2021 but remaining far below the pre-pandemic levels [3]. All Italian regions suffered a sharp reduction in foreign tourist flows, especially the regions having a higher proportion of foreign tourists. Tuscany is one of the most visited regions by foreign tourists, also because of a diversified tourism product including art cities, sun and beach destinations, thermal spas and traditional landscapes. After a decade of growth, tourism demand in Tuscany fell dramatically in 2020 and it was still far from a complete recovery in 2021 [5]. In such a scenario, monitoring the dynamics of tourism demand is important for decision-making in the recovery plan of the tourism sector in the region.

Tourism demand in a destination is usually examined by observing some key indicators: the number of overnight stays, the number of tourist arrivals, the average length-of-stay. The changes in these indicators can be jointly analysed by using the log mean Divisia index decomposition of the variation in tourism demand [4], which breaks down the change in overnight stays by separating the roles of changes in tourist arrivals, tourist-mix and length-of-stay.

The paper is organized as follows: section 2 describes the index decomposition technique for the analysis of changes in tourism demand; section 3 shows the results of the decomposition applied to the changes in tourism demand in the provinces of Tuscany over the 2019-2021 period; section 4 concludes and suggests some directions for future research.

## 2. The index decomposition analysis of changes in tourism demand

The index decomposition analysis is based on the log mean Divisia index decomposition, a technique which is widely used in the studies on energy efficiency to separate the contributions of different factors to the change in energy consumption [1, 2]. Recently, the log mean Divisia index decomposition has been used to obtain a decomposition of the change in tourism demand [4], where tourism demand in a destination is measured by the number of overnight stays in that destination. This decomposition links the change in the number of overnight stays to the changes in the number of tourist arrivals, in the average length-of-stay and in the relative composition of tourism demand.

Let $D_{it}$ and $A_{it}$ be respectively the number of overnight stays and the number of tourist arrivals in destination $i$ in year $t$. Supposing that both $D_{it}$ and $A_{it}$ can be disaggregated into $k$ segments by a segmentation variable (e.g. type of accommodation), we have that $D_{it} = \sum_{j=1}^{k} D_{ijt}$ and $A_{it} = \sum_{j=1}^{k} A_{ijt}$ where $D_{ijt}$ is the number of overnight stays in segment $j$ and $A_{ijt}$ is the number of arrivals in the same segment, with $j = 1, \cdots, k$. The number of overnight stays in segment $j$ can be written as

$$D_{ijt} = A_{it} \frac{A_{ijt}}{A_{it}} \frac{D_{ijt}}{A_{ijt}} = A_{it} M_{ijt} L_{ijt}, \tag{1}$$

where $M_{ijt}$ is the proportion of arrivals in segment $j$ and $L_{ijt}$ is the average length-of-stay of tourists in that segment [4]. The total number of overnight stays in destination $i$ in year $t$ is expressed as

$$D_{it} = \sum_{j=1}^{k} A_{it} M_{ijt} L_{ijt} \tag{2}$$

and the change in overnight stays between two years, namely year 0 and 1, becomes

$$\Delta D_i = D_{i1} - D_{i0} = \sum_{j=1}^{k} A_{i1} M_{ij1} L_{ij1} - \sum_{j=1}^{k} A_{i0} M_{ij0} L_{ij0}. \tag{3}$$

The change in overnight stays in eq. (3) can be split into three components by using the log mean Divisia index decomposition [4]. $W_{ij}$ being the logarithmic mean of $D_{ij1}$ and $D_{ij0}$

$$W_{ij} = \begin{cases} \frac{D_{ij1} - D_{ij0}}{\ln D_{ij1} - \ln D_{ij0}} & if \ D_{ij1} \neq D_{ij0} \\ D_{ij1} & if \ D_{ij1} = D_{ij0} \end{cases}, \tag{4}$$

the change in overnight stays in eq. (3) is broken down into three components:

$$\Delta D_i = \sum_{j=1}^{k} W_{ij} \ln\left(\frac{A_{i1}}{A_{i0}}\right) + \sum_{j=1}^{k} W_{ij} \ln\left(\frac{M_{ij1}}{M_{ij0}}\right) + \sum_{j=1}^{k} W_{ij} \ln\left(\frac{L_{ij1}}{L_{ij0}}\right) = \Delta D_i^A + \Delta D_i^M + \Delta D_i^L, \tag{5}$$

where $W_{ij}$ is the weight of segment $j$ in terms of overnight stays. In eq. (5), $\Delta D_i^A$ is the tourist-flow component measuring the change in overnight stays due to a variation in tourist arrivals, $\Delta D_i^M$ is the tourist-mix component quantifying the change in overnight stays linked to changes in the relative distribution of arrivals among the tourist segments and $\Delta D_i^L$ is the length-of-stay component measuring the effect of changes in average length-of-stay in the segments [4].

## 3. An analysis of the changes in overnight stays in Tuscany

The changes in the number of overnight stays in the provinces of Tuscany are examined over the 2019-2021 period; i.e., from the year before the outbreak of the COVID-19 pandemic in Italy to the year when the tourism sector started to recover. Data on tourist arrivals, country of residence of tourists (Italy or foreign countries) and overnight stays are from the tourism database of the Tuscany Region [6]. The decomposition of the change in the number of overnight stays is obtained by considering two tourist segments, resident and non-resident tourists, in each destination.

Table 1 shows the decomposition of the changes in the number of overnight stays from 2019 to 2021 by province. The variation in the number of overnight stays was negative in each province of Tuscany, but with differences in the magnitude of the decrease. Florence had the largest reduction in overnight stays, a result that one might expect since it was the province with the largest number of overnight stays in 2019, most of which recorded for non-resident tourists. The large decreases in overnight stays occurred in the provinces of Siena, Pistoia and Pisa seem to suggest that the provinces with a tourism product based on cultural and artistic heritage were particularly affected by the COVID-19 pandemic. A graphical analysis of the tourist-flow,

DRAFT

tourist-mix and length-of-stay components may help to compare their respective roles in the change of overnight stays.

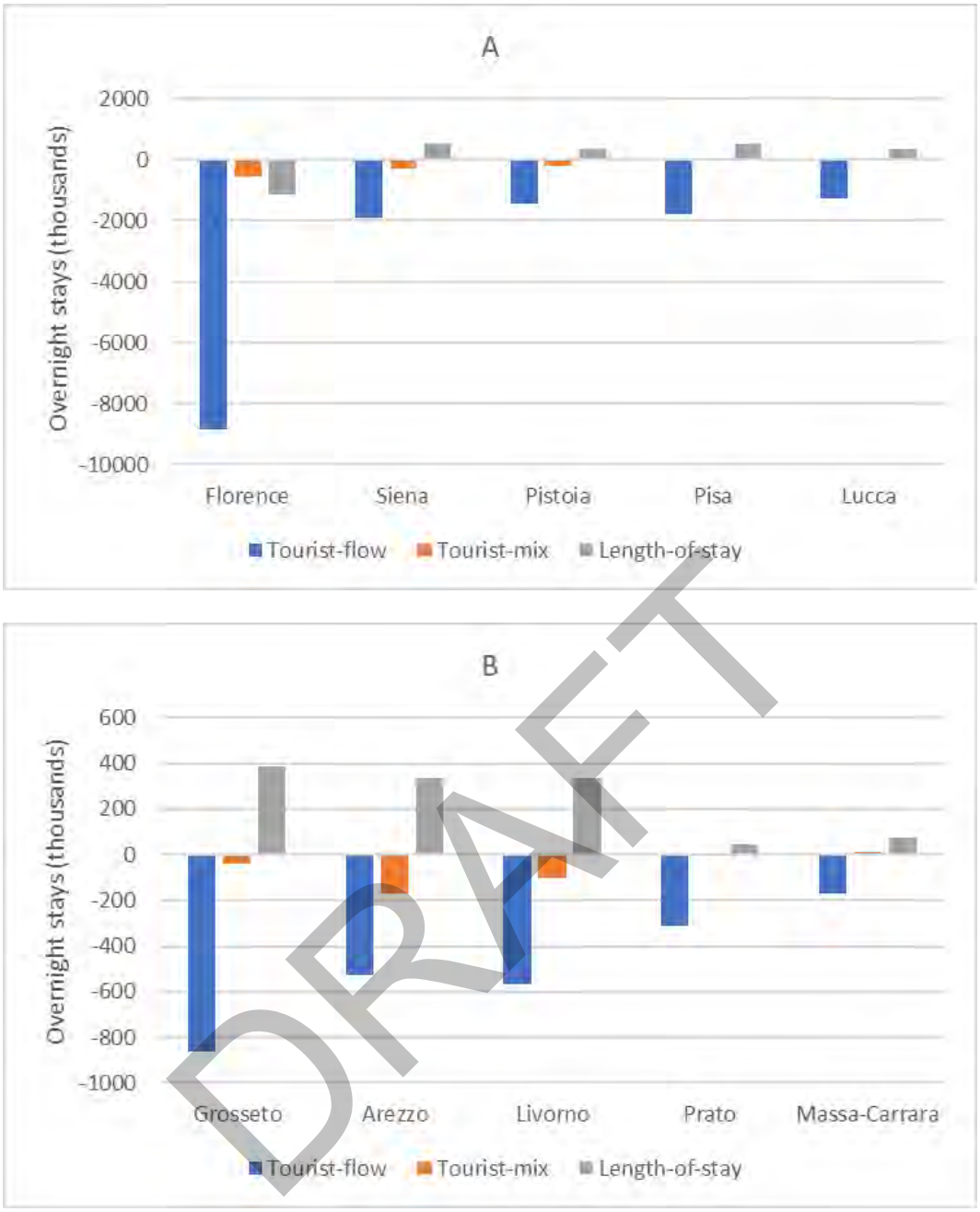
Table 1: Components of the change in overnight stays in the provinces of Tuscany, 2019-2021.

| Provinces | ΔD | Components of ΔD | | |
|---|---|---|---|---|
| | | Tourist-flow | Tourist-mix | Length-of-stay |
| Florence | -10524238 | -8848904.36 | -552665.44 | -1122668.19 |
| Siena | -1636332 | -1894697.08 | -278242.79 | 536607.87 |
| Pistoia | -1307427 | -1451079.87 | -207949.66 | 351602.52 |
| Pisa | -1154175 | -1771075.20 | 71388.56 | 545511.65 |
| Lucca | -928869 | -1274969.85 | -29314.12 | 375414.97 |
| Grosseto | -512546 | -863350.72 | -37955.46 | 388760.18 |
| Arezzo | -359158 | -524906.65 | -171713.22 | 337461.88 |
| Livorno | -333021 | -566559.55 | -100594.33 | 334132.88 |
| Prato | -265244 | -309929.57 | 244.57 | 44441.00 |
| Massa-Carrara | -79761 | -167747.92 | 10833.27 | 77153.65 |

Figure 1 (panels A and B) shows the three components of the change in overnight stays for each province. Florence, the province with the largest decrease in overnight stays, has a negative sign for each of the three components of the change in the number of overnight stays. The decline in tourist arrivals had the most important role in reducing the number of overnight stays in the province of Florence. The loss of overnight stays in the provinces of Siena, Pistoia, Pisa and Lucca (figure 1, panel A) was due to the reduction in tourist arrivals, which was only partially counterbalanced by a positive length-of-stay component while the tourist-mix component was relatively small or almost negligible (in Pisa and Lucca). This suggests that the changes in the proportions of resident and non-resident tourists had a minor impact on the variation in the number of overnight stays.

The number of overnight stays would have increased for effect of the length-of-stay component in the provinces of Grosseto, Arezzo and Livorno (figure 1, panel B), however the decrease in tourist arrivals resulted in a negative tourist-flow component which overcame, in absolute value, the positive contribution of the change in length-of-stay in such provinces. The tourist-mix component contributed to reducing the number of overnight stays in the above three provinces, especially in the province of Arezzo where the proportion of non-resident tourist arrivals had a large decrease. The changes in the number of overnight stays were smaller in the provinces of Prato and Massa-Carrara, but it is worth mentioning that overnight stays in such provinces were lower than those in other provinces at the beginning of the period under consideration.

DRAFT




DRAFT

Figure 1: Decomposition of the changes in overnight stays from 2019 to 2021 by province.

## 4. Conclusion

The analysis of changes in tourism demand in the provinces of Tuscany points out that overnight stays were still below their pre-pandemic levels in 2021, especially where the tourism product is mainly based on cultural and artistic heritage. The decomposition of the changes in the number of overnight stays shows that a decline in the number of tourist arrivals was the main driver of the decrease in overnight stays observed in each province between 2019 and 2021. The length-of-stay component is always positive, except in the province of Florence, suggesting that an increase in the average length-of-stay mitigated to some extent the negative impact of the fall in tourist arrivals. In other words, far fewer tourists arrived in Tuscany in 2021 compared with 2019 but they stayed a little longer. The variation in the proportions of resident and non-resident tourist arrivals had a minor role in changing the number of overnight stays in Tuscany over the 2019-2021 period.

Future research might examine the drivers of the change in tourism demand by segmenting tourists according to the type of accommodation they chose and by extending the period of observation once the

2022 data are available for analysis. A further direction for future research is the development of a decomposition which can take the spatial relationships between destinations into account when breaking down the change in tourism demand of a region.

## References


[1] Ang, B.W., Zhang, F.Q., Choi, K.H.: Factorizing changes in energy and environmental indicators through decomposition. Energy **23**, 489--495 (1998)

[2] Ang, B.W.: The LMDI approach to decomposition analysis: A practical guide. Energy Policy **33**, 867--871 (2005)

[3] Istat: Movimento turistico in Italia | gennaio – settembre 2021. Report - 12 January 2022

[4] Mussini, M.: An index decomposition analysis of tourism demand change. Ann. Tour. Res. **85**, (2020) 102902

[5] Regione Toscana: Il movimento dei clienti negli esercizi ricettivi della Toscana: dati di sintesi 2021. Report - April 2022

[6] Regione Toscana: Banca dati Turismo. Accessed 9 January 2023


DRAFT

# Bayesian networks as a territorial gender impact assessment tool

Flaminia Musella[a], Lorenzo Giammei[b], Fulvia Mecatti[b], and Paola Vicard[c]

[a]Link Campus Unversity, Casale di San Pio V 44 Rome; f.musella@unilink.it
[b]University of Milan-Bicocca, Piazza dell'Ateneo Nuovo 1 Milan; lorenzo.giammei@unimb.it, fulvia.mecatti@unimib.it
[c]Roma Tre University, Via Silvio d'Amico 77 Rome; paola.vicard@uniroma3.it

**Abstract**

Gender impact assessment is a process that evaluates the effect of gender policies and motivates the introduction of subsequent improvement strategies oriented at balancing out gender inequalities. Currently, one of the most employed tools to assess the effect of gender-related policies is the European Gender Equality Index, a composite indicator that summarizes several dimensions of the gender gap. In this paper, a Bayesian network-based approach is proposed as an alternative for combining dimensions of the index. Such modelling, discussed through an application, allows the construction of a network where ingredients of the index interact with socio-economic variables and the inferential engine associated to the network is employed to carry-out both *ex-ante* and *ex-post* analysis. As a result, the method can be a valuable tool for supporting the gender impact assessment.

***Keywords:*** Bayesian Network, GEI, Impact evaluation, Province Level, Value of Information Analysis

DRAFT

## 1. The rationale

Structural gender inequalities are embedded in the society. There is a risk of being gender blind that leads to ignore the specific effects a governmental decision or policy can have on different genders. The European Commission defines gender impact assessment as a tool for detecting if policies are impacting differently between women and men and thus reinforcing or mitigating inequalities ($urly.it/3q_q4$). The process of quantifying the impact of a policy on the gender gap is however unavoidably tied to the measure of gender gap itself. In macroeconomics, composite indicators are popular for their ease of interpretation (8). One of the most used in Europe, is the Gender Equality Index (GEI) developed by the European Institute for Gender Equality (EIGE) for monitoring and compare the gender equality at national level within EU Member States. GEI is based on 6 core domains, measured in 14 sub-domains through 31 variables, here called indicators, playing the role of ingredients of the computation. Ingredients are aggregated through arithmetic means within sub-domains and through geometric means across sub-domains to produce domain-level scores. Finally scores are linearly combined across domains according to a weighting system informed upon experts' advice.
Even though composite indicators like GEI are largely adopted and have an increasingly complex architecture of aggregation, their result is a single value, without any multivariate feature. Indeed, composite indicators do not provide any information about the dependence structure among its ingredient variables and cause-effect relations between specific sets of variables contributing to a certain level of the indicator. In this sense a clear improvement of GEI would be to complement it with a multivariate statistical

approach to holistically handle the gender gap, providing policy makers with a statistical tool to evaluate and compare scenarios of gender policies (7). An integrated approach could result in a richer set of information to orient impact assessment while, at the same time, controlling for collateral negative impact. The aim of the paper is to present a structured framework where GEI is supported by a multivariate graphical statistical model, namely a Bayesian network, and to highlight the advantages of this approach with regard to policy guidance and impact assessment. An empirical application on Italian data at province level is provided to show the functioning of the novel method. We firstly discuss the data availability for reproducing the European Gender Equality Index (GEI) at Italian provincial level (Section 2); in Section 3. we estimate the Bayesian network and we show its potential in Section 4.

## 2. Gender data availabilty

Gender impact assessment promotes transparent decision-making and better governance to achieve a relevant impact on the society and human well-being. A crucial point of the gender impact assessment is the availability of gender-sensitive data, i.e. data collected and analysed by accounting for sex. Moreover, to address geographical heterogeneities, gender sensitive data should be collected with territorial granularity. Undoubtedly, inclusion strategies should be rooted in a territorial specificity to be more effective (2). Nevertheless, gender micro-data may be unavailable, compromising a proper gender impact assessment. United Nations Member States (UN) shared, by an Agenda of Sustainable Development, a set of goals aiming at peace and prosperity for people and the planet but in Italy, according to Openpolis.it, in 2020, 52.5% of gender indicators proposed by UN for monitoring the achievement of the Agenda 2030 were missing.

In this work, firstly, we employ ISTAT data to reconstruct a province level GEI (PV-GEI). This step is similar to what has been made at regional level by (1). Figure 1 compares the data availability of ingredient variables of GEI at regional and provincial level, eventually also denoting proxy variables availability.

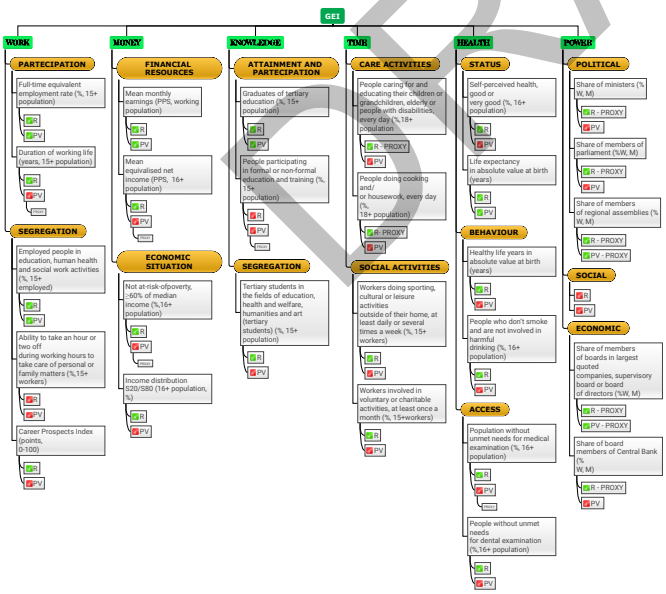



Figure 1: Ingredient variables of GEI: a comparison between data availability at regional and provincial level

The PV-GEI has been built on the basis of available data, so that, with respect to the national GEI, the Time domain is missing. The Italian PV-GEI distribution is depicted in Figure 2. The map shows that highest values of GEI belong to larger cities or northern cities.

DRAFT

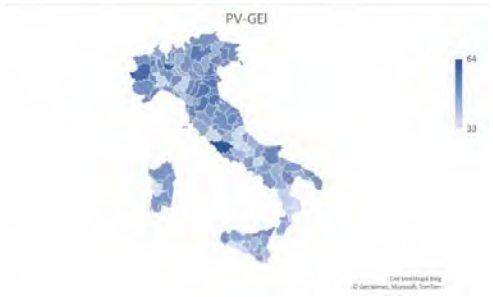



Figure 2: Map chart of GEI at provincial level

## 3. A Bayesian Network for the gender impact assessment

The second phase of our work consists in employing a Bayesian network to support and integrate the measured PV-GEI that, here, has been treated as a variable. The point of departure for this contribution is the model presented in (5). Bayesian networks (BNs, (9)) consist of a directed acyclic graph (DAG) and an associated joint probability distribution, that can be interpreted as a conditional independence map. A DAG is a graph $G(V, E)$ consisting of a set of nodes (or vertices) $V$ and a set $E$ of directed edges that cannot give rise to directed cycles, *i.e.* directed paths starting and ending with the same node following arrow direction. The vertices of the DAG represent random variables and the edges describe the relations between them. A BN is provided with an inferential engine that allows for what-if-analyses. To do so, evidence (i.e. a possible policy) is inserted in the network and then it is propagated throughout it producing the updated marginal distributions. The model can be thus interpretable as a tool for impact assessment.
The proposed model has been developed at Italian provincial level since conducting the analisys at a fine granularity level is crucial to study the determinants of inequalities and thus, for supporting effective decision-making ($urly.it/3ra09$). The statistical model we discuss has been directly learnt from data by means of structural learning (3). Specifically we have employed the PC algorithm (4), a procedure that performs multiple conditional independence tests on data and then translates the obtained independence statements into a DAG. The learning phase has been conducted using the statistical software Hugin. The variables contained in the dataset are the PV-GEI, GEI ingredients and synthetic domains (as depicted in Figure 1), as well as additional structural and socioeconomic features, called extra variables. The latter have been included in the model to obtain a BN that not only captures the relationships among gender-related variables but also investigates if and how other social and economic dimensions can affect the gender gap. Focusing on a wider picture could improve our understanding of how gender-based differences are tied to other domains of our society and provide policy makers with a deeper awareness of the systemic impact of their interventions.

In the learning process, it is assumed that extra variables are not affected by GEI ingredients; consequently, edges from GEI have been forbidden and indicators built to summarize the GEI domains have been forced to be linked to the PV-GEI. The resulting DAG is shown in Figure 3. PV-GEI has been inserted in the network as an additional node, with incoming arcs from each synthetic score domain. The presence of the PV-GEI node allows ranking provinces in terms of gender equality. The colour of the vertices denotes the variables group (orange= extra variables, light green= GEI ingredients, green=GEI domains synthesis, dark green= PV-GEI). The graph provides a clear picture of how all the variables interact under a full multivariate approach.

## 4. Discussion

The model in Figure 3 can be used in both *ex-ante* and *ex-post* perspectives. What-if analysis can provide policy makers with a powerful tool for an *ex-ante* analysis. The simulation of scenarios could help identifying the most effective policy to mitigate gender-related disequities and promote social and

DRAFT

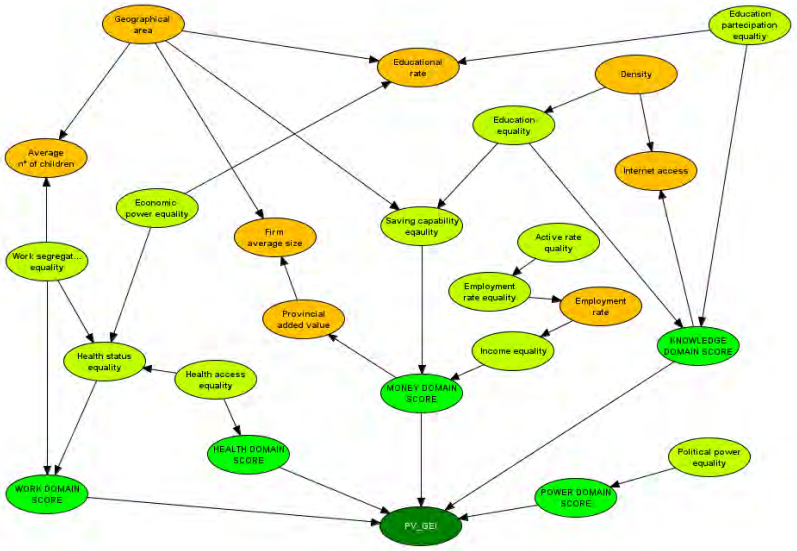



Figure 3: Bayesian network model for territorial gender impact assessment

economic well-being. Moreover, the obtained BN can be used to perform value of information analysis (VOI). VOI allows the user to quantify the mutual information between variables and, thus, to explore the most impacting variables for an area of interest. This may help in defining priorities when considering different improvement actions that target the gender gap. Taking an *ex-post* perspective, the impact of policies can be evaluated by performing a causal analysis on the network (6). A causal use of Bayesian networks requires stronger hypotheses in the structural learning phase (4). However, if the assumptions hold, it allows estimating the impact of policies even when only *ex-post* policy-related observational data are available.

An integrated use of composite indicators and BN can thus enhance the role of statistics in the analysis of gender gap, by providing a tool-set able to simultaneously measure, monitor, predict and evaluate in a unified framework.

DRAFT

## References


[1] Bella, E., Leporatti, L., Gandullia, L., & Maggino, F. Proposing a regional gender equality index (R-GEI) with an application to Italy. Regional Studies, 55(5), 962-973 (2021)

[2] Bozzato, S., Salvatori, F., & Ricci, A. Social Inclusion and Territorial Dynamics. In Territorial Impact Assessment of National and Regional Territorial Cohesion in Italy. Place Evidence and Policy Orientations Towards European Green Deal (pp. 121-131). Patron. (2020).

[3] Daly, R., Shen, Q., Aitken, S. Learning Bayesian Networks: Approaches and Issues. The knowledge engineering review, 26 (2), 99–157 (2011).

[4] Glymour, C.; Spirtes, P.; Scheines, R. Causal Inference. Erkenntnis, 35 (1–3), 151–189 (1991).

[5] Mecatti, F., Vicard, P., Musella, F., & Giammei, L. Bayesian networks versus gender bias. Significance, 19(5), 16-20 (2022)

[6] Morgan, S., & Winship, C. Counterfactuals and Causal Inference: Methods and Principles for Social Research. 2nd ed., Analytical Methods for Social Research (2014)

[7] Musella, F. Giammei, L. Romio, S., Vicard, P., Mecatti, F. Bayesian networks for monitoring the gender gap. 51th SIS Scientific Meeting of the Italian Statistical Society. Book of short papers-SIS 2022. pp. 958–963 (2022)

[8] Joint Research Centre-European Commission. *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing (2008).

[9] Pearl, Judea. Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress 19.2 (2000).

# Can statistics be helpful in detecting electoral fraud?

Massimo Attanasio[a], Vincenzo G. Genova[a], and Michele Tumminello[a]

[a] Department of Economics, Business, and Statistics. University of Palermo; massimo.attanasio@unipa.it,vincenzogiuseppe.genova@unipa.it, michele.tumminello@unipa.it

**Abstract**

This paper presents a case study of electoral fraud in the Italian Overseas Constituencies during the 2018 Italian Election. The study leverages publicly available electoral data obtained from the Central Office for Foreign Circumscriptions at the Rome Court of Appeal to conduct an analysis of voting preferences expressed for the party "Unione Sudamericana Emigrati Italiani" (USEI) and evaluates statistical anomalies in the preferences expressed for the candidate Cario Adriano. The study employs a comprehensive four-step methodology that includes numerical analysis, Monte Carlo simulations, validation, and outlier detection analysis. The results suggest that there is evidence of electoral fraud in the polling stations of Buenos Aires, with specific anomalies identified in the preferences expressed for Cario Adriano.

**Keywords:** Electoral fraud, Italian Overseas Constituencies, Southern American election

DRAFT

## 1. Introduction

The Italian Electoral Law of 2017, also known as Rosatellum Bis, is a mixed electoral system that allocates 37% of seats through a first-past-the-post system and 63% proportional allocation. The Chamber and Senate of the Republic use the largest remainder method to allocate proportional seats, and both houses are elected in a single round of voting [1]. The law replaced the 2005 Porcellum and 2015 Italicum laws, which were declared partly unconstitutional by the Italian Constitutional Court. The Chamber of Deputies has 400 members, with 147 elected in single-member districts, 245 elected through proportional representation, and 8 Overseas Constituencies elected by the Italians abroad [1]. On the other hand, the Senate has 200 elected members, with 74 elected in single member districts, 122 elected through proportional representation, and 4 Overseas [1]. The Senate election uses a single ballot system which includes the district representative elected through a simple majority. The ballot also lists the political parties and party lists that support the representative, used to determine proportional representation with a 3% minimum threshold for party representation. Finally, in the composition of the Senate, a limited number of senators-for-life are appointed for life by the President of the Republic [1]. Concerning the Overseas Constituencies' election, the Italian Parliament has established the Overseas Constituencies to guarantee representation for Italian citizens residing abroad. The right to vote by mail was introduced just for the Italian citizens residing abroad since 2001 [2]. Consular offices are responsible for sending electoral ballots to eligible voters, collecting the ballots, and sending them to the Chamber of Deputies and the Senate of the Republic in Rome. Abroad voters cast their vote by marking the symbol associated with their preferred list, and they can also express a vote of preference by writing the candidate's last name [2]. To write the candidate's name is a prerogative of the Italians residing abroad, while in Italy voters are allowed to mark just the party.

This work regards Fabio Porta, a deputy of the Democratic Party in the Italian Parliament elected in 2022. In 2018, Porta ran as a candidate in the South America constituency for a seat in the Senate of the Republic, which was instead awarded to Adriano Cario. Porta reported Cario for electoral fraud and the Senate conducted a thorough judicial investigation into the preference votes. In October 2021, Porta, on

the advice of his lawyer, requested statistical consulting from the authors of this work to support the findings of previous investigations that had revealed anomalies at the consulate of Buenos Aires. The statistical anomalies founded in this paper have been used by the lawyer to support their legal dispute. In January 2022, Porta replaced Cario as senator and won the legal dispute.

The consultation was organised according to a four-step statistical methodology to uncover anomalies. The first involved a preliminary numerical analysis. In the second step, Monte Carlo simulations were carried out on the Buenos Aires polling stations. In the third step, the results of the Monte Carlo simulations were validated by performing an indirect verification using the polling stations of another consulate as a validation set.

Finally, in the fourth step, an outlier detection analysis was performed on the polling stations in Buenos Aires to identify any preferences that were "outlier," through the Bonferroni test statistics.

The paper is organised as follows: in Section 2, we illustrate the data and aim; in Section 3, we present the four-step statistical procedure; in Section 4, we will offer some conclusions.

## 2. Data and aim

The present study uses electoral data of the Senate, published by the Central Office for Foreign Circumscriptions at the Rome Court of Appeal. This dataset regards the Italian election of the Southern American Overseas Constituencies held on March 4th, 2018. The dataset provides voting preferences expressed for each Overseas candidate at individual polling sections within Consular offices (Table 1 reports just some Polling Sections).

The aim of this paper is to assess statistical anomalies in the preferences expressed for the candidate Cario Adriano in the Buenos Aires consulate. Rosario Consulate is used as a validation set. This examination was carried out while controlling for the preferences expressed for other candidates within the same political party. To achieve the objectives of our analysis, we analyzed voting preferences expressed in Buenos Aires (100 polling sections) and Rosario (47 polling sections) for the candidates of the "Unione Sudamericana Emigrati Italiani" (USEI) party.

Table 1: Preferences for the USEI candidates in Buenos Aires and Rosario consulates by Polling Sections

| Consulate | Polling Section | USEI candidates | | | | **Total** |
|---|---|---|---|---|---|---|
| | | Nardelli | Vicentini | Cario | Moya | |
| Buenos Aires | 1 | 83 | 0 | 3 | 191 | **277** |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 23 | 115 | 2 | 49 | 104 | **270** |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 100 | 12 | 1 | 344 | 6 | **363** |
| **Total** | | **3,645** | **184** | **21,972** | **2,437** | **28,238** |
| Rosario | 1 | 46 | 1 | 0 | 20 | **67** |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 23 | 52 | 3 | 3 | 37 | **95** |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 47 | 30 | 2 | 0 | 38 | **70** |
| **Total** | | **1,860** | **107** | **69** | **1,097** | **3,133** |

## 3. The four-step statistical procedure

### 3.1 Preliminary analysis

In the first step, we built for the two consulates two contingency tables IxJ: the first table (100x2) refers to Buenos Aires, and the second (47x2) to Rosario. In these tables, the columns of the candidates Nardelli, Vicentini, and Moya are collapsed into one (Other USEI). So, we calculate the variances and coefficients of variation of preferences expressed for candidate Cario and other USEI candidates calcu-

DRAFT

lated over the 100 polling sections in Buenos Aires (Table 2). The results indicate that Cario's preferences variance in Buenos Aires (41,734) is higher than that of the other USEI candidates (33,175). This difference is evident in the higher coefficient of variation too: Cario's coefficient is 0.93 and the others' coefficient is 0.64.

Table 2: Variances and coefficient of variation for the USEI candidates

| USEI candidates | Variances | Coefficient of Variation |
|---|---|---|
| Cario | 41,734.7 | 0.93 |
| Other USEI | 33,175.6 | 0.64 |

These findings suggest that Cario's variability over the polling sections is much higher than the others.

### 3.2 Monte Carlo simulation

To detect any potential anomalies in the election data, we compare the real data (observed) to the data obtained via a random ballot allocation (expected). According to the rules adopted in each consulate, each polling section should collect the ballots into a precise number of boxes (100 for Buenos Aires, 47 for Rosario, etc.), just following the order of delivery.

To identify anomalies in the election results, we are conducting a Monte Carlo simulation by generating 10,000 preference tables assuming independence and proportional allocation of preferences to each cell based on fixed marginals [3], given by the collapsed contingency tables (see Section 3.1). The simulation is based on a multivariate hypergeometric distribution [4], which yields 10,000 contingency tables for each consulate. We use these tables to calculate 10,000 variances, one for each simulated table. The empirical distribution of these variances is shown in Figure 1. By comparing the observed variance of preferences expressed for Cairo, which is 41,734.7, to the range of variances [27,000 – 30,000] obtained from the Monte Carlo simulation, it is evident that the observed variance is significantly outside this range.

DRAFT

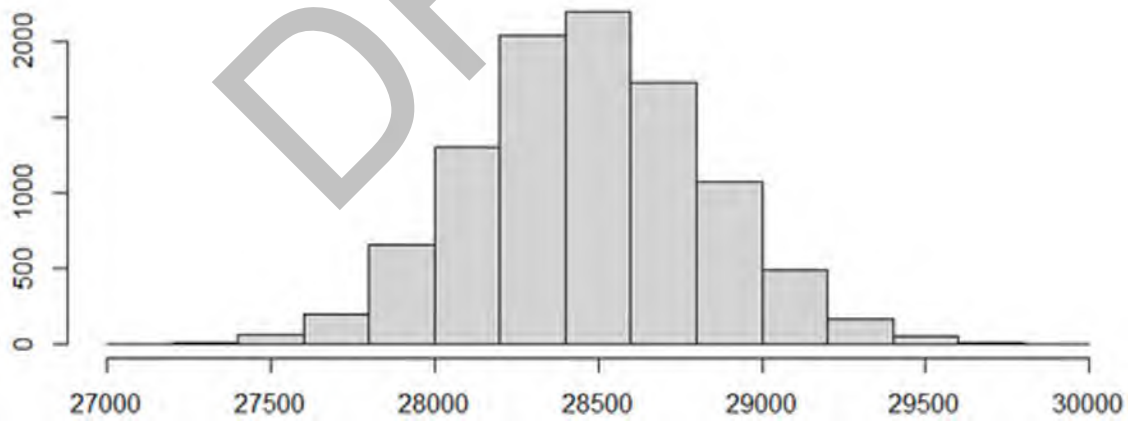



Figure 1: Variance distribution of the 10,000 simulated contingency tables for Buenos Aires

### 3.3 Validation set

We are validating the simulation procedure described in Section 3.2 by replicating the simulation on the validation set in Rosario, which is the second largest Argentinian consulate after Buenos Aires. As with the previous simulation, we generate 10,000 contingency tables, from which we calculate 10,000 variances – one for each simulated table. The empirical distribution of these variances is shown in Figure 2. By comparing the observed variance for Cairo, which is 2.34, to the range of variances [0.5

– 3.0] obtained from the Monte Carlo simulation, it is evident that the observed variance falls perfectly within this range. This result indirectly suggests that the simulated data closely approximate the actual case.

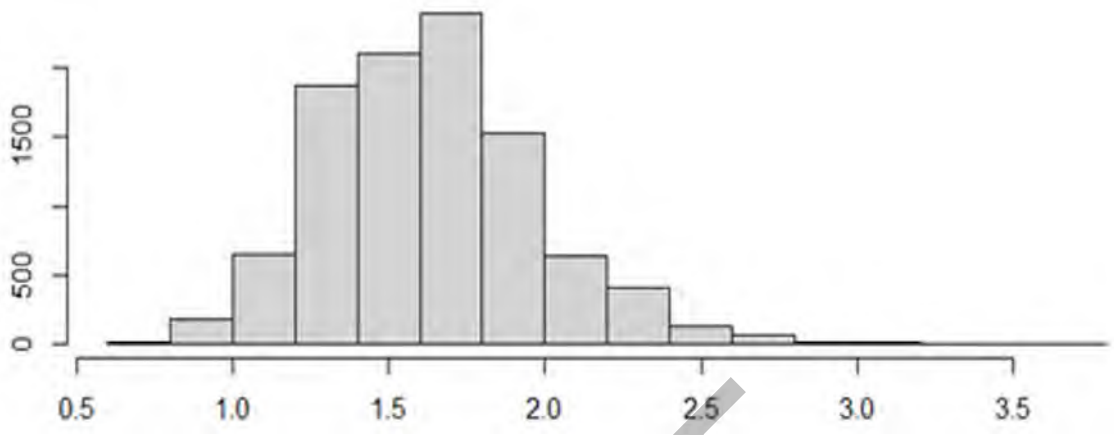



Figure 2: Variance distribution of the 10,000 simulated contingency tables for Rosario.

DRAFT

### 3.4 Outlier Polling Section detection

To identify any tampered polling sections at the Buenos Aires consulate, as suggested by the legal investigations, we have tested a null hypothesis of a random distribution of votes across the different sections. This hypothesis depends on both the total number of votes received by each candidate and the total number of preferences included in each polling section. Specifically, we use the total number of electors ($N$), the total number of votes received by candidate Cario ($K_c$), the total number of votes in the *i-th* section ($n_i$), and the number of votes in the *i-th* section allocated to candidate Cario ($n_{ic}$).

The aim is to test the hypothesis that $n_{ic}$ follows a hypergeometric distribution of parameters $n_i$, $K_c$, and $N$. This means testing the hypothesis that votes are randomly distributed across the sections. A *p-value* can then be calculated to detect any overexpression of votes for the candidate Cairo in each *i-th* section, as follows:

$$P(n_{ic} \geq x_{ic}) = \sum_{n_{ic}=x_{ic}}^{\min(n_i,K_c)} \frac{\binom{n_i}{n_{ic}}\binom{N-n_i}{K_c-n_{ic}}}{\binom{N}{K_c}}, \quad (1)$$

where $x_{ic}$ is the actual number of votes to Cairo included in the *i-th* section.

We associated a p-value with each electoral section of Buenos Aires (100 polling sections). Therefore, statistical significance has evaluated using a correction for multiple hypothesis testing. Since the analysis concerns potentially fraudulent activities, it is advisable to minimize type one errors (false positive rate). Thus, we use the most conservative correction for multiple tests, that is, the Bonferroni correction [5]. Accordingly, by setting a univariate threshold of statistical significance equal to $\alpha = 0.01$, we select all the poling section with a *p-value* lower than $\alpha/100 = 0.0001$. The *p-values* obtained from this procedure indicate the presence of 30 tampered sections. This result is in line with the calligraphic analysis conducted during the process, which identified 32 tampered sections.[1]

[1] Unfortunately, we could not check the overlapping between the 30 polling section that turned out to be suspected according our statistical test and the 32 polling sections identified by calligraphic expertise, since the list of the latter polling sections was not provided to us for privacy reasons.

## 4. Conclusions

The findings of this study suggest the presence of anomalies in the distribution of preferences expressed for Cario in Buenos Aires, as demonstrated by several statistical proofs: the significantly higher variance and coefficient of variation, the Monte Carlo simulation, and the "corrected" Fisher test applied to the contingency tables. These findings can be used as a straightforward application of statistical tools to detect anomalies (fraud!) in electoral data.

## References


[1] Legge 3 novembre 2017, n. 165. Modifiche al sistema di elezione della Camera dei deputati e del Senato della Repubblica. Delega al Governo per la determinazione dei collegi elettorali uninominali e plurinominali. (17G00175) (GU Serie Generale n.264 del 11-11-2017)

[2] Legge 27 dicembre 2001, n. 459. "Norme per l' esercizio del diritto di voto dei cittadini italiani residenti all' estero ", pubblicata sulla Gazzetta Ufficiale n. 4 del 5 gennaio 2002.

[3] Duan, X. G. "Better understanding of the multivariate hypergeometric distribution with implications in design-based survey sampling." arXiv preprint arXiv:2101.00548 (2021)

[4] Sprent, P. (2011). Fisher exact test. In International encyclopedia of statistical science (pp. 524-525). Springer, Berlin, Heidelberg.

[5] Miller, R. G. (1981). Simultaneous Statistical Inference. New York, Springer-Verlag.


DRAFT

# Companies' sustainability disclosure and contrast to hunger: the role of social inclusion

Chiara Di Maria[a] and Rodolfo Damiano[a]

[a]Department of Economics, Business and Statistics, University of Palermo;
chiara.dimaria@unipa.it, rodolfo.damiano@unipa.it

**Abstract**

In order to achieve the sustainable development goals (SDGs) stated in the Agenda 2030, different actors at each level of society are involved. In particular, companies play a key role and have started to disclose their contribution to SDGs through their sustainability reports. In this work, we focus on the second sustainable goal, 'Zero hunger', and analyse which factors are linked with higher sustainable nutrition scores reported by a sample of companies considered as the most influential for the achievement of SDGs. In particular, we investigate the association between nutrition scores, the presence of a corporate social responsibility committee within companies and their reported level of social inclusion, testing if the latter can play a mediating role.

***Keywords:*** SDGs, nutrition, social inclusion, sustainability disclosure, mediation analysis

DRAFT

## 1. Introduction

In September 2015, the 193 countries members of the United Nations signed Agenda 2030 consisting of 17 sustainable development goals (SDGs) to achieve by 2030. These objectives cover several areas, ranging from contrasting poverty and hunger, achieving gender equality, promoting human rights, guaranteeing equal access to education, and protecting the environment. Every country, by means of its government, institutions and citizens, has to give its contribution to fulfilling the SDGs. Indeed, governments cannot achieve SDGs in isolation; they need the support and commitment of several social forces. In this respect, a key role is played by large companies, whose activities affect not only the economy of the countries where they operate but also have social and environmental implications.

Academia and professionals agree about the prominent role of the business community in SDGs achievement; however, research is still needed to comprehend how companies are coping with this role. Large companies have already demonstrated a proactive attitude toward SDGs (3). In recent years, several companies have started to disclose their contribution to SDGs achievement through their sustainability reporting practices. Moreover, companies claim to have integrated these goals into their business strategy to create social value by achieving optimal financial, social, and environmental performance. According to previous studies, sustainability disclosure is a booster for corporate social responsibility performance (6) and can change how organizations think and act (1). However, while companies can use these sustainability-related disclosures to enact and improve their sustainable behaviours towards the achievement of SDGs, in contrast, the accounting literature has also pointed out that these can be used to gain legitimacy, masking business as usual and destructive practices as sustainable ones (10; 8; 15). In this regard, accounting researchers have suggested that these disclosure practices have to be transparent, more inclusive and grounded on principles such as democracy and engagement to be effective

and to improve companies' contributions to sustainable development (4; 5). In this light, and focusing on the second SDG ("Zero hunger", denoted SDG2 in the following), the aim of this work is to test whether companies providing more information regarding their sustainability strategies and more evidence of inclusiveness also achieve a higher level of nutritional sustainability-related disclosure and if social inclusion can act as a mediator in the relationship.

## 2. Data description

We used data from the SDG2000 database, created by the World Benchmarking Alliance (WBA) and containing information about the 2000 most influential companies for the achievement of SDGs. Since our interest lies in the second SDG, contrast to hunger and nutrition improvement, we focused on a subset of SDG2000 called 'Food and Agriculture Benchmark', consisting of 350 leading companies in the nutrition sector. For each company, the data set includes information on the country of its headquarters, on its sector, and the score it obtained in 2021 over a set of indicators, either disaggregated and aggregated into four areas of performance: governance and strategy, environment, nutrition and social inclusion (16; 17). We integrated these variables with financial information retrieved from the annual reports, such as total assets, generally used as a proxy of companies' size, and the return on assets (ROA), as a proxy of profitability. Since annual reports were not available for all firms in the sample, we restricted our focus to only the public ones, moving to a sample size of 211. In addition, we considered variables related to compliance with certain Global standards connected to sustainability reporting. Finally, to account for differences at national levels, we included two variables encoding the national SDG and SDG2 scores related to the country where each company has its headquarters: they allow us to link the national performance in terms of sustainability to the companies' results (11).

In our framework, the three variables of main interest are the presence of a corporate social responsibility committee in the board of each company (PCC), the social inclusion score (SI) and the nutrition score. PCC is a discrete score ranging from 0 to 2, varying by 0.5, where 0 denotes that the company does not disclose information about its governance linked to sustainability, while 2 means that the company has a committee dealing with sustainability issues and discloses information about its remuneration. Social inclusion score ranges from 0 to 18 and is obtained by summing 18 WBA indicators assessing several areas, such as respecting human rights, engaging with potentially negatively affected stakeholders, providing and promoting decent work and acting ethically. Finally, nutrition score measures the extent to which a company contributes to sustainable nutrition; it ranges from 0 to 30 and is assessed over six indicators.

DRAFT

## 3. Data analysis

To investigate the relationship between the variables of interest we first carried out an exploratory analysis and then performed the mediation analysis, which will be discussed in turn in the next sections.

### 3.1 Exploratory analysis

Almost half of the companies (49.3%) show an intermediate PCC score, while the rest is approximately equally distributed between low scores ($\leq 0.5$) and high scores ($\geq 1.5$). As regards inclusion and nutrition scores, many companies present the minimum score (0 for both variables), but none the maximum, and their distributions look positively skewed.

We graphically inspected the interplay of the target variables by means of boxplots. Figures 1 and 2 show how SI and nutrition scores vary as PCC increases, while Figure 3 represents the relationship between nutrition score and inclusion score, the latter categorised through its quartiles. It can be noticed that both the inclusion and nutrition scores increase as PCC increases, moreover nutrition score shows higher medians at the last quartiles of SI. It seems then that these variables are linked and, to further investigate their associations, we moved to the modelling phase, described in the next section.

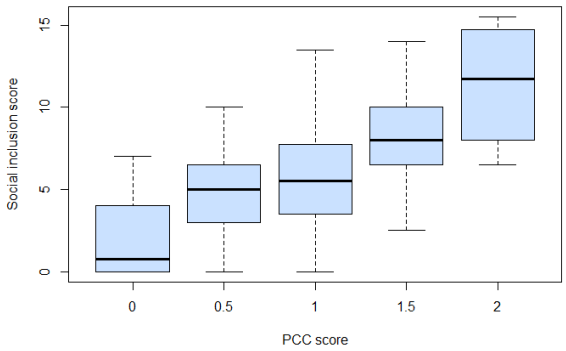



Figure 1: Boxplot of social inclusion score as a function of PCC.

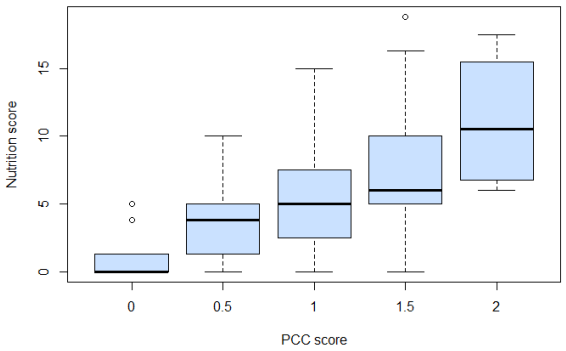



Figure 2: Boxplot of nutrition score as a function of PCC.

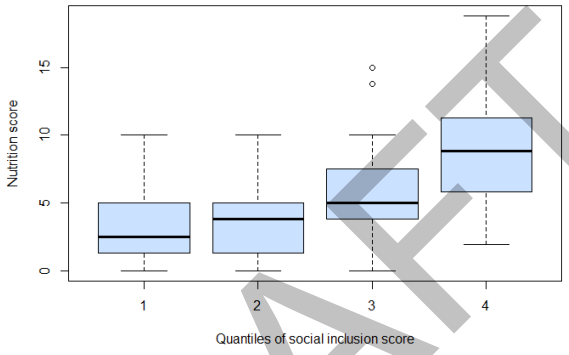



Figure 3: Boxplot of nutrition score as a function of social inclusion score quartiles.

DRAFT

## 3.2 Mediation models

Since both the response variables under investigation are scores, the Beta regression model seems an appropriate modeling option. We transformed the discrete scores into continuous variables ranging from 0 to 1 by dividing them by the maximum value they can take and then using the transformation proposed by (12) to constrain extreme values to fall in (0, 1). For each outcome, we started from a complex model including all variables listed in Table 1 and progressively reduced the number of variables through a stepwise selection procedure based on AIC. The variables included in each model are shown in the last column of Table 1. In both models we used the loglog link function. Results are shown in Table 2.

The disclosed presence of a corporate social responsibility committee is positively and significantly associated with SI, but it is only slightly significant in the outcome model. Larger companies have larger inclusion scores but lower nutrition scores, and the same pattern holds true for ROA, which is however only slightly significant in the mediator model. Adopting sustainability reporting standards is linked to higher inclusion scores, while, among the six industrial sectors only two are significantly associated to SI. The overall and the SDG2 scores at the national level have significant effects on SI, positive and negative respectively. As regards the outcome, only the national overall SDG score is slightly significant and has a negative sign. Companies' industrial sector was not included in the model, since it was not significant, while higher environmental and social inclusion scores are correlated to a higher nutrition score.

These results suggest that SI may act as a mediator in the relationship between PCC and nutrition score. In the traditional mediational setting described by (2) and (9), denoting the exposure by $X$, the mediator by $M$ and the outcome by $Y$, the mediator and the outcome model are assumed to be linear, so that the indirect effect is the product of the regression coefficients lying on the paths from $X$ to $M$ and from $M$ to $Y$. In our setting, both the mediator and the outcome models are characterised by a nonlinear

Table 1: For each variable we reported a short description, its type, and in which model it was included as a predictor after the model selection procedure.

| Variable name | Description | Variable type | Model |
|---|---|---|---|
| PCC | Governance and accountability for sustainable development | Numeric discrete | mediator, outcome |
| SI | Social inclusion score | Numeric discrete | outcome |
| Nutrition score | Nutrition | Numeric discrete | - |
| Region | Geographic macro-region of the country where the company has its headquarters | Six-level factor | - |
| Size | Logarithm of company's total assets | Numeric continuous | mediator, outcome |
| ROA | Company's return on assets | Numeric continuous | mediator, outcome |
| SRS | Number of international standards concerning sustainable disclosure to which a company is subject to | Numeric discrete | mediator |
| Environmental score | Score related to environmental disclosure | Numeric discrete | outcome |
| National_SDG_2020 | SDG score of the country where the company has its headquarters in 2020 | Numeric continuous | mediator, outcome |
| National_SDG2_2020 | Nutrition score of the country where the company has its headquarters in 2020 | Numeric continuous | mediator, outcome |
| Industrial sector | Sector of companies among Agricultural inputs, Agricultural product and commodities, Animal proteins, Food and beverage manufacturers processors, Food retailers and Restaurant and Food service | Six dummy variables | mediator |

DRAFT

link function, thus deriving an expression for the indirect effect is not straightforward. In the associational framework, there are not many examples of mediation analysis in nonlinear settings. (9) discusses the case of binary outcomes and how to estimate the indirect effect through the product method, by rescaling standard errors. (13) focuses on settings with a binary mediator and a binary outcome, deriving an exact formula which allows one to disentangle the direct and indirect effect of the exposure on the outcome. (14) claims that an indirect effect can be considered as a variation in the outcome corresponding to a variation in the exposure via the mediator. In other words, the response variable is a composite function of $X$, since $Y$ depends on $M$ and $M$ is a function of $X$. Analytically, a variation can be expressed through derivatives and, since the expectation of $Y$ is a composite function, using the chain rule from calculus, the indirect effect of $X$ on $Y$ through $M$ can be expressed as

$$\text{Indirect effect} = \frac{\partial \hat{Y}}{\partial M}\frac{\partial \hat{M}}{\partial X}, \tag{1}$$

where $\hat{M} = \mathbb{E}[M|X, Z_M]$ and $\hat{Y} = \mathbb{E}[Y|X, M, Z_Y]$, with $Z_M$ and $Z_Y$ two (possibly coinciding) set of covariates for the mediator and the outcome models, respectively (7). When both models have identity link functions, the indirect effect is a product of coefficients, consistently with the path-analytic framework. In contrast, when at least one between the mediator and the outcome model is nonlinear, there is not a single indirect effect, common to all subjects in the sample, but 'conditional' indirect effects (7), which depend on the values of $X$ and, for some combinations of link functions, on those of $M$. Thus, if $X$ is continuous, it is possible to draw a curve of indirect effects or, in alternative, one can select a set of values of interest and estimate the corresponding indirect effects, where $M$, if present in the formula, is not chosen at random, but assumes the values predicted by the mediator model at each selected $x$.

When $X$ has a discrete support, as in the current analysis, expressing a variation in terms of derivatives is pointless, thus we used finite differences instead. The discrete version of formula (1) is

$$\text{Indirect effect} = \frac{h_Y(x) - h_Y(x')}{k} \tag{2}$$

where $h_Y$ is the inverse of the link function of the outcome model, $x$ and $x'$ two values in the domain of

Table 2: Point estimates and standard errors of Beta regression models for social inclusion and nutrition score. Stars denote p-values: * if $p < 0.05$, ** if $p < 0.01$ and *** if $p < 0.001$.

| | **Social inclusion** | **Nutrition score** |
|---|---|---|
| Intercept | -3.076 (0.614) *** | -0.005 (0.495) |
| PCC | 0.396 (0.065) *** | 0.092 (0.054) . |
| $SI_{Disc}$ | - | 0.608 (0.175) *** |
| Size | 0.094 (0.021) *** | -0.040 (0.017) * |
| ROA | 0.001 (0.000) . | -0.0001 (0.000) * |
| SRS | 0.170 (0.042) *** | - |
| Env | - | 0.057 (0.007) *** |
| Agricultural input | 0.0981 (0.105) | - |
| Agricultural products and commodities | 0.168 (0.065) ** | - |
| Animal proteins | -0.171 (0.072) * | - |
| Food and beverage manufacturers processors | 0.115 (0.086) | - |
| Food retailers | -0.092 (0.077) | - |
| Restaurants and food service | -0.025 (0.112) | - |
| National_SDG_2020 | 0.038 (0.006) *** | -0.011 (0.005) * |
| National_SDG2_2020 | -0.039 (0.004) *** | 0.005 (0.004) |

Table 3: Indirect effects according to the values of PCC. Estimates, standard errors and confidence intervals were obtained through 1000 bootstrap replicates.

| **Values of PCC** | 0 - 0.5 | 0.5 -1 | 1 - 1.5 | 1.5 - 2 |
|---|---|---|---|---|
| **Indirect effect** | 0.041 (0.015) (0.009, 0.068) | 0.049 (0.019) (0.011, 0.085) | 0.055 (0.022) (0.013, 0.099) | 0.059 (0.024) (0.013, 0.107) |

the exposure and $k = x - x'$, assuming that $x > x'$[1]. We estimated the indirect effect for each couple of consecutive values of the exposure. Standards errors and confidence intervals were obtained through non-parametric bootstrap with $B = 1000$ resamples. Results are shown in Table 3. It can be noticed that all effects are significant and, as PCC increases, the variation between two adjacent values yields higher indirect effects; so, for example, the change from PCC = 0 to 0.5 induces an increase of 4.1% in nutrition score through the social inclusion score, while the change of PCC from 1.5 to 2 leads to a 5.9% increase in the nutrition score via SI.

## 4. Conclusions

Achieving goals of Agenda 2030 calls for efforts by all societal actors and companies can play an important role in the process. In this work, we focused on the second sustainable development goal, contrast to hunger and nutrition improvement, and analysed the determinants of nutrition scores using a sample of 211 companies considered as the most influential for fulfilling SDG2. Our results show that the presence of a corporate social responsibility committee within the company is slightly associated to a higher nutrition score and that this effect is mediated by the social inclusion score. Specifically, moving from a PCC score to the subsequent one leads to an increase in the nutrition score mediated by the social inclusion score, and this positive change becomes larger as the two compared values of PCC increase.

[1]The function $h_Y$ depends also on $M$, but, as discussed, the mediator value depends itself on the value of the exposure.

DRAFT

This suggests that companies more transparent in their board commitment to sustainability and more inclusive in their reporting practices also provide evidence of achieving better nutrition conditions for everyone.

## References


[1] Adams, C. A.: Conceptualising the contemporary corporate value creation process. Account. Audit. Account. J. **30**(4), 906–931 (2017)

[2] Baron, R. M. and Kenny, D. A.: The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. J. Personal. Soc. Psychol. **51**(6), 1173–1182 (1986)

[3] Bebbington, J., Unerman, J.: Advancing research into accounting and the UN Sustainable Development Goals. Account. Audit. Account. J. **33**(7), 1657–1670 (2020)

[4] Bellucci, M., Simoni, L., Acuti, D., Manetti, G.: Stakeholder engagement and dialogic accounting. Account. Audit. Account. J. **32**(5), 1467–1499 (2019)

[5] Brown, J.: Democracy, sustainability and dialogic accounting technologies: Taking pluralism seriously. Crit. Persp. Account. **20**(3), 313–342 (2009)

[6] Di Vaio, A., Varriale, L., Di Gregorio, A., Adomako, S.: Corporate social performance and non-financial reporting in the cruise industry: Paving the way towards UN Agenda 2030. Corp. Soc. Responsib. Environ. Manag. **29**(6), 1931–1953 (2022)

[7] Geldhof, G. J., Anthony, K. P., Selig, J. P., Mendez-Luck, C. A.: Accommodating binary and count variables in mediation: A case for conditional indirect effects. Int. J. Behav. Develop. **42**(2), 300–308 (2018)

[8] Gray, R.: Is accounting for sustainability actually accounting for sustainability...and how would we know? An exploration of narratives of organisations and the planet Account. Organiz. Soc., **35**(1), 47–62 (2010)

[9] MacKinnon, D. P.: Introduction to Statistical Mediation Analysis. Taylor and Francis Group, New York (2008)

[10] Milne, M. J.: Phantasmagoria, sustain-a-babbling in social and environmental reporting. In: Jack, L., Davison, J., Craig, R. (eds.) The Routledge companion to accounting communication, pp. 135–153. Routledge, London (2013)

[11] Sachs, J. D., Lafortune, G., Kroll, C., Fuller, G., Woelm, F. Sustainable Development Report 2022. From Crisis to Sustainable Development: the SDGs as Roadmap to 2030 and Beyond. Cambridge University Press, Cambridge (2022)

[12] Smithson, M., Verkuilen J.: A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. Psychol. Methods **11**(1), 54–71 (2006)

[13] Stanghellini, E., Doretti, M.: On marginal and conditional parameters in logistic regression models. Biom. **106**(3), 732–739 (2016)

[14] Stolzenberg, R. M.: The measurement and decomposition of causal effects in nonlinear and nonadditive models. Sociol. Methodol. **11**, 459–488 (1980)

[15] Tregidga, H., Milne, M., Kearins, K.: (Re)presenting 'sustainable organizations'. Account. Organ. Soc. **39**(6), 477–494 (2014)

[16] World Benchmarking Alliance. 2021 Food and Agriculture Benchmark - scoring guidelines (Tech. Rep.). Retrieved from https://www.worldbenchmarkingalliance.org/research/food-and-agriculture-methodology/

[17] World Benchmarking Alliance. Methodology for the 2021 Food and Agriculture Benchmark (Tech. Rep.). Retrieved from https://assets.worldbenchmarkingalliance.org/app/uploads/2021/09/2021-Food-and-Agriculture-Benchmark-scoring-guidelines.pdf


DRAFT

# A passing network-based performance indicator in football: evidence from UEFA Champions League 2016-2017

Riccardo Ievoli[a], Lucio Palazzo[b], and Giancarlo Ragozini[b]

[a]Department of Chemical, Pharmaceutical and Agricultural Sciences, University of Ferrara; riccardo.ievoli@unife.it
[b]Department of Political Sciences, University of Naples Federico II; lucio.palazzo@unina.it, giragoz@unina.it

DRAFT

**Abstract**

This work exploits the possibility to combine information retrieved from network analysis and the methodology to construct a performance indicator of the team play in football. The main contribution is to propose a unique index able to summarize the passing behaviour of a football team from a network perspective. The applicability of the proposed approach is shown through an empirical application involving 192 network from 96 matches of UEFA Champions League 2016-2017.

***Keywords:*** passing networks, performance indicator, network intensity, ball possession

## 1. Introduction

In the last decade, Network Analysis (NA) became a well-known and useful tool to visualize and analyse the football passes. At the team level, the seminal contribution of Grund (6) shown the existence of a relationship between some network-based variables and the football outcome, measured in terms of scored goals of two competing teams. Following this approach, a set of network summary measures has been used to model the probability of winning the match (7) or the football outcome measured in terms of scored goals or difference in goals (8). From a different perspective, a hierarchical clustering approach is also proposed to identify different tactics based on the community structure retrieved from NA (4).

Despite the availability of studies considering network indices and team strategies in football (1), the possibility to develop a comprehensive index of the passing behaviour (also viewed as the quality of team play) at the team level has not been fully explored in the literature, to the best of our knowledge. The aim of this paper is to exploit the methodology regarding composite indicators (CI's) (5) to propose an *ex post* performance indicator to compare and rank the passing behaviour of the football teams in a tournament. Nine variables at a team level are involved to construct the CI and four combinations of normalization and aggregation methods are compared using data from the Season 2016-2017 of UEFA Champions League.

The paper is organized as follows: Section 2. presents the considered network-related variables while the methodology used to construct the CI is illustrated in Section 3. Section 4. presents the main results of the empirical application concerning UEFA Champions League data. Finally, Section 5. concludes with some remarks and possible advances.

## 2. Passing Network Variables

The passing network distribution of a football team can be expressed as a weighted and directed network in the form of *adjacency matrix* P. The names of both columns and rows involve the football players of a team in a match. The generic cell $p_{ij}$ contains the number of passes given by the $i$-th player to the $j$-th player and, at the same time, the cell expresses the number of passes that $j$-th player receives from $i$-th one. Starting from this P and considering passing information usually freely available at the end of the match, we consider variables capturing the complexity of network topology, also having a meaningful interpretation for football (2; 8). These variables are described in the following list

***Pass Accuracy*:** it is defined as the sum of ratios between completed and attempted passes, representing a measure of the overall technical passing skills of a football team.

***Ball Possession*:** is the ratio between the time in which a team plays the ball and the total time of a match.

***Short, medium and long passes*:** these three variables depict the raw counts of short, medium and long passes, respectively, completed by a team in each match.

***Network Intensity*:** considering T as the actual ball possession time in minutes and $p_{ij}$ as the generic element of P, the expression: $\mathrm{T}^{-1} \sum_i \sum_j p_{ij}$, quantifying the level of passing speed of a team (6; 8).

***Network Diameter*:** it is the maximum length between two vertices of a graph, without taking into account the weights, expressing the ability to generate a variety of direct connections (7; 8).

***Reciprocity*:** it is computed as the proportion of mutual connections in a directed graph.. In football, it measures the overall ability of players to have mutual connections with each other, also evaluating the balance of a team in terms of passing directions (8).

***Median of Average Nearest Neighbors (MANN)*:** this measure is the median of an individual measure computed as follows: (2):

$$ANN_i = \frac{\sum_j (p_{ij} + p_{ji})(p_{j\cdot} + p_{\cdot j})}{2(p_{i\cdot} + p_{\cdot i})},$$

where $p_{i\cdot}$ and $p_{\cdot i}$ are, respectively, the row and column marginal sums of P. The team-level indicator MANN measures the cohesion in terms of passing behaviour (8).

DRAFT

## 3. Constructing The Performance Indicator

To construct a unique performance indicator to measure the quality of team play, four normalization (and aggregation) methods are considered. Let be $g = 1, \dots G$ the networks (in our case the football teams in the matches) and the previously mentioned variables $k = 1, \dots, K = 9$, the *min-max* method used to compute the normalized variables $I_{gk}$ can be summarized as follows:

$$I_{gk} = \frac{x_{gk} - \max_g(x_{gk})}{\max_g(x_{gk}) - \min_g(x_{gk})}, \tag{1}$$

where $x_{gk}$ is the value of the $k$-th variable collected in for the $g$ network, and $\max_g(x_{gk})$ and $\min_g(x_{gk})$ are the maximum and the minimum of the $k$-th variable, respectively, over the $G$ networks.

The second normalization method, i.e., the *z-score*, can be expressed as:

$$z_{gk} = \frac{x_{gk} - \overline{x}_k}{s_k}, \tag{2}$$

where $\overline{x}_k$ and $s_k$ stand for the average and the standard deviation of the $k$-th variable over the $G$ networks, respectively. For both normalization methods, the considered aggregating function to compute the CI is the arithmetic mean:

$$\mathrm{CI}_g^{mm} = \frac{1}{K} \sum_{k=1}^{K} I_{gk}; \qquad \mathrm{CI}_g^{z} = \frac{1}{K} \sum_{k=1}^{K} z_{gk}, \tag{3}$$

Table 1: Descriptive statistics of the passing network variable

| Variable | Mean | Median | Min | Max | S.D. | CV (%) |
|---|---|---|---|---|---|---|
| Pass Accuracy (%) | 84.38 | 85.69 | 65.99 | 96.09 | 5.80 | 6.87 |
| Ball Possession (%) | 50.00 | 50.00 | 28.00 | 72.00 | 9.94 | 19.89 |
| Short Passes (n) | 113.58 | 107.00 | 35.00 | 238.00 | 38.68 | 34.06 |
| Medium Passes (n) | 277.83 | 266.00 | 70.00 | 661.00 | 110.74 | 39.86 |
| Long Passes (n) | 41.03 | 38.00 | 18.00 | 88.00 | 12.85 | 31.32 |
| Network Intensity (n/T) | 13.80 | 13.73 | 7.65 | 18.78 | 2.13 | 15.41 |
| Network Diameter (n) | 5.82 | 6.00 | 4.00 | 10.00 | 1.37 | 23.62 |
| Reciprocity (%) | 68.56 | 68.93 | 42.59 | 85.00 | 8.44 | 12.31 |
| MANN | 20.28 | 19.74 | 14.22 | 31.06 | 3.07 | 15.13 |

S.D.: Standard Deviation

where $\text{CI}_g^{mm}$ is computed using the min-max method and $\text{CI}_g^{z}$ through the z-score method.

Moreover, the two illustrated methods are considered compensatory approaches (9) in the literature regarding CI's, suffering from the so called substitutability issue. Thus, we decide to apply two partially compensatory approaches that are able to take into account the possible unbalances arising between the original variables. The first method is called Mazziotta-Pareto index (MPI) (3) and the normalization step is based on a modification of the *z-score*:

$$z_{gk}^{\star} = 100 + \frac{x_{gk} - \overline{x}_k}{s_k} \cdot 10. \tag{4}$$

The last method is called Adjusted Mazziotta-Pareto index (AMPI) (9). With respect to MPI, AMPI allows proper comparison of CI's over time, and the normalization step is based on a modification of the *min-max*, as follows:

$$I_{gk}^{\star} = \frac{x_{gk} - max_k(x_g)}{max_g(x_k) - min(x_{gk})} \cdot 60 + 70. \tag{5}$$

When the composite indicator has a positive direction, i.e., it increases in the case of positive changes in the phenomenon of interest (i.e., the passing behaviour at the team level), the considered aggregating functions to compute the CI's for MPI and AMPI are:

$$\text{CI}_g^{MPI} = \frac{1}{K}\sum_{k=1}^{K} I_{gk}^{\star} - S_{I_g^{\star}} \cdot CV_{I_g^{\star}}; \qquad \text{CI}_g^{AMPI} = \frac{1}{K}\sum_{k=1}^{K} z_{gk}^{\star} - S_{z_g^{\star}} \cdot CV_{z_g^{\star}}, \tag{6}$$

where $S_{(.)_g}$ and $CV_{(.)_g}$ are the standard deviation and the Coefficient of Variation (CV) of the $K$ normalized indicators for the network $g$ obtained using (4) or (5). In practice, these kind of CI's penalizes each statistical unit $g$ exploiting its horizontal variability, i.e., the variability of the $K$ normalized indicators.

We remark that in case of negative direction ("polarity") of one or more variables involved in the CI, i.e., a negative sign in the relationship with the phenomenon of interest, the expressions (1), (2), (4) and (5) should be properly modified.

## 4. Empirical Application

Data for the application are collected using freely available press kits from the official UEFA website[1] and include $G = 192$ passing networks retrieved for 32 European teams. The matches come from the first phase of the tournament, named Group Stage, in the Season 2016-2017. This phase involves 8 groups composed by 4 teams, for a total of 96 matches.

Firstly, the descriptive statistics of the $G$ networks are depicted in Table 1. The variability ranges from the 7% of the pass accuracy to the 40% of the completed passes of medium length. To confirm the positive directions of the variables with respect to the team quality of passes, an exploratory data

[1] www.uefa.com

DRAFT

analysis is carried out. First of all, correlation analysis using Spearman coefficients shows that all the observed correlations between variables are greater than 0. Secondly, a Principal Component Analysis (PCA) is performed to investigate the relationships between the $K = 9$ variables. The first component (or dimension) explains the 61% of the overall variability and all variables are positively correlated with them. In particular, this correlation is greater than 0.8 for both short and medium passes, network intensity, pass accuracy, and ball possession. The largest contributions to the first dimension can be registered for medium passes (16%), network intensity (15%), ball possession (14%) and accuracy (14%). Furthermore, all variables show a pairwise positive relationship with the winning of the match and a (pairwise) positive correlation with the number of scored goals. Thus, following a data-driven approach, the results of PCA confirm the positive direction of the $K = 9$ variables with respect to the phenomenon of interest.

The ranking of European Football teams in terms of passing behaviour is summarized in Table 2 that contains the best and the worst 10 networks associated to a team in a match, where the match number is depicted in brackets. All the methods identify the network of the German team Bayern Munich (in its fourth match against PSV Eindhoven with a final results 1-2 for the "Bavarians") as the best in terms of quality of passes. The same team, coached by Carlo Ancelotti (that succeeded Pep Guardiola) appears at least four times in the considered top 10 rankings. Another relevant performance is represented by the sixth match of Barcelona, against Borussia M'Gladbach, ended with the result of 4-0, followed by the performance of Borussia Dortmund in the match won 8-4 against Legia Warsaw. In addition, all the teams in the top 10 of each method, except Seville, were able to pass the Group Stage and were qualified to the Round of Sixteen. Regarding the worst team in terms of passing behaviour, the Russian team of Rostov appears at least two times in the four considered worst ten. In addition, the bad performance of the Dutch team PSV Eindhoven against Bayern Munich arises from this analysis.

Table 2: Performance indicator for the quality of passes: ranking of different CI's

| Rank | Team | $CI^{mm}$ | Team | $CI^{z}$ | Team | $CI^{MPI}$ | Team | $CI^{AMPI}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Bayern Munich (4) | 0.81 | Bayern Munich (4) | 1.84 | Bayern Munich (4) | 116.62 | Bayern Munich (4) | 116.22 |
| 2 | Barcelona (6) | 0.80 | Barcelona (6) | 1.78 | Bayern Munich (1) | 113.34 | Bayern Munich (1) | 111.26 |
| 3 | Bayern Munich (1) | 0.76 | Bayern Munich (1) | 1.62 | Bayern Munich (3) | 111.05 | Barcelona (6) | 106.94 |
| 4 | Paris Saint-Germain (2) | 0.74 | Paris Saint-Germain (2) | 1.56 | Borussia Dortmund (5) | 110.38 | Bayern Munich (6) | 106.91 |
| 5 | Bayern Munich (3) | 0.73 | Bayern Munich (3) | 1.45 | Barcelona (6) | 109.97 | Borussia Dortmund (5) | 102.65 |
| 6 | Seville (3) | 0.72 | Barcelona (1) | 1.39 | Bayern Munich (6) | 109.97 | Bayern Munich (3) | 101.91 |
| 7 | Barcelona (1) | 0.72 | Seville (3) | 1.38 | Juventus (2) | 108.71 | Paris Saint-Germain (3) | 101.88 |
| 8 | Bayern Munich (6) | 0.71 | Bayern Munich (6) | 1.34 | Manchester City (5) | 107.46 | Real Madrid (1) | 101.00 |
| 9 | Paris Saint-Germain (3) | 0.70 | Bayern Munich (5) | 1.28 | Barcelona (1) | 107.38 | Club Atletico Madrid (3) | 100.94 |
| 10 | Borussia Dortmund (5) | 0.70 | Borussia Dortmund (5) | 1.27 | Bayern Munich (5) | 107.17 | Juventus (6) | 99.97 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 183 | Bayer Leverkusen (4) | 0.20 | Bayer Leverkusen (4) | -1.22 | Dinamo Zagreb (4) | 83.98 | Celtic Glasgow (1) | 72.50 |
| 184 | CSKA Moskow (2) | 0.19 | CSKA Moskow (2) | -1.28 | PSV Eindhoven (4) | 83.43 | PSV Eindhoven (2) | 71.98 |
| 185 | CSKA Moskow (6) | 0.17 | CSKA Moskow (6) | -1.37 | Rostov (1) | 82.28 | Manchester City (4) | 71.45 |
| 186 | Dinamo Zagreb (3) | 0.17 | PSV Eindhoven (4) | -1.44 | Bayer Leverkusen (4) | 81.06 | PSV Eindhoven (4) | 71.35 |
| 187 | PSV Eindhoven (4) | 0.16 | Dinamo Zagreb (3) | -1.45 | Rostov (4) | 79.58 | Rostov (5) | 70.67 |
| 188 | Dinamo Zagreb (4) | 0.15 | Dinamo Zagreb (4) | -1.50 | Rostov (3) | 78.57 | Borussia M'Gladbach (6) | 70.55 |
| 189 | Rostov (1) | 0.15 | Rostov (1) | -1.52 | Dinamo Zagreb (2) | 77.52 | PSV Eindhoven (3) | 70.49 |
| 190 | Rostov (4) | 0.13 | Rostov (4) | -1.61 | Rostov (6) | 75.17 | Dinamo Zagreb (2) | 69.33 |
| 191 | Rostov (3) | 0.13 | Rostov (3) | -1.63 | Rostov (5) | 71.92 | Rostov (6) | 67.67 |
| 192 | Rostov (5) | 0.03 | Rostov (5) | -2.10 | Dinamo Zagreb (3) | 64.36 | Dinamo Zagreb (3) | 54.36 |

A comparison between the four methods is carried out through a correlation analysis using the Spearman coefficient and considering the mean absolute deviation between rankings. As expected, the correlation between compensatory methods (min-max and z-score) is practically equal to 1, while the MPI method results highly correlated with both compensatory methods exhibiting ranking correlation greater than 0.95. Moreover, the AMPI appears less correlated with the two compensatory methods, with Spearman coefficient equal to 0.88, and presents correlation equal to 0.92 with the other partially compensatory method (MPI). The mean absolute difference shows that changes in rank between min-max and z-score are equal to 1.4 position, on average. This difference increases when the compensatory methods are compared with the partially compensatory methods, reaching values approximately equal to 12 positions in both cases. The largest differences can be observed in the comparison with respect to the AMPI, with values ranging from 18 to 21 positions. Moreover, averaging the CI's by 32 teams in the sixth matches, all methods confirm the excellent performance of Bayern Munich, followed by Barcelona, Paris Saint-German, Real Madrid, Borussia Dortmund, Juventus and Naples, and the bad performance of teams such

DRAFT

as Rostov, Dinamo Zagreb and CSKA Moskov.

Table 3: Comparison of rankings in terms of different CI's

| Spearman Correlation Coefficient | | | | |
|---|---|---|---|---|
| Method | Min-Max | Z-score | MPI | AMPI |
| Min-Max | 1.000 | 0.999 | 0.953 | 0.882 |
| Z-score | 0.999 | 1.000 | 0.955 | 0.879 |
| MPI | 0.953 | 0.955 | 1.000 | 0.915 |
| AMPI | 0.882 | 0.879 | 0.915 | 1.000 |
| Mean absolute difference of rank | | | | |
| Method | Min-Max | Z-score | MPI | AMPI |
| Min-Max | 0.000 | 1.417 | 11.708 | 20.969 |
| Z-score | 1.417 | 0.000 | 11.510 | 21.156 |
| MPI | 11.708 | 11.510 | 0.000 | 18.260 |
| AMPI | 20.969 | 21.156 | 18.260 | 0.000 |

To show the usefulness of the proposed approach, a logistic regression is performed to model the probability of win the game. Four specifications of the model are compared using the four types of performance indicators. In addition, the following seven in-match variables are included in the model: a) number of shots on target b) number of corners c) number of fouls committed d) number of fouls suffered e) yellow cards f) red cards and g) distance covered in kilometres. Multicollinearity is checked through Variance Inflation Factor (VIF): all variables present values ranging between 1 and 1.3.

Coefficients of logistic regression models are depicted in Table 4. In line with previous studies (8; 7), shots on target appears the most relevant variable in terms of statistical significance. Moreover, while the other in-field covariates result not statistically different from zero in each model specification, the proposed indicator shows a statistically significant positive impact on the probability of winning the match. This relationship is weaker using the MPI ($p = 0.055$) and stronger using the AMPI ($p = 0.008$). Thus, conventional measures such as Bayesian Information Criterion (BIC), McFadden's Pseudo $R^2$, and in-sample accuracy show that the model including $CI^{AMPI}$ slightly outperforms the others. We remark that it is possible to run variable selection methods, such as the *stepwise* selection, to remove redundant variables. Indeed, the method based on the AIC (Akaike Information Criterion) helps to select only three variables in all of the model specifications. These variables are shots on target, the new performance indicator and number of corners.

DRAFT

Table 4: Results of Logistic Regressions

| | CI methods | | | |
|---|---|---|---|---|
| Variable | Min-Max | Z-score | MPI | AMPI |
| Intercept | -4.381 | -2.713 | -8.442 | -8.051 |
| Shots on Target | 0.376 (***) | 0.377 (***) | 0.388 (***) | 0.384 (***) |
| **New Indicator** | 3.743 (*) | 0.721 (*) | 0.056 (.) | 0.068 (**) |
| Corners | -0.100 | -0.099 | -0.091 | -0.103 |
| Fouls committed | -0.047 | -0.047 | -0.051 | -0.052 |
| Fouls suffered | 0.041 | 0.042 | 0.040 | 0.040 |
| Yellow cards | -0.074 | -0.076 | -0.087 | -0.089 |
| Red cards | -0.555 | -0.548 | -0.547 | -0.559 |
| Distance covered (km) | 0.009 | 0.009 | 0.012 | 0.005 |
| BIC | 232.740 | 233.020 | 235.617 | 231.837 |
| Pseudo $R^2$ | 0.253 | 0.252 | 0.242 | 0.257 |
| Accuracy (%) | 80.7 | 79.7 | 79.7 | 80.7 |

Statistical significance; (.): $p < 0.1$; (*): $p < 0.05$; (**): $p < 0.01$; (***): $p < 0.001$

## 5. Concluding Remarks

Measuring the passing behaviour of a team will be an interesting challenge for statisticians and sport scientists, even from a network perspective, to give a unique measure of the team play. Indeed, comprehensive and easily interpretable indicators such as the expected goals ("XG") are very used in football by practitioners, experts, journalists and trainers.

The present paper explored the possibility to construct a reasonable *ex post* performance indicator to measure the passing behaviour of a team also considering variables retrieved from NA. Four different normalization-aggregation methods have been considered allowing to rank teams in a tournament. Despite the differences arising between rankings obtained through compensatory (min-max and z-scores) and partially compensatory methods (MPI and AMPI), these comprehensive indicators appear to be statistically significant to model the probability of win the match in logistic regression models, even including very relevant information such as the shots on target.

Several advances will be carried out to improve the proposed approach. First of all, relevance of selected variables to measure the team play will be investigated using a larger set of matches, e.g., considering different tournaments (or championships) or several seasons of the same tournament. Secondly, other normalization methods and/o linear combinations of the variables should be implemented, e.g., fully non compensatory methods (such as the geometric mean or the Jevons index). Lastly, the assumption of equal weights for each variable can be relaxed using different weighting schemes. A first proposal can be the use of the PCA results in the aggregating function (i.e., the weighted average), considering the observed contributions of the variables (in percentage).

DRAFT

## References


[1] Caicedo-Parada, S., Lago-Peñas, C., Ortega-Toro, E.: Passing networks and tactical action in football: A systematic review. Int. J. Environ. Res. Public Health, **17**(18), 6649 (2020).

[2] Clemente, F. M., Martins, F. M. L., Mendes, R. S.: Social network analysis applied to team sports analysis. Springer (2016).

[3] De Muro, P., Mazziotta, M., Pareto, A. (2011). Composite indices of development and poverty: An application to MDGs. Soc. Indic. Res., **104**, 1–18.

[4] Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: An innovative method to cluster networks. Stat. Modelling, **19**(1), 28–54 (2019).

[5] Greco, S., Ishizaka, A., Tasiou, M., Torrisi, G.: On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. Soc. Indic. Res., **141**, 61–94 (2019).

[6] Grund, T. U.: Network structure and team performance: The case of English Premier League soccer teams. Soc. Networks, **34**(4), 682–690 (2012).

[7] Ievoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. Knowl. Based Syst., **222**, 106997 (2021).

[8] Ievoli, R., Gardini, A., Palazzo, L.: The role of passing network indicators in modeling football outcomes: an application using Bayesian hierarchical models. AStA Adv. Stat. Anal., **107**(1-2), 153–171 (2023).

[9] Mazziotta, M., Pareto, A: Measuring well-being over time: The adjusted Mazziotta-Pareto index versus other non-compensatory indices. Soc. Indic. Res., **136**, 967–976 (2018).

# An Importance Sampling Algorithm For Bayesian Logistic Regression with Independent Gaussian Scale Mixture Prior

Paolo Onorati[a] and Brunero Liseo[a]

[a]Sapienza University of Rome; p.onorati@uniroma1.it, brunero.liseo@uniroma1.it

**Abstract**

We propose an importance sampling algorithm for producing a posterior sample for quantities of interest in the logit model when the prior distribution on the $\beta$ coefficients follows a Student-$t$ prior with independent components. We also show that the proposed method allows one to easily compute the marginal density of the data, which facilitates model selection procedures through the Bayes factor. The algorithm then avoids the use of Markov Chain Monte Carlo methods, at least when the sample size is less than $500$.

***Keywords:*** Unified Skew-Normal Distributions, Bayes Factor, Kolmogorov Distribution, Weakly Informative Prior.

DRAFT

## 1. Introduction

Binary regression is among the most popular and routinely used statistical methods in applied science. Standard Bayesian approaches and off-the-shelf packages are today available, see for example the R suites `brms` or `rstanarms`. However, there is no general consensus on the choice of prior distributions and on how to select among different link functions.

A remarkable contribution in this direction has been provided by Durante (6) who shows that the SUN family of densities (2) can be used as a conjugate prior for the probit regression model, that is the posterior distribution of the regression coefficients in a probit setting still belongs to the SUN family. Durante (6) himself provides methods for efficiently sampling from the SUN distribution. The above methodology is particularly useful in the case when $p >> n$.

Following this line of research, we have introduced a larger class of distributions, namely the *perturbed SUN* (pSUN hereafter) family (9), which is obtained by replacing the two Gaussian laws appearing in the SUN definition with two scale mixtures of Gaussian distributions. This double scaling produces a larger class of densities which can be particularly useful in Bayesian binary regression.

However, the use of scale mixtures leads to some increasing in computational complexity; indeed, for a SUN distribution it is possible to obtain an exact posterior sample, through Botev (5) algorithm, as long as the sample size is not too large, that is no more than few hundreds of observations, say $500$: see Durante (6) and Botev (5) for more details; nonetheless for a pSUN distribution the above method cannot be used and, in general, we must introduce a Gibbs sampler step.

On the other hand, the availability of an exact posterior simulator would lead to a computational gain in efficiency with respect to the use of MCMC methods. For this reason, we have introduced an importance sampling algorithm that works for a pSUN posterior distribution in the case of a logistic regression with a Gaussian prior for the $\beta$ coefficients. In (9) we present an example where the Gibbs

sampler takes 26 mins in order to get $10^4$ posterior draws while the importance sampling strategy only takes 33 secs for the same number of draws.

The gist of this paper is to construct an importance sampling algorithm that works for a larger class of priors; in particular, we concentrate on a Student-$t$ prior with independent components as suggested by Gelman et al. (8). However our approach can be easily adapted to any scale mixtures of Gaussian densities.

The rest of the paper is organized as follows: in section 2. we illustrate the construction of SUN and pSUN parametric families; then we introduce the Bayesian binary regression model with emphasis to the logit link; in section 3. we describe the importance sampling method for the Student-$t$ with independent components prior. Section 4. is devoted to some numerical examples.

## 2. Conjugate Prior for Bayesian Binary Regression Model

Arellano-Valle and Azzalini (2) introduce the Unified Skew-Normal (SUN) class of densities which includes many of the several proposals appeared in the literature. It is based on the introduction of a given number $m$ of latent variables; a $d$-dimensional random vector $\beta$ has a SUN distribution, i.e. $\beta \sim \mathrm{SUN}_{d,m}(\Gamma, \Delta, \gamma, \xi, \Omega)$, if its density function is

$$f_\beta(\beta) = \phi_\Omega(\beta - \xi) \frac{\Phi_{\Gamma - \Delta'\bar{\Omega}^{-1}\Delta}\left(\gamma + \Delta'\bar{\Omega}^{-1}\mathrm{diag}^{-\frac{1}{2}}(\Omega)(\beta - \xi)\right)}{\Phi_\Gamma(\gamma)},$$

where $\Gamma$ is a $m$-correlation matrix, $\Delta$ is $d \times m$ matrix, $\gamma \in \mathbb{R}^m$,, $\xi \in \mathbb{R}^d$, $\Omega$ is a $d$-covariance matrix, and $\bar{\Omega} = \mathrm{diag}^{-\frac{1}{2}}(\Omega)\,\Omega\,\mathrm{diag}^{-\frac{1}{2}}(\Omega)$. Here we use a slightly different parametrization, namely

$$\begin{aligned} \Theta &= \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right), \\ A &= \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\Delta'\bar{\Omega}^{-1}, \\ b &= \mathrm{diag}^{-\frac{1}{2}}\left(\Gamma - \Delta'\bar{\Omega}^{-1}\Delta\right)\gamma\,; \end{aligned} \tag{1}$$

in this case we denote $\beta \sim \mathrm{SUN}^*_{d,m}(\Theta, A, b, \xi, \Omega)$ and the following stochastic representation holds:

$$\beta = \xi + \mathrm{diag}^{\frac{1}{2}}(\Omega)Z \,|\, (T \le AZ + b)\,, \tag{2}$$

where $Z \sim N_d(0, \bar{\Omega})$ is independent of $T \sim N_m(0, \Theta)$. An extension of the above family is obtained by considering a scale mixture of Gaussian densities replacing $T$ and $Z$. We will use the following notation: let $\Sigma$ be a generic covariance matrix and let $Q(\cdot)$ be a CDF with positive values in the positive orthant; then we set

$$\begin{aligned} \phi_{\Sigma,Q}(u) &= \int_{\mathbb{R}^d} \prod_{i=1}^{d} \left(W_i^{-\frac{1}{2}}\right) \phi_\Sigma\left(\mathrm{diag}^{-\frac{1}{2}}(W)\,u\right) dQ(W)\,, \\ \Phi_{\Sigma,Q}(u) &= \int_{\mathbb{R}^d} \Phi_\Sigma\left(\mathrm{diag}^{-\frac{1}{2}}(W)\,u\right) dQ(W)\,, \\ \Psi_{Q_V,\Theta,A,Q_W,\bar{\Omega}}(b) &= \mathrm{P}(T - AZ \le b)\,, \\ T \sim \Phi_{\Theta,Q_V}(\cdot) &\perp\!\!\!\perp Z \sim \Phi_{\bar{\Omega},Q_W}(\cdot)\,. \end{aligned}$$

Let $V$ and $W$ be a $m$-dimensional and a $d$-dimensional random vectors, respectively, both defined on their positive orthant. With a little abuse of notation, for a generic $d$-dimensional vector $H$, let $\mathrm{diag}(H)$ be the $d$-dimensional diagonal matrix with same entries as $H$. Assume that

$$\begin{aligned} T|V \sim N_m(0, \Theta_V) \quad &\perp\!\!\!\perp \quad Z|W \sim N_d(0, \bar{\Omega}_W)\,, \\ V \sim Q_V(\cdot) \quad &\perp\!\!\!\perp \quad W \sim Q_W(\cdot)\,; \end{aligned} \tag{3}$$

DRAFT

where $\Theta_V = \text{diag}^{\frac{1}{2}}(V)\Theta\,\text{diag}^{\frac{1}{2}}(V)$ and $\bar{\Omega}_W = \text{diag}^{\frac{1}{2}}(W)\bar{\Omega}\,\text{diag}^{\frac{1}{2}}(W)$. Then the pSUN class of distributions is defined by the stochastic representation of equation (2) with the assumptions introduced in formula (3) on $Z$ and $T$; in this case we denote

$$\beta \sim \text{pSUN}_{d,m}\left(Q_V, \Theta, A, b, Q_W, \Omega, \xi\right)$$

and it is easy to see that the density is

$$f_\beta(\beta) = \phi_{\Omega,Q_W}(\beta-\xi)\frac{\Phi_{\Theta,Q_V}\left(A\,\text{diag}^{-\frac{1}{2}}(\Omega)(\beta-\xi)+b\right)}{\Psi_{Q_V,\Theta,A,Q_W,\bar{\Omega}}(b)}\,.$$

As we will see, the SUN and pSUN families play a central role in Bayesian inference for binary output. Consider a general version of the model as

$$\begin{aligned} Y_i|p_i &\overset{\text{ind}}{\sim} Be(p_i), \quad \forall i = 1,2,\ldots,n \\ p_i &= \Lambda(\eta(X_i))\,, \end{aligned}$$

where $\Lambda : \mathbb{R} \to [0,1]$ is a known link function, $\eta(\cdot)$ is a calibration function, and $X_i \in \mathbb{R}^p$ is the $i$-th row of the design matrix $X$. Typically $\Lambda(\cdot)$ is an univariate CDF of some random variable, symmetric about 0, and $\eta(x)$ takes the simple linear form, $x'\beta$; we refer to this case as the linear symmetric binary regression model (LSBR). Let $\Lambda_n(x) = \prod_{i=1}^n \Lambda(x_i)$, $x \in \mathbb{R}^n$ and $B_x = [2\,\text{diag}(x) - I_n]$ for $x \in \{0,1\}^n$, where $I_n$ is the identity matrix of size $n$; the likelihood function of a LSBR model can be written as

$$L\left(\beta; y\right) = \Lambda_n(B_y X\beta)\,.$$

Durante (6) shows that if $\beta$ has a SUN prior distribution and $\Lambda(x) = \Phi(x)$ then $\beta|Y$ is still SUN distributed; notice that the Gaussian distribution is a special case of the SUN so if $\beta$ is given a Gaussian prior, then the posterior is known to be a SUN distribution.

In order to extend the result of Durante (6) to the logit model, i.e. $\Lambda(x) = \left(1+\exp(-x)\right)^{-1}$, we show that if $\beta$ is given a pSUN prior and $\Lambda(x)$ is a CDF of a random variable that admits a representation in terms of scale mixture of centered Gaussian random variables, then $\beta|Y$ still belongs to the $psun$ family. The logit link satisfies this condition; see Andrews and Mallows (1) and Stefanski (11).

Here we concentrate on the case of independent Student-$t$ distributions; if $\beta_i \overset{i.i.d.}{\sim} T(\nu, \xi_i, \Omega_{i,i})\,,\ i = 1,2,\ldots,p$ then

$$\beta|Y \sim \text{pSUN}_{p,n}(Q^n_{V^*}, I_n, B_Y X\text{diag}^{\frac{1}{2}}(\Omega), B_Y X\xi, Q^p_{W^*}, \xi, \Omega)\,, \tag{4}$$

where $Q^n_{V^*}(\cdot)$ denotes the CDF of $n$ i.i.d. random variables distributed as 4 times the square of a Kolmogorov distribution, $Q^p_{W^*}(\cdot)$ denotes the CDF of $p$ i.i.d. random variables with $\text{Inv.Gamma}(\nu/2, \nu/2)$ and $\Omega$ is a diagonal matrix with components $\Omega_{i,i}\,,\ i = 1,2,\ldots,p$. See Onorati and Liseo (9) for details.

## 3. Importance Sampling Algorithm

In this section we illustrate an importance sampling algorithm in order to compute posterior quantities of interest from distribution (4). Notice that if one sets $V$ and $W$ equal to some constant, say $\widetilde{V}$ and $\widetilde{W}$, then

$$\beta|Y, \widetilde{V}, \widetilde{W} \sim \text{SUN}^*_{p,n}(I_n, \text{diag}^{-\frac{1}{2}}(\widetilde{V})B_Y X\text{diag}^{\frac{1}{2}}(\Omega)\text{diag}^{\frac{1}{2}}(\widetilde{W}), \text{diag}^{-\frac{1}{2}}(\widetilde{V})B_Y X\xi, \xi, \Omega_W)\,. \tag{5}$$

Therefore, we set $\widetilde{V}_i = \text{E}(V_i) = \pi^2/3\,,\ i = 1,2,\ldots,n$; this value is chosen since it represents the variance of the centered Gaussian random variable that minimizes the Kullback-Leibler divergence from

DRAFT

a standard logistic distribution; for $\widetilde{W}_i\,, i = 1, 2, \ldots, p$ we choose the median of an Inverse Gamma distributed random variable with density

$$\text{Inv.Gamma}\left(\frac{\nu+1}{2}, \frac{\nu + (\widehat{\beta}_{i,MaP} - \xi_i)^2/\Omega_{i,i}}{2}\right)\,, \tag{6}$$

Expression (6) denotes the distribution of $W_i$ when conditioned to $\beta_i$ set equal to its posterior mode.

However, the ratio between densities (4) and (5) is unbounded. To circumvent this problem, we replace the distribution (5) with a scale mixtures based on an Inverse Gamma distribution. More in detail

$$\begin{aligned}
&\zeta_1 = \xi + \sqrt{S}\text{diag}^{\frac{1}{2}}(\Omega)\text{diag}^{\frac{1}{2}}(\widetilde{W})\zeta_0\,,\\
&S \sim \text{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\right)\,,\\
&\zeta_0 \sim \text{SUN}^*_{p,n}(I_n, \text{diag}^{-\frac{1}{2}}(\widetilde{V})B_Y X\text{diag}^{\frac{1}{2}}(\Omega)\text{diag}^{\frac{1}{2}}(\widetilde{W}), \text{diag}^{-\frac{1}{2}}(\widetilde{V})B_Y X\xi, 0, \bar{\Omega})\,,\\
&S \perp\!\!\!\perp \zeta_0\,.
\end{aligned}$$

This way it is easy to show that the density ratio between (4) and the above importance density is bounded if some mild conditions hold; in particular, the existence of the MLE is a sufficient condition. The parameter value $\nu_0$ is set to 5; see Onorati and Liseo (9) for other details.

## 4. Examples

We evaluate the performance of the proposed importance algorithm; in particular we present a simulation study for evaluating the frequentist coverage of the resulting posterior sample. We also discuss a real data example, where we compare posterior means and credible intervals obtained with our method with those computed using the Polya-Gamma Gibbs sampler (10).

In all settings we use the prior described in Gelman et al. (8); i.e. $\nu = 1$, $\xi = 0$, $\Omega_{1,1} = 100$ and $\Omega_{i,i} = 6.25\,,\ i = 2, 3, \ldots, p$; in this case $\beta_1$ is the intercept and all the other covariates are centered and scaled in order to have a standard deviation equals to 0.5. In all settings we produce $10^5$ draws both in the Gibbs sampler and in the importance sampling set up.

### 4.1 Simulation Study

We first show the simulation study to evaluate the finite sample performance of the algorithm. Let $G = 3000$ be the number of iterations. We implement the following procedure:

for $g = 1, 2, \ldots, G$;

- set the first column of $X$ equals to a vector of ones; sample the other covariate values independently, i.e. $X_{ij}^{(g)} \overset{i.i.d.}{\sim} N(0,1)\,,\ i = 1, 2, \ldots, n\,,\ j = 2, 3, \ldots, p$;
- center and scale each column, but the first, of $X^{(g)}$ in order to have a standard deviation equal to 0.5;
- sample $\beta^*_{(1,g)} \sim Ca(0, 10)$ and $\beta^*_{(i,g)} \overset{i.i.d.}{\sim} Ca(0, 2.5)\,,\ i = 2, 3, \ldots, p$;
- sample $Y_i^{(g)} \overset{ind}{\sim} Be\big(\Lambda(X_i'^{(g)}\beta^*_{(g)})\big)$;
- draw $10^5$ values through importance sampling;
- compute the empirical quantiles of level $\gamma \in \{1/10 \times j, j = 1, 2, \ldots, 9\}$;
  $\Longrightarrow$ evaluate the frequentist coverage comparing the quantiles with $\beta^*_{(g)}$.

We set the sample size $n = 100$ and $p = 10$. Table 1 reports the results.

### 4.2 Iris Dataset

We consider a real data analysis on the Iris dataset, (7). There are $n = 150$ observations with 4 covariates, i.e. length and width for sepal and petal of 3 species, namely "setosa", "versicolor" and "virginica". We

DRAFT

Table 1: Average Frequentist Coverage of the 10 Parameters

| **Theoretical** | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|
| **Empirical** | 10.93% | 21.05% | 30.94% | 40.41% | 49.84% | 59.2% | 68.76% | 78.44% | 88.52% |

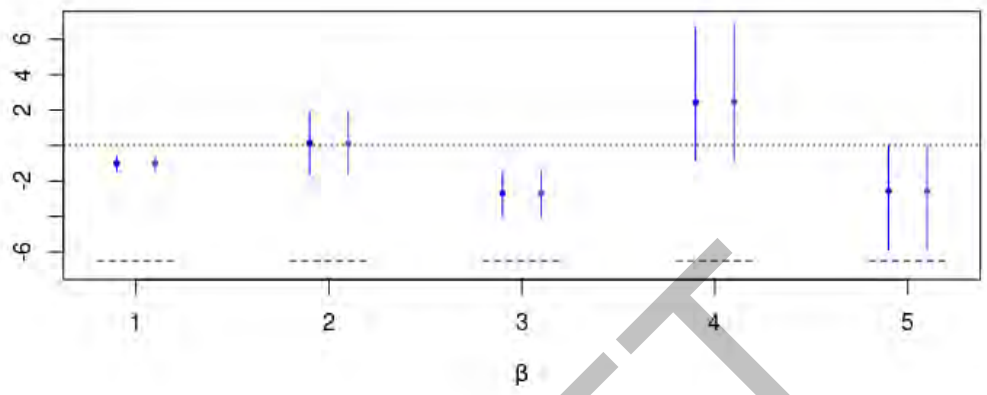



Figure 1: Iris dataset, logit model with independent Cauchy priors: Posterior means and 95% credible sets of the coefficients for the 4 covariates and the intercept $\beta_1$: Polya-Gamma (blue); importance sampling (purple)

DRAFT

set $y_i = 1$ if the output is "versicolor" and $y_i = 0$ otherwise. As before, we center and scale the covariates in order to have a standard deviation equal to 0.5; we also add the intercept so that the total number of parameters is $p = 5$.

We compute the posterior means and 95% credible intervals using both the proposed importance sampling and the Gibbs sampler of (10). Figure 1 compares the results of the two methods. One can notice that they are essentially similar.

We have also implemented a variable selection procedure based on the posterior probabilities of all possible subsets of covariates, adopting a uniform prior over the model space and compatible independent Cauchy priors within each model. Using the importance sampling algorithm, one is able to quickly compute, for each model, the normalizing constant of the posterior distribution. Assuming, as it is customary (4), that the intercept is included in all models, there are $2^{p-1} = 2^4 = 16$ possible combinations of covariates. See Onorati and Liseo (9) for more details about the implemented procedure for Bayes factor and computation of posterior inclusion probabilities.

Table 2 reports the results for each covariate; if a covariate is selected when its posterior inclusion probability is larger than 0.5 (3), we only select the predictor "Sepal Width", i.e. the third one.

Table 2: Posterior Probabilities

| Sepal Length | Sepal Width | Petal Length | Petal Width |
|---|---|---|---|
| 0.1549 | $> 0.9999$ | 0.2575 | 0.2710 |

## 5. Conclusions

We have proposed a new method for computing quantities of interest of the posterior distribution in a Bayesian logistic regression model using independent Student-$t$ prior. We have also shown how to compute the marginal density of data, opening the way to perform variable selection and model choice using Bayes factor.

Furthermore, when the sample size is not too large, one can completely avoid the use of Markov Chain simulations. Indeed, in logistic regression, MCMC techniques show a poor performance when the number of parameters is very large. On the other hand, with some refinements of the proposed algorithm, the importance sampling strategy can adequately manage this case. Here we have only discussed a specific Student $t$ prior. However, the algorithm can be easily adapted to any prior that can be expressed as a scale mixture of Gaussian densities, for instance in the case of a LASSO prior. Also, all the results can be easily adapted to other link functions, including the probit case.

## References


[1] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B*, 36:99–102, 1974.

[2] R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scand. J. Statist.*, 33(3):561–574, 2006.

[3] M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32: 870–897, 2004.

[4] M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550 – 1577, 2012.

[5] Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(1):125–148, 2017.

[6] D. Durante. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779, 2019.

[7] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

[8] A. Gelman, A. Jakulin, M. G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383, 2008.

[9] P. Onorati and B. Liseo. An extension of the unified skew-normal family of distributions and application to bayesian binary regression. *arXiv2209.03474*, 2022.

[10] N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.*, 108(504):1339–1349, 2013.

[11] L. A. Stefanski. A normal scale mixture representation of the logistic distribution. *Statist. Probab. Lett.*, 11(1):69–70, 1991.


DRAFT

# Binomial Extended Stochastic Block Model for Brain Networks

Valentina Ghidini, Sirio Legramanti, Raffaele Argiento

**Abstract** One of the main goals in the statistical analysis of brain networks is clustering nodes, which represent brain regions, based on their connectivity patterns. This can be assisted by node covariates, such as lobe memberships. In the present paper, we cluster the nodes of a weighted brain network through a binomial extension of the extended stochastic block model by Legramanti et al. (2022b), which was originally designed for binary networks.

**Key words:** Bayesian nonparametrics, Gibbs-type priors, Weighted networks

DRAFT

## 1 Introduction

One of the main goals when analyzing brain networks, in which each node corresponds to a brain region, is obtaining clusters of nodes that are homogeneous with respect to their connectivity patterns. Since each brain region belongs to a lobe (macro-area) and brain regions in the same lobe typically exhibit similar connectivity, lobe memberships can be leveraged to assist the above-mentioned clustering.

Clustering nodes is not limited to brain networks, and several algorithms are available for this task (e.g. Newman, 2004; Von Luxburg, 2007; Blondel et al., 2008), including some that exploit node covariates (e.g. Combe et al., 2015; Binkiewicz et al., 2017). However, algorithmic techniques have major drawbacks, such as the need to choose the number of clusters in advance, instead of learning it from data, and the absence of uncertainty quantification around the estimated partition. For

**Valentina Ghidini**
Bocconi University, Milan, e-mail: valentina.ghidini@unibocconi.it
**Sirio Legramanti**
University of Bergamo, e-mail: sirio.legramanti@unibg.it
**Raffaele Argiento**
University of Bergamo, e-mail: raffaele.argiento@unibg.it

these reasons, in this work we consider a model-based approach. In particular, we build on the extended stochastic block model (ESBM) by Legramanti et al. (2022b), which already proved effective on brain data (Legramanti et al., 2022a) but was originally implemented for binary networks. To fully exploit the information in the considered brain network, instead of dichotomizing its integer edge weights, we extend the ESBM by introducing a binomial likelihood, as detailed in the next section.

## 2 Binomial Extended Stochastic Block Model

Consider an undirected integer-weighted network with $V$ nodes and its $V \times V$ symmetric adjacency matrix $Y$. Each element $Y_{uv}$ of $Y$ is such that $Y_{uv} = Y_{vu} \in \{0, 1, \ldots, K\}$, and contains the weight of the edge linking nodes $u$ and $v$. Since here the goal is clustering, self-loops are neglected and, consequently, $Y_{vv}$ is set to 0 for each $v = 1, \ldots, V$. Moreover, let $\mathbf{x}_v = (x_{v1}, \ldots, x_{vp})$ be the $p$-dimensional row vector of the covariates associated to node $v$, and let $X$ be the $V \times p$ matrix obtained by stacking all the $\mathbf{x}_v$'s. Finally, let $\mathbf{z} = (z_1, \ldots, z_V) \in \{1, \ldots, H\}^V$ be the vector containing the node memberships associated to a partition of the $V$ nodes into $H$ mutually-exclusive clusters, such that $z_v = h$ if and only if node $v$ belongs to cluster $h$. We are now ready to define our binomial extended stochastic block model (bESBM):

$$\begin{aligned} (Y_{uv} | z_u = h, z_v = k, \phi_{hk}) &\overset{ind.}{\sim} \text{Binomial}(K, \phi_{hk}) && \text{for } 1 \leq u < v \leq V, \\ (\phi_{hk} | \mathbf{z}) &\overset{i.i.d.}{\sim} \text{Beta}(a, b) && \text{for } h, k = 1, \ldots, H, \\ (\mathbf{z} | X) &\sim \text{supervised Gibbs-type}(\sigma, X). \end{aligned} \tag{1}$$

In words, each edge weight is modeled with a conditionally independent binomial whose probability depends solely on the cluster memberships of the involved nodes. Independent Beta priors with common positive hyperparameters $a$ and $b$ are placed on such binomial probabilities. Finally, cluster memberships are given a Gibbs-type prior with (discount) parameter $\sigma < 1$ and supervised by node covariates, i.e.

$$p(\mathbf{z} | X) \propto \mathscr{W}_{V,H} \prod_{h=1}^{H} (1 - \sigma)_{n_h - 1} g(X_h^*)^{\beta}, \tag{2}$$

where $n_h$ is the cardinality of cluster $h$, $g(\cdot)$ is the so-called cohesion function, and $X_h^*$ is the $n_h \times p$ matrix obtained by stacking the covariates of the nodes in cluster $h$, while $\{\mathscr{W}_{V,H} : 1 \leq H \leq V\}$ is a collection of non-negative weights such that $\mathscr{W}_{V,H} = (V - H\sigma)\mathscr{W}_{V+1,H} + \mathscr{W}_{V+1,H+1}$ and $\mathscr{W}_{1,1} = 1$. The parameter $\beta \geq 0$ allows to adjust the impact of the covariates on the partition process.

Several specifications of the cohesion $g(\cdot)$ have been proposed (e.g. Dahl, 2008; Müller et al., 2011; Page and Quintana, 2016); among these, we follow the recipe for categorical settings in Müller et al. (2011), motivated by the application in Section 4. In particular, considering the case in which each node attribute is a single categorical variable with $C$ levels, we choose a cohesion function that is proportional to the

DRAFT

marginal likelihood of a Dirichlet-multinomial model on the covariates:

$$g(X_h^*) = \frac{1}{\Gamma(n_h + \alpha_0)} \prod_{c=1}^{C} \Gamma(n_{hc} + \alpha_c), \tag{3}$$

where $n_{hc}$ is the number of nodes in cluster $h$ with covariate $c$, while $\alpha_c > 0$ ($c = 1, \ldots, C$) are the Dirichlet concentration parameters, and $\alpha_0 = \sum_{c=1}^{C} \alpha_c$ is their sum.

## 3 Posterior Inference

In this section, we outline a collapsed Gibbs sampler for the posterior distribution $p(\mathbf{z}|Y,X)$ of the bESBM in (1). First of all, the cluster-specific probabilities $\phi_{hk}$, which are not of direct interest here, are marginalized out from (1), yielding

$$\begin{aligned} p(Y|\mathbf{z}) = \prod_{h=1}^{H} \prod_{k=1}^{h-1} \Bigg\{ \prod_{u,v: z_u=h;\, z_v=k} \binom{K}{y_{uv}} \Bigg\} \frac{B(m_{hk}+a, Kn_h n_k - m_{hk} + b)}{B(a,b)} \cdot \\ \cdot \Bigg\{ \prod_{u<v:\, z_u=z_v=h} \binom{K}{y_{uv}} \Bigg\} \frac{B(m_{hh}+a, (K/2)(n_h-1)n_h - m_{hh} + b)}{B(a,b)}, \end{aligned} \tag{4}$$

where $m_{hk}$ is the sum of the weights of the edges connecting nodes in cluster $h$ and nodes in cluster $k$, while $m_{hh}$ is the sum of edge weights within cluster $h$. From (4), we derive a collapsed Gibbs sampler that, at every iteration, updates the cluster membership of each node according to its full-conditional distribution

$$p(z_v = h|\mathbf{z}^{(-v)}, X, Y) \propto p(z_v = h|\mathbf{z}^{(-v)}, X) \frac{p(Y|\mathbf{z}^{(-v)}, z_v = h)}{p(Y^{(-v)}|\mathbf{z}^{(-v)})}, \tag{5}$$

where the superscript $(-v)$ denotes quantities computed excluding node $v$. The final ratio in (5) is computed from (4) while, for a bESBM with the cohesion in (3),

$$p(z_v = h|\mathbf{z}^{(-v)}, X) \propto \begin{cases} [(n_{hx_v}^{(-v)} + \alpha_{x_v})/(n_h^{(-v)} + \alpha_0)]^{\beta} \mathscr{W}_{V,H^{(-v)}}(n_h^{(-v)} - \sigma), & \text{for } h = 1, \ldots, H^{(-v)}, \\ (\alpha_{x_v}/\alpha_0)^{\beta} \mathscr{W}_{V,H^{(-v)}+1}, & \text{for } h = H^{(-v)} + 1, \end{cases}$$

where $\mathscr{W}_{V,H}$ and $\sigma$ are determined by the Gibbs-type prior of choice; see, e.g., Legramanti et al. (2022b). In particular, for the application in Section 4, we will employ a Gnedin prior (Gnedin and Pitman, 2004), which corresponds to $\sigma = -1$ and $\mathscr{W}_{V,H} = (\gamma)_{V-H} \prod_{h=1}^{H} (h^2 - \gamma h) / \prod_{v=1}^{V} (v^2 + \gamma v)$.

DRAFT

## 4 Brain Network Application

The motivating dataset is the result of the `NKI1` pilot study of the Enhanced Nathan Kline Institute-Rockland Sample project[1], and contains brain imaging data for 20 patients. From this data, we obtain a binary network for each subject, with nodes representing the 68 brain regions in the Desikan atlas (Desikan et al., 2006) and binary edges denoting the presence of white matter fibers between two regions. Since we are not interested in subject-specific structures, we sum the 20 individual binary adjacency matrices obtaining a single integer-valued adjacency matrix $Y$, whose entries $Y_{uv} \in \{0,1,\dots,20\}$ count how many out of the 20 considered subjects show a connection between brain regions $u$ and $v$. The resulting matrix corresponds to a single integer-weighted network, and clearly motivates our binomial extension of the ESBM. Finally, the lobe membership of each brain region is leveraged as a node covariate. Figure 1 shows the resulting integer-weighted network and the corresponding adjacency matrix, with nodes colored according to lobe memberships.

We sample from the posterior distribution of $\mathbf{z}$ by running the Gibbs sampler described in Section 3 for 10'000 iterations, with a burn-in of 1'000. Within the bESBM in (1), we employ a Gnedin prior with $\gamma = 0.3$, corresponding to a prior expectation of 17 clusters, while for the cohesion function in (3) we set $\alpha_c = 1$ for $c = 1,\dots 6$ (the six lobes in the human brain). In Figure 2 we represent the posterior

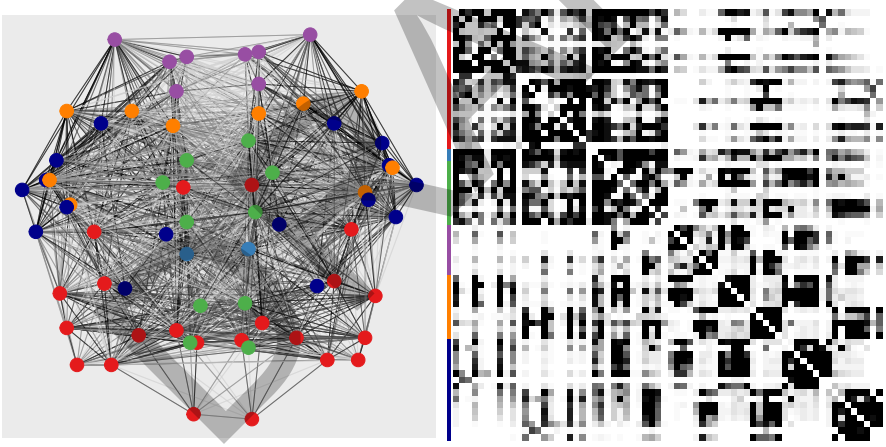


Fig. 1: Graphical representation (left) and adjacency matrix (right) of the considered integer-weighted brain network. Each node represents a brain region and is placed in its anatomical location. Node and side colors correspond to the human brain lobes: frontal (red), inter-hemispheric (light blue), limbic (green), occipital (purple), parietal (orange), or temporal (blue). The darkness of network edges and adjacency-matrix cells represents edge weights (the darker, the higher).

[1] `http://fcon_1000.projects.nitrc.org/indi/enhanced/`

DRAFT

point estimate of the node clustering (see Wade and Ghahramani, 2018) for $\beta = 0$ (no supervision), $\beta = 10$ (mild supervision) and $\beta = 1000$ (strong supervision). As expected, by increasing $\beta$, we obtain a partition that is more and more coherent with the node covariates, i.e. with the lobes illustrated in Figure 1.

## References

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008(10):1–12, 2008.

D. Combe, C. Largeron, M. Géry, and E. Egyed-Zsigmond. I-Louvain: An attributed graph clustering method. In *International Symposium on Intelligent Data Analysis*, pages 181–192. Springer, 2015.

D. B. Dahl. Distance-based probability distribution for set partitions with applications to Bayesian nonparametrics. *JSM Proceedings*, 2008.

R. Desikan, F. Ségonne, B. Fischl, B. Quinn, B. Dickerson, D. Blacker, R. Buckner, A. Dale, R. Maguire, B. Hyman, M. Albert, and R. Killiany. An automated labeling system for subdiving the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31:968–80, 2006.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov*, 325:83–102, 2004.

S. Legramanti, T. Rigon, and D. Durante. Bayesian clustering of brain regions via extended stochastic block models. *Book of short papers SIS*, pages 45–51, 2022a.

S. Legramanti, T. Rigon, D. Durante, and D. B. Dunson. Extended stochastic block models with application to criminal networks. *The Annals of Applied Statistics*, 16(4):2369–2395, 2022b.

P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

M. E. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):9, 2004.

G. L. Page and F. A. Quintana. Spatial product partition models. *Bayesian Analysis*, 11(1):265–298, 2016.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

S. Wade and Z. Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

DRAFT

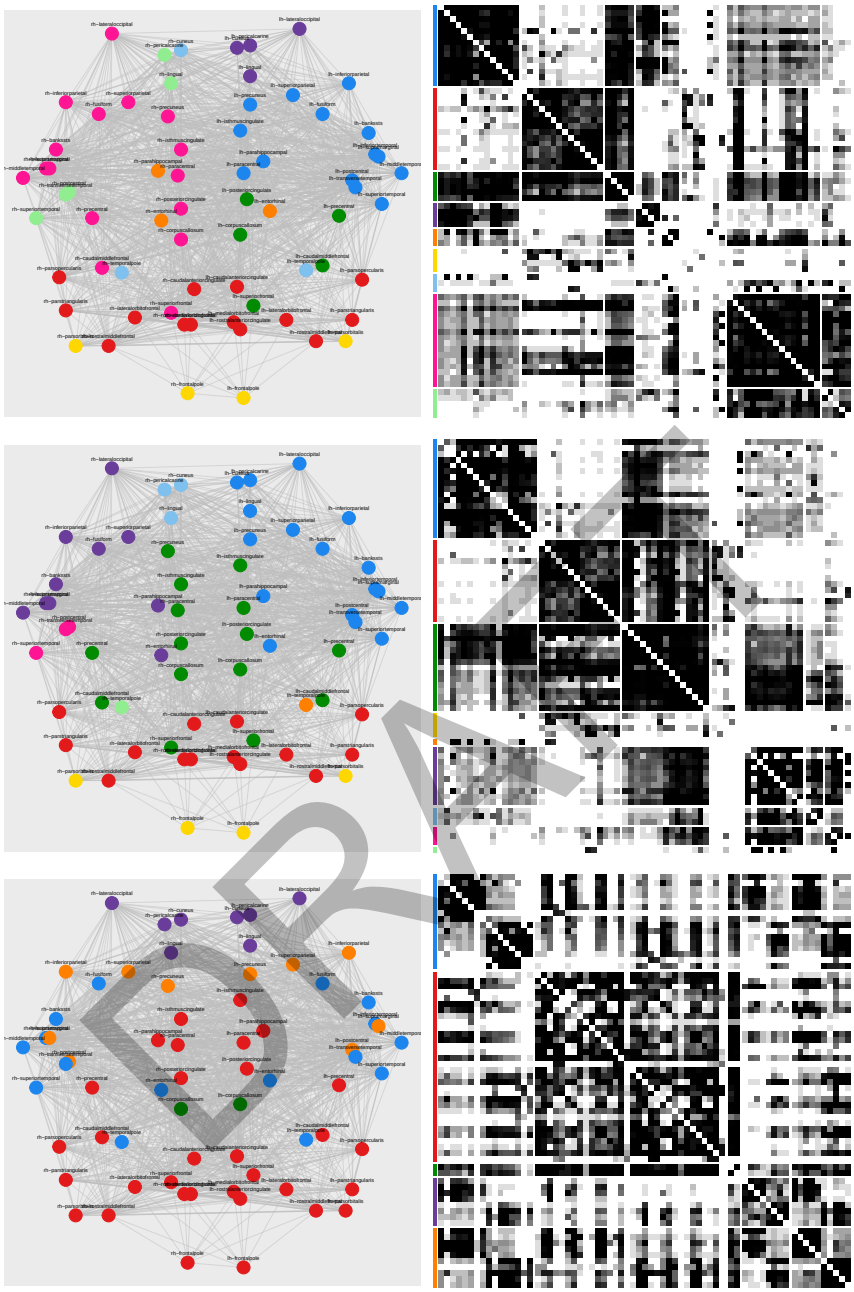


Fig. 2: Graphical representation (left) and adjacency matrix (right) of the considered brain network, after being clustered under the proposed binomial extended stochastic block model, endowed with a Gnedin prior supervised by lobe memberships, with increasing supervision: $\beta = 0$, i.e. no supervision (top); $\beta = 10$, i.e. balanced supervision (middle); $\beta = 1000$, i.e. strong supervision (bottom). Nodes are colored and adjacency matrix rows/columns are reordered according to the point posterior estimate of the node clustering.

DRAFT

# Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures

Giulia Capitoli[a], Simone Colombara[b], Alessia Cotroneo[b], Francesco De Caro[b], Riccardo Morandi[b], Chiara Schembri[b], Alfredo G. Zapiola[b], and Francesco Denti[c]

[a]Univerisity of Milan-Bicocca; giulia.capitoli@unimib.it

[b]Politecnico of Milan

[c]Università Cattolica del Sacro Cuore - Milan

DRAFT

**Abstract**

Mass spectrometry methods can record biomolecule abundance for a broad set of molecular masses given a sample of a specific biological tissue. In particular, the MALDI-MSI technique produces imaging data where, for each pixel, a mass spectrum is recorded. There is the urge to rely on suited statistical methods to model these data, fully addressing their morphological characteristics. Here, we investigate the use of Bayesian mixture models to segment these real biomedical images. We aim to detect groups of pixels that present similar patterns to extract interesting insights, such as anomalies that one cannot capture from the original pictures. This task is particularly challenging given the high dimensionality of the data and the spatial correlation among pixels. To account for the spatial nature of the dataset, we rely on Hidden Markov Random Fields.

***Keywords:*** Mass spectrometry, Bayesian mixture models, Potts model, Brain imaging.

## 1. Introduction

Mass spectrometry imaging (MSI) is an emerging technology capable of mapping various biomolecules within their native spatial context. In this work, we describe an application to spatial multi-omics data obtained via the state-of-the-art MALDI-MSI technology. MALDI-MSI uses a laser energy-absorbing matrix to create ions from large molecules with minimal fragmentation. In other words, this technology visualizes the distribution of molecules such as peptides, lipids, and glycans in a biological sample [10]. Here, we focus on the sequential MALDI-MSI of lipids on a single mouse brain formalin-fixed paraffin-embedded (FFPE) tissue section. A slice of this brain is partitioned into a grid of pixels: see panel (a) of Figure 1 for a low-resolution depiction of the biological sample we consider. Then, MALDI-MSI acquires a mass spectrum for each pixel. In particular, MALDI-MSI creates pixel-by-pixel spectra by measuring the mass-to-charge ($m/z$) values representing analytes of interest along the $x$-axis and the corresponding abundance along the $y$-axis [3]. Finally, a three-dimensional MSI dataset is obtained by arranging the mass spectra and the grid of coordinates for each pixel to be investigated in further analyses. Due to experimental limitations, the MALDI-MSI spectra can contain noise affecting subsequent statistical analyses. To mitigate this issue, a well-established biological preprocessing step has to be performed. In this contribution, we aim to segment the pixels of images obtained with the MALDI-MSI

technique into biologically meaningful clusters employing Bayesian mixture models. Our article proceeds as follows. In the next subsection, we introduce the dataset we analyze. In Section 2, we describe our modeling approach, while in Section 3, our results are presented and discussed. Finally, in Section 4, we discuss future directions and conclude.

## 1.1 Data description and statistical preprocessing

As we already mentioned, the raw dataset produced by the MALDI-MSI technology undergoes some initial preprocessing steps to filter out the noise. These steps are baseline correction, smoothing, normalization, spectra alignment, peak detection and extraction. In addition, the preprocessing needs to reduce the intra-sample variability and correct for analytical and instrumental variability following sample preparation to ensure accurate *m/z* localization. For more details, see, for example, [4]. The resulting dataset for our statistical analysis has the following characteristics. The lipid image of the biological tissue is analyzed with a raster of 50 $\mu m$ resulting in a total of about 18,000 laser shots (defining the pixels). In other words, we obtain thousands of spectra, each with its pixel coordinates ($s_1$,$s_2$). Recall that, for each pixel, the abundance of the lipids as a function of different values of molecular masses (*m/z*) is recorded, resulting in potentially hundreds of variables. Therefore, given the large dimensionality of the data, as the first step, we perform pixel-wise functional principal component analysis [9]. We then consider only the first functional principal component scores (fPCs) value for each pixel. We report the distributions of fPCs in panel (b) of Figure 1. The shape of the distribution suggests using Gaussian mixtures to segment the pixels of the image.

DRAFT

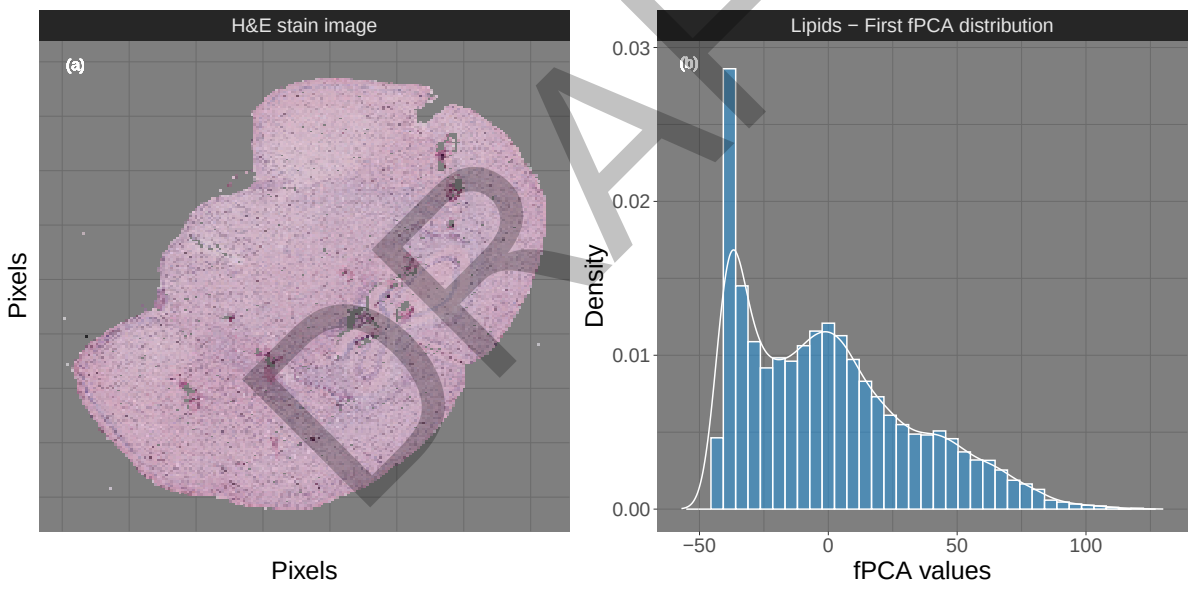



Figure 1: Panel (a): Hematoxylin and Eosin (H&E) staining image (low resolution). The reported mouse brain slice underwent digital scanning. Panel (b): The distribution of the first fPCs scores considered for our analysis. The shape suggests using a mixture of normal distributions to model the data.

## 2. Gaussian Mixtures and Potts Models

Here, we describe the model-based clustering approach we adopt for the preprocessed data presented in the previous section. We estimate clustering solutions via mixture models. Consider the vector $\boldsymbol{y} = (y_1, \ldots, y_n)$, where $y_i \in \mathbb{R}$ denotes the value of the first fPCs assumed by $i$-th pixel. Moreover, denote with $\mathcal{N}_i$ the set of all the neighboring (adjacent) pixels to the $i$-th one. Assuming the different pixels are

fully exchangeable (i.e., ignoring any spatial relation), one can specify the following basic, univariate Gaussian mixture likelihood with a fixed number of components $K$:

$$p(y_i \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} w_k \phi(\mu_k, \sigma_k^2), \qquad i = 1, \ldots, n, \tag{1}$$

where $\phi(\mu_k, \sigma_k^2)$ denotes a Normal density with mean $\mu_k$ and variance $\sigma_k^2$ and $\boldsymbol{w} = (w_1, \ldots, w_K)$ is the collection of mixture weights. To complete our model specification in a Bayesian setting, exploiting conjugacy, we adopt a Dirichlet prior for the mixture weights and Normal and Inverse Gamma priors for the mixture components' means and variances, respectively. In formulas, $\boldsymbol{w} \sim Dir_K(\boldsymbol{\lambda})$ and for $1, \ldots, K$, we assume $\mu_k \sim Normal(m, s^2)$ and $\sigma_k^2 \sim InvGamma(a, b)$. As customary with mixtures, we can augment the likelihood specification (1) by adding a set of $n$ latent membership labels $\{z_i\}_{i=1}^n$. Each membership label has a discrete distribution with support over $1, \ldots, K$, and $z_i = k$ implies that the $i$-th observation has been assigned to the $k$-th cluster. The augmented likelihood can be expressed as, for $i = 1, \ldots, n$:

$$y_i \mid z_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim Normal(\mu_{z_i}, \sigma_{z_i}^2), \qquad p(z_i|\boldsymbol{w}) = \sum_{k=1}^{K} w_k \, \delta_k(z_i), \tag{2}$$

where $\delta_i(j)$ is the delta function, equal to 1 when $i = j$ and 0 otherwise. Note how $y_i$ only depends on its correspondent latent membership $z_i$. It becomes evident how the addition of the latent variables considerably simplifies inference, usually carried out via MCMC techniques. We apply model (2) to set a benchmarking result.

Although the Bayesian Gaussian mixture model (GMM) presents a fully probabilistic extension to the classical $K$-means algorithm, the above specification disregards the available spatial information. Intuitively, especially when performing image segmentation, we expect neighboring pixels to have a higher probability of belonging to the same cluster. To introduce such spatial structure in the model, we can rely on Hidden Markov Random Fields. In particular, following [1; 2], we can defined a Potts model by modifying the distribution of the membership labels, introducing dependence of each pixel $i$ on the sets of its neighbors $\mathcal{N}_i$. We assume a Gibbs distribution for the vector $\boldsymbol{z}$, which can be specified via the following conditional probability statement:

$$p(z_i|\boldsymbol{z}_{-i}, \beta) \propto \exp(\beta \sum_{j \in \mathcal{N}_i} \delta_{z_i}(z_l)), \tag{3}$$

where $\boldsymbol{z}_{-i}$ denotes all the variables in $\boldsymbol{z}$ without the $i$-th, and $\beta$ is the *inverse temperature parameter*, which defines the strength of the spatial connection. Tuning and simulating the $\beta$ parameter is challenging due to the doubly-intractable nature of its full conditional distributions. More importantly, the Potts model undergoes a phase transition, switching from a disordered to an ordered state. The phase transition happens as $\beta$ exceeds a threshold $\beta^*$, also called *critical value*. The critical value for a regular 2D lattice is given by $\beta = \log(1 + \sqrt{K})$ [5; 8]. For simplicity, in this application, we assume a fixed value of inverse temperature, setting $\beta = \beta^*$ . To fit both the GMM and the Hidden Potts model (HPM), we rely on the R package `bayesImageS` [6].

## 3. MALDI-MSI Bayesian Segmentation Results

All the results described in this section are summarized in Figure 2. Panel (a) shows the fPCs values across the pixels. Spatial patterns are apparent, highlighting the hippocampal formation (light blue pixels), the white matter (high-intensity blue pixels), and a combination of gray matter and cerebral cortex (white pixels). Moreover, from a biological perspective, we expect to distinguish the Thalamus and Hypothalamus within the hippocampal formation. Led by this biological information, we set the number of mixture components $K$ equal to 5.

DRAFT

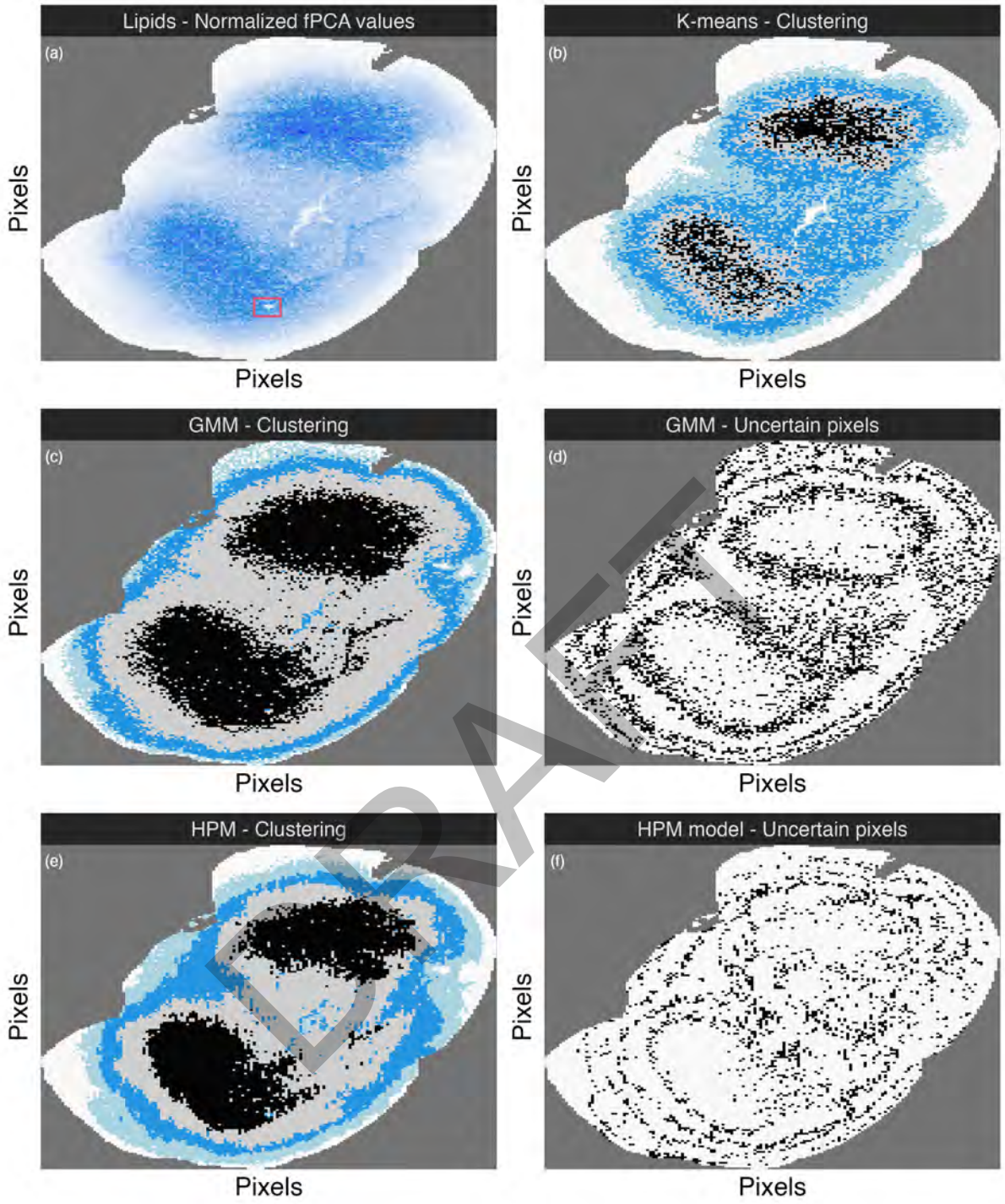



Figure 2: Each panel presents a pixelated image of the considered brain slice colored according to different results. (a) Distribution of the (normalized) first fPCs; (b) $K$-means clustering results; (c) GMM modal clustering results and (d) corresponding most uncertain pixels (e) HPM modal clustering results and (f) corresponding most uncertain pixels.

To benchmark our mixture models' results, we also fit a $K$-means algorithm on the first fPCs values. The resulting partition is reported in panel (b). The partition is fuzzy in the central parts of the slices, and no clear pattern is evident. We then fit both the GMM and the HPM. No label switching problem was detected: therefore, we estimate the clustering membership for each pixel as $\hat{c}_i = \arg\max_k f_{i,k}$, where $f_{i,k}$ is the proportion of iterations in which the MCMC allocated pixel $i$ in cluster $k$. In panels (c) and

DRAFT

(e), we report the results of the GMM and the HPM, respectively. The addition of spatial information helps the model detect clearer boundaries across the different regions of the picture.

Moreover, we compute a Gini-Simpson index for each pixel as $g_i = 1 - \sum_{k=1}^{K} f_{i,k}^2$, summarizing the uncertainty of the model in the allocation of a specific pixel. Indeed, a large value of $g_i$ implies that, across all the MCMC iterations, the pixels have been assigned to multiple clusters with similar probability, making $\hat{c}_i$ an unreliable estimate. In panels (d) and (f), we flag as uncertain (in black) all the pixels for which $g_i > 0.4$. These last two panels showed a discrete separation between the inter-brain formations (hippocampal white matter regions) and the external ones (gray matter). Since uncertain pixels correspond to cells that fall in the separation border between clusters, the separation enhances as we add spatial information. In fact, the number of pixels flagged as uncertain decreases from panels (d) to (f). The presence of uncertain pixels has a biological explanation. The laser diameter used by the MALDI-MSI in this analysis was 50 $\mu m$, while the average dimension of a single cell is $\approx 10\mu m$. Therefore, pixels at the border between two brain regions may contain cells of multiple natures that hinder their classifications.

Finally, the histological sample shows some anomalies (see, for example, the pixels highlighted by the red rectangle in panel (a) of Figure 2). These pixels correspond to pieces of tissue that detached from their original position and moved during sample preparation and analysis. Both the GMM and the HPM correctly classify these pixels exploiting their histological information. While retaining these anomalies, we can appreciate how the two Bayesian models eliminate sampling impurities better than the $K$-means.

## 4. Conclusions

The results presented in this contribution are promising and pave the way for many future directions to pursue. First, one can enhance the modeling aspect by allowing the estimation of a stochastic $\beta$ using, for example, the pseudo-likelihood approach or the exchange algorithm [7]. Second, one could extend the likelihood specification to handle multivariate measurements. This extension would allow the joint modeling of multiple functional principal components, increasing the amount of information preserved after our dimensionality reduction step. Alternatively, these findings also encourage the development of statistical methods that can be used to jointly model lipids, peptides, and glycans, uncovering hidden molecular patterns resulting from the relationship between these multiple molecular levels. Integrating the three molecular data sets could improve the separation across distinct histopathological regions of interest of the mouse brain sections. Lastly, one could focus on the original image of the biological tissue, modeling its RGB encoding to detect anomalies in its morphological characteristics patterns. We plan to explore these research avenues in the future.

DRAFT

## References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48:259–302, 1986.

[3] K. Boggio, E. Obasuyi, K. Sugino, S. Nelson, N. Agar, and J. Agar. Recent advances in single-cell maldi mass spectrometry imaging and potential clinical impact. *Expert Rev Proteomics*, 8(5): 591–604, 2011. doi:10.1586/epr.11.53.

[4] V. Denti, G. Capitoli, I. Piga, F. Clerici, L. Pagani, L. Criscuolo, G. Bindi, L. Principi, C. Chinello, G. Paglia, F. Magni, and A. Smith. Spatial Multiomics of Lipids, N-Glycans, and Tryptic Peptides on a Single FFPE Tissue Section. *Journal of Proteome Research*, 21(11):2798–2809, 2022. ISSN 15353907. doi: 10.1021/acs.jproteome.2c00601.

[5] M. Moores, G. Nicholls, A. Pettitt, and K. Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model. *Bayesian Analysis*, 15(1):1–27, 2020. ISSN 19316690. doi: 10.1214/18-BA1130.

[6] M. Moores, D. Feng, and K. Mengersen. bayesImageS: Bayesian Methods for Image Segmentation using a Potts Model. *R package (v0.6-1)*, 2021. URL https://cran.r-project.org/package=bayesImageS.

[7] I. Murray, Z. Ghahramani, and D. MacKay. MCMC for doubly-intractable distributions. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pages 359–366, 2006.

[8] R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1):106–109, 1952. ISSN 14698064. doi: 10.1017/S0305004100027419.

[9] J. Ramsay and B. Silverman. *Functional data analysis*. Springer New York, NY, New York, 2nd edition, 2005.

[10] T. Rohner, D. Staab, and M. Stoeckli. Maldi mass spectrometric imaging of biological tissue sections. *Mech Ageing Dev*, 126(1):177–185, 2005. doi:10.1016/j.mad.2004.09.032.

DRAFT

# Efficient expectation propagation for posterior approximation in high-dimensional probit models

Augusto Fasano[a], Niccolò Anceschi[b], Beatrice Franzolini[c], and Giovanni Rebaudo[a,d]

[a]Collegio Carlo Alberto, Turin, IT; augusto.fasano@carloalberto.org
[b]Duke University, Durham, USA; niccolo.anceschi@duke.edu
[c]Agency for Science, Technology and Research (A*STAR), Singapore, SG; beatricef@sics.a-star.edu.sg
[d]University of Turin, Turin, IT; giovanni.rebaudo@unito.it

DRAFT

**Abstract**

Bayesian binary regression is a prosperous area of research due to the computational challenges encountered by currently available methods either for high-dimensional settings or large datasets, or both. In the present work, we focus on the expectation propagation (EP) approximation of the posterior distribution in Bayesian probit regression under a multivariate Gaussian prior distribution. Adapting more general derivations in Anceschi et al. (2023), we show how to leverage results on the extended multivariate skew-normal distribution to derive an efficient implementation of the EP routine having a per-iteration cost that scales linearly in the number of covariates. This makes EP computationally feasible also in challenging high-dimensional settings, as shown in a detailed simulation study.

***Keywords:*** Probit Model, Expectation Propagation, Bayesian Inference, Extended Multivariate Skew-Normal Distribution

## 1. Introduction and literature review

The past few years have seen florid research in Bayesian inference for the probit model [4; 8] as well as its extensions to dynamic [7; 6] and multinomial [5; 9; 10] settings and beyond [1; 11]. This has been driven, among others, by computational challenges that may arise in high-dimensional settings. See [3] for an excellent review of Bayesian computations for binary regression. Here, we focus on the expectation propagation (EP) approximation of the posterior of the Bayesian probit model

$$\begin{aligned} y_i \mid \boldsymbol{\beta} &\overset{ind}{\sim} \text{BERN}\left(\Phi\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}\right)\right), \quad i = 1, \ldots, n, \\ \boldsymbol{\beta} &\sim \text{N}_p(\mathbf{0}, \nu^2 \mathbf{I}_p), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown vector of parameters, $\mathbf{x}_i \in \mathbb{R}^p$ is the covariate vector associated with observation $i$ and $\mathbf{I}_p$ denotes the identity matrix of dimension $p$. $\Phi(t)$ denotes instead the cumulative distribution function of a standard Gaussian random variable evaluated at $t$. Similarly, $\phi_p(\mathbf{t}, \mathbf{S})$ will denote the density of a $p$-variate Gaussian random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}$, evaluated at $\mathbf{t}$. [4] showed that the posterior distribution for model (1) is a unified skew-normal (SUN) and that, thanks to characterization properties of the SUN family, one can obtain i.i.d. samples from it via a linear combination of $p$-variate Gaussian samples and $n$-variate truncated Gaussian samples. As the computational bottleneck is represented by the truncated normal component, such i.i.d. sampler is well-suited for high-dimensional problems with small-to-moderate sample sizes but may become computationally hard for larger sample sizes. To overcome such limitation, [8] developed a partially-factorized variational (PFM-VB) approximation of the posterior distribution which, for any fixed $n$, converges to the true posterior distribution as $p$ diverges. Crucially, PFM-VB does not require dealing with any multivariate truncated Gaussian since the corresponding density component is replaced with a product of univariate truncated Gaussian densities, which do not represent a computational problem. This approximation has a pre-processing cost of $\mathcal{O}(pn \cdot \min\{p, n\})$ and cost-per-iteration of $\mathcal{O}(n \cdot \min\{p, n\})$, making it computationally tractable also in large $p$ and large $n$ settings. Empirically, the approximate posterior moments closely match the ones obtained via i.i.d. sampling for $p \geq 2n$. The possible over-shrinkage of the posterior moments towards zero for smaller $p$ motivates the investigation of efficient implementations of other approximation techniques that may be more accurate in those settings, like EP, at the price of a higher computational cost. Adapting more general results obtained for a broad class of models in [1], we show how the EP routine for posterior inference under the multivariate Gaussian prior in (1) can be implemented at per-iteration-cost of $\mathcal{O}(pn \cdot \min\{p, n\})$, which, although higher than the one of PFM-VB, improves over the cost $\mathcal{O}(p^2 n)$ reported in [3], leading to sensible computational advantages and making EP computationally feasible also in settings with $p$ of the order of tens of thousands. Considering the goodness of the EP approximation [1; 3], the possibility to extend the number of scenarios where it can be effectively implemented represents a major contribution to Bayesian binary regression computations.

DRAFT

## 2. Expectation propagation for the probit model

In this section, we present an implementation of EP for the probit model (1) which leverages results on multivariate extended skew-normal (SN) random variables (see [2]). Calling $\mathbf{y} = (y_1, \ldots, y_n)$, in EP we approximate $p(\boldsymbol{\beta} \mid \mathbf{y})$ with $q(\boldsymbol{\beta}) \propto \prod_{i=0}^n q_i(\boldsymbol{\beta})$, where $q_0(\boldsymbol{\beta}), \ldots, q_n(\boldsymbol{\beta})$ are probability density functions and, in particular, $q_0(\boldsymbol{\beta}) = p(\boldsymbol{\beta})$ and $q_i(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}\boldsymbol{\beta}^\intercal \mathbf{Q}_i \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_i\}$ for $i = 1, \ldots, n$. Hence, writing $q_0(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}\boldsymbol{\beta}^\intercal \mathbf{Q}_0 \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal \mathbf{r}_0\}$, with $\mathbf{r}_0 = \mathbf{0}$ and $\mathbf{Q}_0 = \nu^{-2}\mathbf{I}_p$, we immediately note that $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \mathbf{Q}^{-1}\mathbf{r}, \mathbf{Q}^{-1})$, where $\mathbf{r} = \sum_{i=0}^n \mathbf{r}_i$, $\mathbf{Q} = \sum_{i=0}^n \mathbf{Q}_i$.

EP proceeds by updating each site $i = 1, \ldots, n$ (we do not update the site of the prior), by iteratively matching the first two moments of the global approximation $q(\boldsymbol{\beta})$ and the hybrid distribution

$$h_i(\boldsymbol{\beta}) \propto p(y_i \mid \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}) = \Phi((2y_i - 1)\mathbf{x}_i^\intercal \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}). \tag{2}$$

To compute the moments of (2), instead of proceeding as [3], we can exploit the fact that some easy algebraic manipulations show that (2) is the kernel of a multivariate extended skew-normal distribution

$\text{SN}_p(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}_i, \tau_i)$ (see [2]), with

$$\boldsymbol{\xi}_i = \mathbf{Q}_{-i}^{-1}\mathbf{r}_{-i}, \quad \boldsymbol{\Omega}_i = \mathbf{Q}_{-i}^{-1},$$
$$\boldsymbol{\alpha}_i = (2y_i - 1)\boldsymbol{\omega}_i\mathbf{x}_i, \quad \tau_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1/2}\mathbf{x}_i^\intercal\boldsymbol{\xi}_i,$$

where $\mathbf{Q}_{-i} = \sum_{j\neq i}\mathbf{Q}_j$, $\mathbf{r}_{-i} = \sum_{j\neq i}\mathbf{r}_j$ and $\boldsymbol{\omega}_i = [\text{diag}\,(\boldsymbol{\Omega}_i)]^{1/2}$.

After noticing this, exploiting formulae (5.71) and (5.72) in [2], we can immediately obtain the first two moments of $h_i(\boldsymbol{\beta})$:

$$\boldsymbol{\mu}_{h_i} = \mathbb{E}_{h_i(\boldsymbol{\beta})}[\boldsymbol{\beta}] = \boldsymbol{\xi}_i + \zeta_1(\tau_i)s_i\boldsymbol{\Omega}_i\mathbf{x}_i$$
$$\boldsymbol{\Sigma}_{h_i} = \text{var}_{h_i(\boldsymbol{\beta})}[\boldsymbol{\beta}] = \boldsymbol{\Omega}_i + \zeta_2(\tau_i)s_i^2(\boldsymbol{\Omega}_i\mathbf{x}_i)(\boldsymbol{\Omega}_i\mathbf{x}_i)^\intercal,$$

where $s_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1/2}$, $\zeta_1(x) = \phi(x)/\Phi(x)$ and $\zeta_2(x) = -\zeta_1(x)^2 - x\zeta_1(x)$. Hence, when updating site $i$, the EP moment-matching condition implies that the updated quantities $\mathbf{r}_i^{\text{NEW}}$ and $\mathbf{Q}_i^{\text{NEW}}$ must be such that

$$\begin{cases} (\mathbf{Q}_{-i} + \mathbf{Q}_i^{\text{NEW}})^{-1}(\mathbf{r}_{-i} + \mathbf{r}_i^{\text{NEW}}) = \boldsymbol{\mu}_{h_i} \\ (\mathbf{Q}_{-i} + \mathbf{Q}_i^{\text{NEW}})^{-1} = \boldsymbol{\Sigma}_{h_i}, \end{cases}$$

from which it immediately follows

$$\begin{cases} \mathbf{r}_i^{\text{NEW}} = (\mathbf{Q}_{-i} + \mathbf{Q}_i^{\text{NEW}})\,\boldsymbol{\mu}_{h_i} - \mathbf{r}_{-i} \\ \mathbf{Q}_i^{\text{NEW}} = \boldsymbol{\Sigma}_{h_i}^{-1} - \mathbf{Q}_{-i}. \end{cases}$$

The direct computation of $\boldsymbol{\Sigma}_{h_i}^{-1}$ can be avoided since, by Woodbury's identity

$$\begin{aligned}
\mathbf{Q}_i^{\text{NEW}} &= \boldsymbol{\Omega}_i^{-1} - \zeta_2(\tau_i)s_i^2(1 + \zeta_2(\tau_i)s_i^2\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1}\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Omega}_i\mathbf{x}_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\boldsymbol{\Omega}_i^{-1} - \mathbf{Q}_{-i} \\
&= -\zeta_2(\tau_i)s_i^2(1 + \zeta_2(\tau_i)s_i^2\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1}\mathbf{x}_i\mathbf{x}_i^\intercal = -(\zeta_2(\tau_i)^{-1}s_i^{-2} + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1}\mathbf{x}_i\mathbf{x}_i^\intercal \\
&= -\frac{\zeta_2(\tau_i)}{1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i + \zeta_2(\tau_i)\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i^\intercal = k_i^{\text{NEW}}\mathbf{x}_i\mathbf{x}_i^\intercal,
\end{aligned}$$

with $k_i^{\text{NEW}} = -\zeta_2(\tau_i)/\,(1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i + \zeta_2(\tau_i)\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)$. Moreover,

$$\begin{aligned}
\mathbf{r}_i^{\text{NEW}} &= \mathbf{Q}_{-i}\boldsymbol{\mu}_{h_i} + \mathbf{Q}_i^{\text{NEW}}\boldsymbol{\mu}_{h_i} - \mathbf{r}_{-i} = \mathbf{Q}_{-i}\mathbf{Q}_{-i}^{-1}\mathbf{r}_{-i} + \zeta_1(\tau_i)s_i\mathbf{Q}_{-i}\boldsymbol{\Omega}_i\mathbf{x}_i + \mathbf{Q}_i^{\text{NEW}}\boldsymbol{\mu}_{h_i} - \mathbf{r}_{-i} \\
&= \zeta_1(\tau_i)s_i\mathbf{x}_i + \mathbf{Q}_i^{\text{NEW}}\boldsymbol{\mu}_{h_i} = \zeta_1(\tau_i)s_i\mathbf{x}_i + k_i^{\text{NEW}}\mathbf{x}_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{r}_{-i} + k_i^{\text{NEW}}\zeta_1(\tau_i)s_i\mathbf{x}_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i \\
&= [\zeta_1(\tau_i)s_i + k_i^{\text{NEW}}(\boldsymbol{\Omega}_i\mathbf{x}_i)^\intercal\mathbf{r}_{-i} + k_i^{\text{NEW}}\zeta_1(\tau_i)s_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i]\mathbf{x}_i = m_i^{\text{NEW}}\mathbf{x}_i,
\end{aligned}$$

DRAFT

---

**Algorithm 1:** Probit EP - $\mathcal{O}(p^2n)$ cost per iteration

---

**Initialization:** $\mathbf{Q} = \nu^{-2}\mathbf{I}_p$; $\mathbf{Q}^{-1} = \nu^2\mathbf{I}_p$; $\mathbf{r} = \mathbf{0}$; $k_i = 0$ and $m_i = 0$ for $i = 1, \ldots, n$.

**for** *t from 1 until convergence* **do**

  **for** *i from 1 to n* **do**

    $\mathbf{Q}_{-i} = \mathbf{Q} - k_i\mathbf{x}_i\mathbf{x}_i^\intercal$

    $\mathbf{r}_{-i} = \mathbf{r} - m_i\mathbf{x}_i$

    $\boldsymbol{\Omega}_i = \mathbf{Q}^{-1} + k_i/\,(1 - k_i\mathbf{x}_i^\intercal\mathbf{Q}^{-1}\mathbf{x}_i)\,(\mathbf{Q}^{-1}\mathbf{x}_i)\,(\mathbf{Q}^{-1}\mathbf{x}_i)^\intercal$

    $s_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)^{-1/2}$

    $\tau_i = s_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\,\mathbf{r}_{-i}$

    $k_i = -\zeta_2(\tau_i)/\,(1 + \mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i + \zeta_2(\tau_i)\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i)$

    $m_i = \zeta_1(\tau_i)s_i + k_i(\boldsymbol{\Omega}_i\mathbf{x}_i)^\intercal\mathbf{r}_{-i} + k_i\zeta_1(\tau_i)s_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i$

    $\mathbf{Q} = \mathbf{Q}_{-i} + k_i\mathbf{x}_i\mathbf{x}_i^\intercal$

    $\mathbf{r} = \mathbf{r}_{-i} + m_i\mathbf{x}_i$

    $\mathbf{Q}^{-1} = \boldsymbol{\Omega}_i + \zeta_2(\tau_i)s_i^2(\boldsymbol{\Omega}_i\mathbf{x}_i)(\boldsymbol{\Omega}_i\mathbf{x}_i)^\intercal$

**Output:** $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \mathbf{Q}^{-1}\mathbf{r}; \mathbf{Q}^{-1})$

---

where $m_i^{\text{NEW}} = \zeta_1(\tau_i)s_i + k_i^{\text{NEW}}(\boldsymbol{\Omega}_i\mathbf{x}_i)^\intercal\mathbf{r}_{-i} + k_i^{\text{NEW}}\zeta_1(\tau_i)s_i\mathbf{x}_i^\intercal\boldsymbol{\Omega}_i\mathbf{x}_i$. Hence, we can implement EP by storing only the scalar quantities $k_i$ and $m_i$, $i = 1, \ldots, n$. In practice, they are initialized to zero, so that the initial global approximation is the prior distribution. Combining the above results with Woodbury's identity, we obtain

$$\boldsymbol{\Omega}_i = \mathbf{Q}_{-i}^{-1} = (\mathbf{Q} - k_i\mathbf{x}_i\mathbf{x}_i^\intercal)^{-1} = \mathbf{Q}^{-1} + \frac{k_i}{1 - k_i\mathbf{x}_i^\intercal\mathbf{Q}^{-1}\mathbf{x}_i}\left(\mathbf{Q}^{-1}\mathbf{x}_i\right)\left(\mathbf{Q}^{-1}\mathbf{x}_i\right)^\intercal,$$

which can be computed avoiding explicit matrix inversions, since $\mathbf{Q}^{-1}$ is known from the beginning. Finally, the update of the inverse of the EP precision matrix is immediate as $(\mathbf{Q}^{\text{NEW}})^{-1} = (\mathbf{Q}_{-i} + \mathbf{Q}_i^{\text{NEW}})^{-1} = \boldsymbol{\Sigma}_{h_i}$.

Putting it all together, we obtain the EP implementation in Algorithm 1. Its core part coincides with the EP derivations presented in [3], and implemented in the `EPprobit` function in the `R` package `EPGLM`. However, we arrived at it by exploiting results on SNs that leverage more general derivations for a broader class of models presented in [1]. We also avoided the computation of the normalizing constants for the unnormalized densities, $Z_i$, $i = 1, \ldots, n$, which can be used for the computation of the approximate marginal likelihood, since the approximated posterior moments can be computed also without them. Algorithm 1 has per-iteration cost $\mathcal{O}(p^2n)$, which, although avoiding explicit $p \times p$ matrix inversions, might be impractical in high-dimensional settings. Adapting more general results presented in [1], we thus derive in Section in full detail an implementation of EP for the Bayesian probit model having per-iteration cost $\mathcal{O}(pn^2)$.

## 3. Efficient expectation propagation for large $p$ settings

The crucial part to obtain a per-iteration-cost that is linear in $p$ is to note that we can avoid handling $p \times p$ matrices, as, by close inspection of Algorithm 1, the whole EP routine can be written by working out directly the updates of the $p$-dimensional vectors $\mathbf{w}_i = \boldsymbol{\Omega}_i\mathbf{x}_i = \mathbf{Q}_{-i}^{-1}\mathbf{x}_i$ and $\mathbf{v}_i = \mathbf{Q}^{-1}\mathbf{x}_i$, $i = 1, \ldots, n$. As for the former, we have

$$\begin{aligned}\mathbf{w}_i = \mathbf{Q}_{-i}^{-1}\mathbf{x}_i &= (\mathbf{Q} - \mathbf{Q}_i)^{-1}\mathbf{x}_i = \mathbf{Q}^{-1}\mathbf{x}_i + (1 - k_i\mathbf{x}_i^\intercal\mathbf{Q}^{-1}\mathbf{x}_i)^{-1}k_i(\mathbf{Q}^{-1}\mathbf{x}_i)(\mathbf{Q}^{-1}\mathbf{x}_i)^\intercal\mathbf{x}_i \\ &= \mathbf{v}_i + k_i(1 - k_i\mathbf{x}_i^\intercal\mathbf{v}_i)^{-1}\mathbf{v}_i\mathbf{v}_i^\intercal\mathbf{x}_i = \left[1 + (1 - k_i\mathbf{x}_i^\intercal\mathbf{v}_i)^{-1}(k_i\mathbf{x}_i^\intercal\mathbf{v}_i)\right]\mathbf{v}_i = d_i\mathbf{v}_i,\end{aligned}$$

where $d_i = (1 - k_i\mathbf{x}_i^\intercal\mathbf{v}_i)^{-1}$. As for the $\mathbf{v}_i$'s, each time a site $i$ is updated $\mathbf{Q}$ changes and thus all the $\mathbf{v}_j$'s, $j = 1, \ldots, n$, should be modified accordingly as

$$\begin{aligned}\mathbf{v}_j^{\text{NEW}} &= (\mathbf{Q}^{\text{NEW}})^{-1}\mathbf{x}_j = (\mathbf{Q} - \mathbf{Q}_i + \mathbf{Q}_i^{\text{NEW}})^{-1}\mathbf{x}_j = [\mathbf{Q} + (k_i^{\text{NEW}} - k_i)\mathbf{x}_i\mathbf{x}_i^\intercal]^{-1}\mathbf{x}_j \\ &= [\mathbf{Q}^{-1} - (k_i^{\text{NEW}} - k_i)[1 + (k_i^{\text{NEW}} - k_i)\mathbf{x}_i^\intercal\mathbf{Q}^{-1}\mathbf{x}_i]^{-1}\mathbf{Q}^{-1}\mathbf{x}_i\mathbf{x}_i^\intercal\mathbf{Q}^{-1}]\mathbf{x}_j \\ &= \mathbf{Q}^{-1}\mathbf{x}_j - [(k_i^{\text{NEW}} - k_i)^{-1} + \mathbf{x}_i^\intercal\mathbf{v}_i]^{-1}\mathbf{v}_i\mathbf{x}_i^\intercal\mathbf{v}_j = \mathbf{v}_j - c_i(\mathbf{x}_i^\intercal\mathbf{v}_j)\mathbf{v}_i,\end{aligned}$$

where $c_i = (k_i^{\text{NEW}} - k_i)/(1 + (k_i^{\text{NEW}} - k_i)\mathbf{x}_i^\intercal\mathbf{v}_i)$. Instead of cycling over $j$, these updates can be performed in block by defining a $p \times n$ matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$. Accordingly, $\mathbf{V}^{\text{NEW}} = \mathbf{V} - c_i\mathbf{v}_i\mathbf{x}_i^\intercal\mathbf{V}$. This operation is the most expensive per site update, being of order $\mathcal{O}(pn)$. Accordingly, each EP iteration has cost $\mathcal{O}(pn^2)$. Contrarily to Algorithm 1, once the procedure has reached convergence we still need to calculate the inverse of the global precision matrix $\mathbf{Q}^{-1}$. The explicit calculation can be avoided as follows, obtaining a post-processing cost of $\mathcal{O}(p^2n)$. First, $\mathbf{Q} = \mathbf{Q}_0 + \sum_{i=1}^n k_i\mathbf{x}_i\mathbf{x}_i^\intercal = \nu^{-2}\mathbf{I}_p + \mathbf{X}^\intercal\mathbf{K}\mathbf{X}$ with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$ $\mathbf{K} = \text{diag}(k_1, \ldots, k_n)$. Calling $\boldsymbol{\Lambda} = (\mathbf{I}_n + \nu^2\mathbf{K}\mathbf{X}\mathbf{X}^\intercal)^{-1}$, so that, by Woodbury's identity, $\mathbf{Q}^{-1} = \nu^2\mathbf{I}_p - \nu^4\mathbf{X}^\intercal\boldsymbol{\Lambda}\mathbf{K}\mathbf{X}$, one obtains that $\mathbf{V} = \mathbf{Q}^{-1}\mathbf{X}^\intercal = \nu^2\mathbf{X}^\intercal\left[\mathbf{I}_n - \nu^2\boldsymbol{\Lambda}\mathbf{K}\mathbf{X}\mathbf{X}^\intercal\right] = \nu^2\mathbf{X}^\intercal\boldsymbol{\Lambda}\left[\boldsymbol{\Lambda}^{-1} - \nu^2\mathbf{K}\mathbf{X}\mathbf{X}^\intercal\right] = \nu^2\mathbf{X}^\intercal\boldsymbol{\Lambda}$ and thus $\mathbf{Q}^{-1} = \nu^2\mathbf{I}_p - \nu^2\mathbf{V}\mathbf{K}\mathbf{X}$. Notice that, if the interest is only in approximate posterior means and variances, this expression for $\mathbf{Q}^{-1}$ allows doing it at reduced post-processing cost of $\mathcal{O}(pn)$. The whole routine is summarized in Algorithm 2.

DRAFT

**Algorithm 2:** Efficient probit EP for large $p$ - $\mathcal{O}(p \cdot n^2)$ cost per iteration

**Initialization:** $\mathbf{r} = \mathbf{0}$; $k_i = 0$ and $m_i = 0$ for $i = 1, \dots, n$; $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n] = \nu^2 \mathbf{X}^\intercal$.

**for** *$t$ from 1 until convergence* **do**

  **for** *$i$ from 1 to $n$* **do**

    $\mathbf{w}_i = (1 - k_i \mathbf{x}_i^\intercal \mathbf{v}_i)^{-1} \mathbf{v}_i$

    $\mathbf{r}_{-i} = \mathbf{r} - m_i \mathbf{x}_i$

    $s_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal \mathbf{w}_i)^{-1/2}$

    $\tau_i = s_i \mathbf{w}_i^\intercal \, \mathbf{r}_{-i}$

    $k_i^{\text{NEW}} = -\zeta_2(\tau_i) / \left(1 + \mathbf{x}_i^\intercal \mathbf{w}_i + \zeta_2(\tau_i) \mathbf{x}_i^\intercal \mathbf{w}_i\right)$

    $m_i = \zeta_1(\tau_i) s_i + k_i^{\text{NEW}} \mathbf{w}_i^\intercal \mathbf{r}_{-i} + k_i^{\text{NEW}} \zeta_1(\tau_i) s_i \mathbf{x}_i^\intercal \mathbf{w}_i$

    $k_i = k_i^{\text{new}}$

    $\mathbf{r} = \mathbf{r}_{-i} + m_i \mathbf{x}_i$

    $\mathbf{V} = \mathbf{V} - \mathbf{v}_i \left[(k_i^{\text{NEW}} - k_i) / \left(1 + (k_i^{\text{NEW}} - k_i) \mathbf{x}_i^\intercal \mathbf{v}_i\right)\right] \mathbf{x}_i^\intercal \mathbf{V}$

$\mathbf{Q}^{-1} = \nu^2 \mathbf{I}_p - \nu^2 \mathbf{V} \mathbf{K} \mathbf{X}$

**Output:** $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \mathbf{Q}^{-1}\mathbf{r}; \mathbf{Q}^{-1})$

## 4. Simulation study

We conclude with a simulation study where probit regression is applied to multiple simulated datasets, with $n = 100$ and $p = 50, 100, 200, 400$ and $800$. We investigate the performances of EP when the efficient implementations presented in Algorithm 1 and Algorithm 2 are used when $p < n$ and $p \geq n$, respectively. Such implementation, denoted EP-EFF in the following, is compared with PFM-VB in terms of running time and quality of the approximation. The latter is measured by the median absolute difference between the approximate posterior means and standard deviations and the ones computed via 2000 i.i.d. samples, for $\nu^2 = 25$. The moderate sample size is taken so that the i.i.d. sampler is computationally efficient, but the approximate methods could be used in more challenging settings, as in all scenarios they both give almost immediate outputs. To show the computational gains with respect to standard EP implementations, we also compare the running time needed to obtain the EP approximation with the `R` function `EPprobit` from the package `EPGLM`, which implements the EP derivations reported in [3]. As it emerges from Table 1, EP-EFF leads to a dramatic reduction of the computational effort with respect to the standard `EPprobit` in high dimensions. This results in a drop of the running time by more than three orders of magnitude in the setting $p = 800$, with a computational gain increasing with $p$, as expected. The EP-EFF running times, although generally much lower than the ones of `EPprobit`, are still higher than the ones of PFM-VB in most cases. Nevertheless, if one looks at the quality of the approximation of the two posterior moments in Figure 1, EP-EFF gives consistently accurate approximations across different dimensions of $p$, while PFM-VB gets similar accuracy for $p \gtrsim 2n$. This shows the importance of developing efficient implementations for EP like the ones in this paper, so make it computationally feasible in challenging high-dimensional settings where routine implementations are impractical. Code can be found at https://github.com/augustofasano/EPprobit-SN.

Table 1: Running time, in seconds, to compute posterior means and standard deviations with the EP approximation as in Algorithms 1 and 2 (EP-EFF), with the EP approximation computed via the `R` function `EPprobit` (`EPprobit`) and with the PFM-VB approximation (PFM-VB) for probit regression with $n = 100$ and $\nu^2 = 25$.

| | | $p$ | | | | |
|---|---|---|---|---|---|---|
| | *Method* | 50 | 100 | 200 | 400 | 800 |
| Running time (seconds) | EP-EFF | 0.11 | 0.02 | 0.03 | 0.05 | 0.09 |
| | `EPprobit` | 0.07 | 0.42 | 3.18 | 24.36 | 140.24 |
| | PFM-VB | 0.11 | 0.06 | 0.01 | 0.01 | 0.01 |

DRAFT

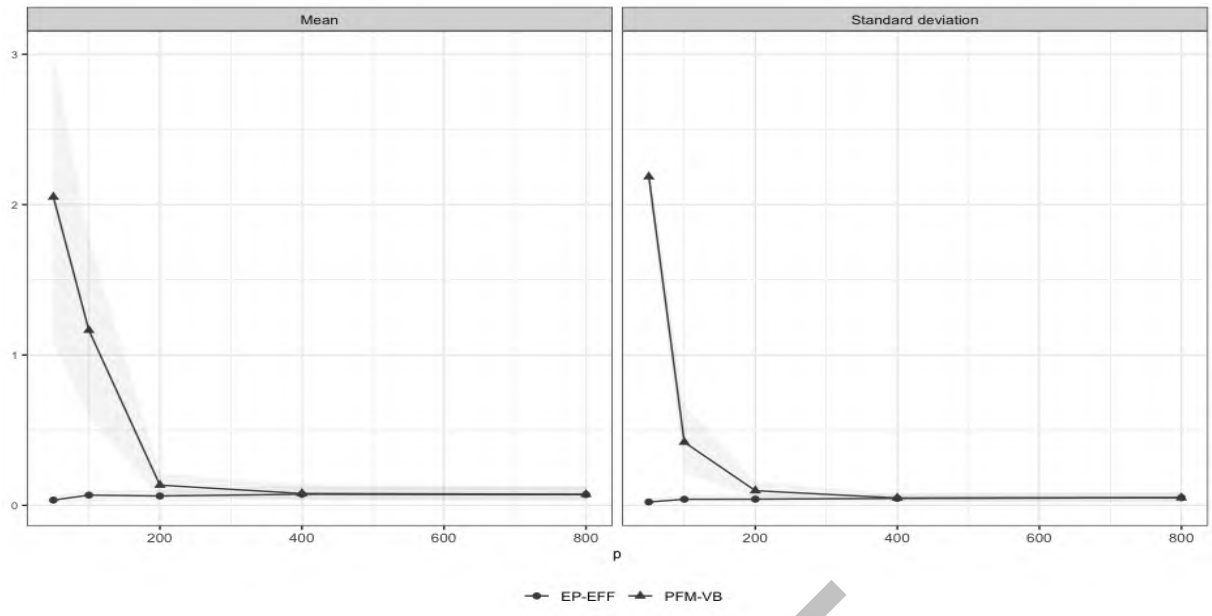



Figure 1: For varying $p$, median absolute difference between the $p$ posterior means and standard deviations resulting from 2000 i.i.d. samples and the ones arising from EP-EFF and PFM-VB for probit regression with $n = 100$ and $\nu^2 = 25$. Grey areas denote the first and third quartiles.

**Acknowledgments** The authors wish to thank D. Durante for carefully reading a preliminary version of this manuscript and providing insightful comments.

DRAFT

## References

[1] Anceschi, N., Fasano, A., Durante, D. and Zanella, G.: Bayesian conjugacy in probit, tobit, multinomial probit and extensions: a review and new results. *Journal of the American Statistical Association* [in press] (2023)

[2] Azzalini, A. and Capitanio, A.: *The Skew-Normal and Related Families*. Cambridge University Press (2014)

[3] Chopin, N. and Ridgway, J.: Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Statistical Science*, **32**, 64–87 (2017)

[4] Durante, D.: Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, **106**, 765–779 (2019)

[5] Fasano, A. and Durante, D.: A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research*, **23**, 1–16 (2022)

[6] Fasano, A. and Rebaudo, G.: Variational inference for the smoothing distribution in dynamic probit models. *Book of Short Papers - SIS 2021*, 1076-1081 (2021)

[7] Fasano, A., Rebaudo, G., Durante, D., and Petrone, S.: A closed-form filter for binary time series. *Statistics and Computing*, **31**, 1–20 (2021)

[8] Fasano, A., Durante, D. and Zanella, G.: Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika*, **109**, 901–919 (2022)

[9] Fasano, A., Rebaudo, G. and Anceschi, N.: Bayesian inference for the multinomial probit model under Gaussian prior distribution. *Book of Short Papers - SIS 2022*, 871–876 (2022)

[10] Loaiza-Maya, R. and Nibbering, D.: Fast variational Bayes methods for multinomial probit models. *Journal of Business & Economic Statistics* [online version] (2022)

[11] Loaiza-Maya, R., Smith, M. S., Nott, D. J. and Danaher, P. J.: Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics*, **230**, 229–362 (2022)

# Model-based clustering of non-stationary time series with common historical change times

Riccardo Corradin[a], Luca Danese[b], Wasiur KhudaBukhsh[a], and Andrea Ongaro[b]

[a]University of Nottingham; riccardo.corradin@nottingham.ac.uk,
wasiur.khudabukhsh@nottingham.ac.uk

[b]University of Milan-Bicocca; l.danese1@campus.unimib.it,
andrea.ongaro@unimib.it

DRAFT

**Abstract**

In this work we propose a way to cluster time series in which we assume as unique commonality that two time series belong to the same group if structural changes in their behaviour happen at the same times. Our object of interest is then a latent random partition of the time series generated by a discrete distribution whose weights are distributed as a Dirichlet distribution. To obtain a posterior estimate of the partition we propose a collapsed Gibbs sampler on which we implement an acceleration step based on the split-and-merge scheme that improves the mixing performances of the algorithm. We show with a simulation on synthetic data that the model is capable of detecting correctly the clustering structure.

***Keywords:*** Bayesian Statistics, Change Points, Model Based Clustering, Time Series Analysis

## 1. Introduction

In statistics an important role is played by methods for time series analysis. Since many phenomena evolve through time, it is crucial to take into account the temporal dependence of data. In this context an important and recently popular class of methods consists of detecting structural changes in the dynamic of the time series. This class of methods is called change points analysis. A change point corresponds to a time instant in which the data-generating distribution changes. If we assume that observations are generated by the same general law, then a change point will occur when the parameters of this law mutate. In literature methods for change points detection can be found both with frequentist and Bayesian frameworks.

Another important class of methods in statistics are those for clustering analysis. Grouping together observations allows us to find possible latent structures in the data. One of the main approaches to perform clustering consists in assuming that data within each cluster are generated by the same model, and then we can define to which group an observation belongs by studying its generating model.

In this work we propose a model based clustering method for time series based on the time occurrence of their change points. That is, two time series belong to the same group if they have changes that occur approximately in the same moment. To motivate our proposal, consider two time series that measure the profit of two companies, one that sells energy and another one whose production system relies on a significant amount of energy. A potential crisis in the energy supplies will influence both of them, but

in opposite terms - one negatively and the other positively. Therefore we can state that both companies have some sort of similarity since they both experience a shock from the same event at the same time, even though with opposite consequences.

The rest of the short paper is organized as follows. In Sect. 2. we present the model specification. Sect. 3. gives a glimpse on the computational aspects and the key quantities needed to perform posterior inference on the clustering structure. Finally, in Sect. 4. we show an application on synthetic data. Future directions with a possible applications on real data are deferred to Sect. 5.

## 2. Model

We denote with $\boldsymbol{y}_i = \{y_{i,1}, \dots y_{i,T}\}$ the generic $i$th observation that takes values in the discrete time range $1, \dots, T$. Then our sample is a vector of time series denoted by $\{\boldsymbol{y}_1, \dots \boldsymbol{y}_n\}$ where $\boldsymbol{y}_i = \{y_{i,l}\}_{l=1}^{T}$ and the generic $y_{i,l}$ takes values on $(\mathbb{Y}, \mathcal{Y})$, a Polish space endowed with its $\sigma$-algebra. We assume that each time series is indexed by a sequence of parameters $\theta_{i,1}, \dots, \theta_{i,T}$ associated to each observation. We further assume that two observations, say $y_{i,j}$ and $y_{i,l}$, belong to the same group $h$ if they share the same value of the parameter, i.e. $\theta_{i,j} = \theta_{i,l} = \theta^*_{i,h}$. A change point occurs whenever two subsequent observations belong to different groups. The sequence $\theta^*_{i,1} \dots \theta^*_{i,h} \dots \theta^*_{i,k_i}$ denotes the unique values of the parameters associated to the $k_i$ different regimes of the generic time series $\boldsymbol{y}_i$. We consider as distribution to model the sequence of parameters the product of two independent terms

$$\mathcal{L}(\theta_{i,1}, \dots, \theta_{i,T}) = \mathcal{L}(\rho_i)\mathcal{L}(\theta^*_{i,1}, \dots, \theta^*_{i,k_i}), \tag{1}$$

where $\mathcal{L}(\theta^*_{i,1}, \dots, \theta^*_{i,k_i})$ is the distribution of the sequence of unique values and $\mathcal{L}(\rho_i)$ is the distribution of the random order $\rho_i$. The random order $\rho_i$ is defined as a sequence of disjoint subsets of the observational time indices with an ordering constant and it is fully characterizing the change points. Specifically, we have that $\rho_i = \{A_{i,1}, \dots, A_{i,k_i}\}$ with $A_{i,j} \cap A_{i,l} = \emptyset$ for any $j \neq l$ and $\bigcup_{j=1}^{k_i} A_{i,j} = \{1, \dots, T\}$ and $i < s$ for any $i \in A_{i,j}$ and $s \in A_{i,l}$ with $j < l$. In other words, $A_{i,1}$ is a set containing all the time indices $t$ ranging from the first observational time 1 to the time $t^*_{i,1}$, corresponding to the first change point for the $i$th time series, i.e. $A_{i,1} = \{1, 2, \dots, t^*_{i,1}\}$. The generic block $A_{i,j}$ contains the observational times ranging from the $(j-1)$th to the $j$th change points, with $A_{i,j} = \{t^*_{i,j-1} + 1, t^*_{i,j-1} + 2, \dots, t^*_{i,j}\}$. A graphical representation of a single time series with corresponding latent parameters and blocks is given in Figure 1.

DRAFT

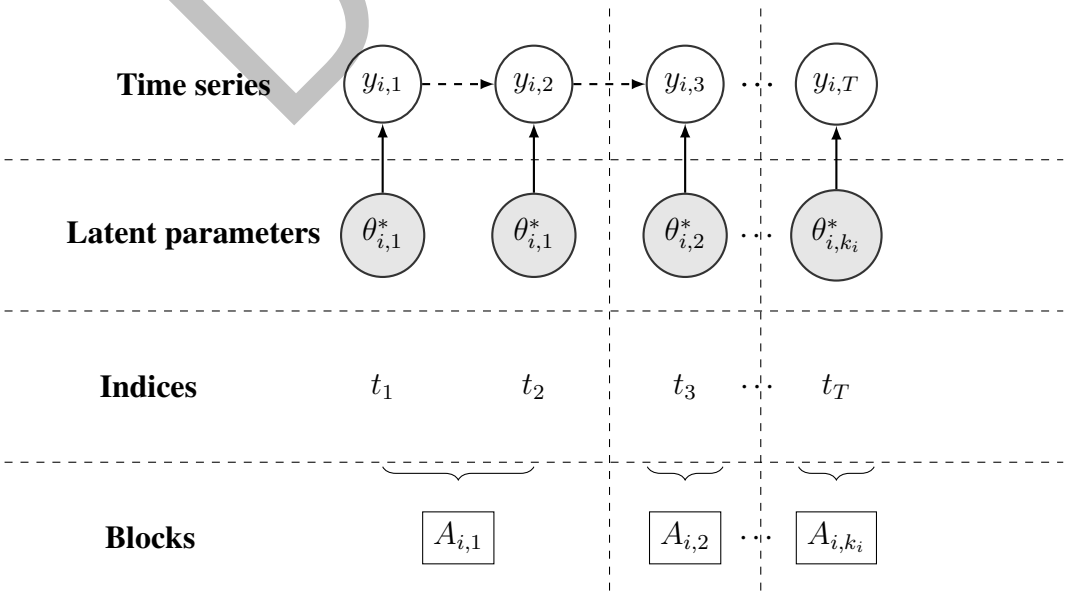



Figure 1: graphical representation of a time series $\boldsymbol{y}_i$ divided into $k_i$ blocks $A_{i,1}, A_{i,2}, \dots, A_{i,k_i}$

This way of defining the distribution of a random order for change points detection problems was proposed for the first time in (1) considering general product partition models. For a more recent non-parametric approach see (3), where the authors defined a probability distribution for random orders by considering the restriction of an exchangeable partition probability function from the partitions' space to the orders' space. Recent extensions also investigated the multivariate setting, see (2), and the multiple parameter case in (5).

Our goal is to perform model-based clustering of time series with respect to their latent random orders. In other words if two time series $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ have the same latent random order, i.e $\rho_i = \rho_j$, they belong to the same $j$th cluster with corresponding order $\rho_j^*$. The generic $i$th observation is also characterised by a unique sequence of parameters $\theta_{i,1}^*, \dots, \theta_{i,k}^*$. However our interest is not in such sequence of parameters, but only in the clustering structure, namely the sequence of blocks $A_1, \dots, A_k$ that characterises each cluster $\rho_j^*$.

We consider $\rho_1, \dots, \rho_n$ as a sequence of exchangeable realisations from a discrete distribution that we denote with $\tilde{p}$. The atoms of $\tilde{p}$ are the possible latent orders for the time series. Since we do not want to make assumptions on the parametric form of this distribution we consider $\tilde{p}$ as a realisation from a random discrete probability measure. Then we have that, according to De Finetti's theorem, the exchangeable sequence $\rho_1, \dots, \rho_n$ conditioned on $\tilde{p}$ can be represented as an $iid$ sequence precisely from $\tilde{p}$, i.e. $\rho_i \mid \tilde{p} \overset{iid}{\sim} \tilde{p}$, where $\tilde{p}$ is a discrete distribution of the form

$$\tilde{p} = \sum_{r=1}^{2^{T-1}} \pi_r \delta_{\rho_r^*}(\cdot), \tag{2}$$

with $\rho_1^*, \dots, \rho_{2^{T-1}}^*$ denoting all the possible orders of T elements. The distribution of $\tilde{p}$ is then fully specified by letting the weights $\pi_1, \dots, \pi_{2^{T-1}}$ distributed a priori as a Dirichlet distribution. The discrete structure of (2) implies the possible presence of ties among the orders, that means that there is a positive probability of having groups with more than one time series sharing the same structure of change points. We assume a Markovian dependence structure within each block of the $i$th time series and the likelihood term then can be factorised in the product of sub-sequential time instants. The final model is given by

$$\begin{aligned}
\boldsymbol{y}_i \mid \rho_i, \boldsymbol{\theta}_i^* &\sim \prod_{j=1}^{|\rho_i|} \prod_{t \in A_{i,j}} \mathcal{L}(y_{i,t} \mid y_{i,t-1}, \theta_{i,j}^*) \\
\theta_{i,j}^* &\overset{iid}{\sim} Q_0(d\theta) \\
\rho_1, \dots, \rho_n \mid \tilde{p} = p &\overset{iid}{\sim} \tilde{p} \\
\tilde{p} &= \sum_{r=1}^{2^{T-1}} \pi_r \delta_{\rho_r^*}(\cdot) \\
(\pi_1, \dots \pi_R) &\sim \text{Dir}(\alpha_1, \dots, \alpha_R),
\end{aligned} \tag{3}$$

with the proviso that if $t' = \min(A_{i,j})$ then $\mathcal{L}(y_{i,t'} \mid y_{i,t'-1}, \theta_{i,j}^*) = \mathcal{L}(y_{i,t'} \mid \theta_{i,j}^*)$.

## 3. Computational details

We denote with $\boldsymbol{Y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_n\}$ the observed data, with $\boldsymbol{\theta} = \{\theta_{1,1}, \dots \theta_{t,n}\}$ the whole latent parameters, with $\boldsymbol{\rho} = \{\rho_1, \dots, \rho_n\}$ the latent random orders and with $\boldsymbol{\pi} = \{\pi_1, \dots \pi_{2^{T-1}}\}$ the weights of (2). The joint distribution of the model can be obtained by exploiting the conditional distributions,

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{j=1}^{|\rho_i|} \prod_{l \in A_{i,j}} \mathcal{L}(y_{i,l} \mid y_{i,l-1}, \theta_{i,j}^*) \prod_{i=1}^{n} \prod_{j=1}^{|\rho_i|} Q_0(\theta_{i,j}^*) \prod_{i=1}^{n} \tilde{p}(\rho_i) \frac{1}{B(\boldsymbol{\alpha})} \prod_{r=1}^{2^{T-1}} \pi_r^{\alpha_r - 1}, \tag{4}$$

DRAFT

where $B(\boldsymbol{\alpha})$ is the multivariate beta function with parameter $\boldsymbol{\alpha}$ and $Q_0$ is a non-atomic probability measure whose support coincides with the support of the latent parameters. Since we are not interested both in the specific values of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ we integrate them out obtaining

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{\rho}) = \frac{\Gamma(\alpha^+)}{\Gamma(\alpha^+ + n)} \prod_{r=1}^{2^{T-1}} \frac{\Gamma(\alpha_r + n_r)}{\Gamma(\alpha_r)} \prod_{\{i:\rho_i=\rho_r^*\}} \prod_{j=1}^{|\rho_r^*|} \mathcal{M}(\{y_{i,l} \mid l \in A_{i,j}\}), \tag{5}$$

with $\alpha^+ = \sum_{r=1}^{2^{T-1}} \alpha_r$, $n_r = |\{i : \rho_i = \rho_r^*\}|$ and where $\mathcal{M}(\{y_{i,l} \mid l \in A_{i,j}\})$ is the marginal distribution of the $i$th time series in the $j$th block. In equation (5) we have the convolution of the block specific marginal distributions, through the marginal distribution of a Dirichlet-Multinomial model, and the marginal distribution for each specific block of each observed time series. We exploit the distribution of $\rho_i \mid \boldsymbol{\rho_{(i)}}$, where $\boldsymbol{\rho}_{(i)} = \{\rho_1 \dots \rho_{i-1}, \rho_{i+1}, \dots, \rho_n\}$ denotes the random orders associated to the observed time series without the $i$th element, to update the allocation of the time series to different components of $\tilde{p}$. That corresponds to a Pólya-Urn scheme of a Dirichlet-Multinomial process, and we have that

$$\begin{aligned} P(\rho_i = \rho_r^* \mid \boldsymbol{y}_i, \boldsymbol{\rho}_{(i)}) \propto & \sum_{r \in \mathcal{S}_{(i)}^c} \frac{\alpha_r}{\alpha^+ + n - 1} \prod_{j=1}^{|\rho_r^*|} \mathcal{M}\left(\{y_{i,l} \mid l \in A_{r,j}^*\}\right) \\ & + \sum_{r \in \mathcal{S}_{(i)}} \frac{\alpha_r + n_{(i),r}}{\alpha + n - 1} \prod_{j=1}^{|\rho_r^*|} \mathcal{M}\left(\{y_{i,l} \mid l \in A_{r,j}^*\}\right), \end{aligned} \tag{6}$$

where $A_{r,j}^*$ is the $j$th block of $\rho_r^*$, $\mathcal{S}_{(i)}$ denotes the indices in $\{1, \dots, 2^{T-1}\}$ whose clusters are non-empty after removing the $i$th element, i.e. $r \in \mathcal{S}_{(i)}$ if $|\{\ell \neq i : \rho_\ell = \rho_r^*\}| > 0$, $\mathcal{S}_{(i)}^c = \{1, \dots, 2^{T-1}\} \setminus S_{(i)}$ and $n_{(i),r}$ the number of time series (except $\boldsymbol{y}_i$) that belong to $\rho_r^*$.

In order to sample from equation (6) we resort to a collapsed Gibbs sampler in which we add an acceleration step to improve the mixing ability of the sampler. The acceleration step updates at each iteration the unique values $\rho_j^*$ and it is based on a split-and-merge strategy fully inspired by the work (3).

DRAFT

## 4. Illustration

We perform a simulation on synthetic data to verify if the model correctly identifies groups of time series generated by the same underlying distribution. We consider a scenario of normally distributed time series, each one with $T = 300$ time instants, grouped in $k = 3$ clusters respectively with size $n_1 = 3$, $n_2 = 4$ and $n_3 = 3$. We include a markovian dependency by considering each time observation dependent on the previous one. The $j$th time series at time $i$ is defined as

$$y_{ij} = \begin{cases} \mu_{ij} + N(0, 0.25) & \text{if } i = 1 \\ \gamma\, y_{i-1,j} + (1-\gamma)\, \mu_{ij} + N(0, 0.25) & \text{otherwise.} \end{cases}$$

We fix $\gamma = 0.5$ and set $\mu_{ij}$ according to the scheme

$$\textbf{Group 1} \quad (j = 1, \dots, 3) \quad \begin{cases} \mu_{ij} = 2 & \text{if } i = 1, \dots, 100 \\ \mu_{ij} = 1 & \text{if } i = 101, \dots, 200 \\ \mu_{ij} = 2 & \text{if } i = 201, \dots, 300 \end{cases}$$

$$\textbf{Group 2} \quad (j = 4, \dots, 7) \qquad \begin{cases} \mu_{ij} = 0 & \text{if } i = 1, \dots, 25 \\ \mu_{ij} = 1 & \text{if } i = 26, \dots, 50 \\ \mu_{ij} = 0 & \text{if } i = 51, \dots, 70 \\ \mu_{ij} = -1 & \text{if } i = 71, \dots, 160 \\ \mu_{ij} = -2 & \text{if } i = 161, \dots, 300 \end{cases}$$

$$\textbf{Group 3} \quad (j = 8, \dots, 10) \qquad \begin{cases} \mu_{ij} = -1 & \text{if } i = 1, \dots, 30 \\ \mu_{ij} = -2 & \text{if } i = 31, \dots, 150 \\ \mu_{ij} = 0 & \text{if } i = 151, \dots, 300. \end{cases}$$

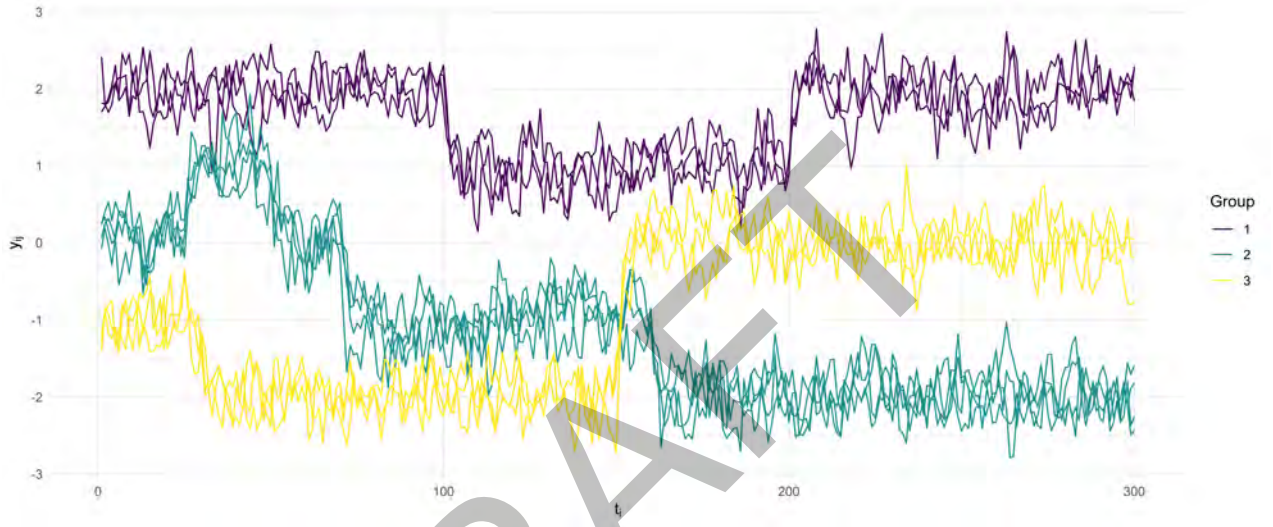



Figure 2: Posterior clustering of the synthetic time series.

Figure 2 shows a realisation from this model. We run the algorithm for 1000 iterations preceded by 500 burn-in iterations. Since sampling from all the potential $2^{T-1}$ partitions is infeasible, especially when $T$ is large, we sample from a $b$-dimensional subset of $S_{(i)}$. The dimension $b$ of this subset from which we sample the proposals of new unique values for $\rho_i$ is definitely a parameter that determines the quality of the approximation. We choose $b = 25$ so we can gain a good trade-off between approximation consistency and computational time. According to the *Variation of Information* loss function (4) the algorithm selects as optimal partition $\{\{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3\}, \{\boldsymbol{y}_4, \boldsymbol{y}_5, \boldsymbol{y}_6, \boldsymbol{y}_7\}, \{\boldsymbol{y}_8, \boldsymbol{y}_9, \boldsymbol{y}_{10}\}\}$ which corresponds to the real one.

## 5. Future directions

In this work we presented a way to effectively perform a clustering of time series according to their changes over time. Future directions consist in applying this model to real data in different contexts. As mentioned in the introduction a first possible application might be in analyzing the effect of gas prices on different companies, or in general how the prices of raw materials affect the economies. Another interesting usage might be in studying real estate market, especially how it reacts to the global market changes and to inflation. Finally a notable application is in the field of infectious disease epidemiology, specifically for grouping survival functions for the time of infection of patients affected by a particular disease.

DRAFT

## References


[1] Barry, D., Hartigan, J.A.: Product partition models for change point problems. Ann. Statist. **20**(1), 260–279 (1992). URL https://doi.org/10.1214/aos/1176348521

[2] Corradin, R., Danese, L., Ongaro, A.: Bayesian nonparametric change point detection for multivariate time series with missing observations. Internat. J. Approx. Reason. **143**, 26–43 (2022). URL https://doi.org/10.1016/j.ijar.2021.12.019

[3] Martínez, A.F., Mena, R.H.: On a nonparametric change point detection model in Markovian regimes. Bayesian Anal. **9**(4), 823–857 (2014). URL https://doi.org/10.1214/14-BA878

[4] Meilă, M.: Comparing clusterings—an information based distance. J. Multivariate Anal. **98**(5), 873–895 (2007). URL https://doi.org/10.1016/j.jmva.2006.11.013

[5] Pedroso, R.C., Loschi, R.H., Quintana, F.A.: Multipartition model for multiple change point identification (2021). URL https://arxiv.org/abs/2107.11456


DRAFT

# A functional Ground Motion Model for Italy built with a weighted analysis of reconstructed seismic curves

Teresa Bortolotti[a], Riccardo Peli[a], Giovanni Lanzano[b], Sara Sgobba[b], and Alessandra Menafoglio[a]

[a]MOX, Department of Mathematics, Politecnico di Milano; teresa.bortolotti@polimi.it, riccardo.peli@polimi.it, alessandra.menafoglio@polimi.it
[b]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano; giovanni.lanzano@ingv.it, sara.sgobba@ingv.it

**Abstract**

In the context of modelling earthquake-induced ground motion for the Italian territory, we consider the response of a Ground Motion Model as an element of a functional space. We propose to reconstruct the missing trajectories of the functional response, and to adopt a routine which estimates the functional coefficients of the Ground Motion Model by taking into account that some values of the functional responses have been directly observed, and other values have been reconstructed. To this end, we present a novel methodology that introduces a system of weights giving full weight to estimation errors made on originally observed values, and less weight to errors made on reconstructed values.

***Keywords:*** Functional data, weighted analysis, partially observed data, ground motion model

DRAFT

## 1. Introduction

Securing buildings from seismic damage is an issue of great importance all over the world, and especially in the Italian context, characterised by a medium-high seismic hazard level. To optimize the action of the institutions in terms of civil protection planning and seismic adaptation, it is pivotal to quantify the impact that the occurrence of a prospective earthquake would have on the territory. Ground Motion Models (GMMs) are regression models that estimate the expected value of some ground motion intensity measure, conditionally on seismic parameters that describe a given seismic scenario. Spectral acceleration is an intensity measure commonly used as response variable in GMMs, which describes the maximum acceleration that a damped harmonic oscillator (*i.e.*, a building) is subjected to during a seismic sequence. By construction, spectral acceleration is defined over a domain of natural vibration periods. This implies that the response variable of a GMM can be embedded in three alternative analytical frameworks: scalar, multivariate, or functional. The scalar GMM proposed in (3), which we refer to as ITA18, is a highly interpretable benchmark for the italian context, but necessarely suffers from the two main drawbacks of a scalar approach: *(i)* it provides only discrete estimates at some specific periods of interest, and *(ii)* no correlation between periods is included in the model. In this communication, we report the main methodological idea and results discussed by (1), which considers spectral acceleration as a function defined over the domain of vibration periods, and in this framework provides estimates of the regression coefficients of the ITA18 model.

Embedding spectral acceleration in a functional space implies coping with the fact that a non-negligible fraction of SA profiles are only partially observed over the domain of vibration periods, and need to be reconstructed. In the framework of functional data analysis, we propose a methodology that

produces estimates of the functional regression coefficients of the GMM, by accounting for the varying degrees of confidence that we have on different parts of the reconstructed curves of spectral acceleration. This is done by defining curve-specific weighting functions which enter the estimation routine by giving full weight to the errors made on the parts of the curves which are directly observed, and less weight to the errors made on parts of the curves that are reconstructed and thus more uncertain. The routine consists of two cascading steps – smoothing and concurrent linear regression – whose classical formulation and use (2; 5) are generalized in (1), in order to include observation-specific weighting functions.

## 2. Data and Model

Most of the data is provided by the ITalian ACcelerometric Archive (ITACA; (6)), which collects the manually-revised waveforms recorded by the main seismic networks in Italy. The dataset collects 5607 records of 146 earthquakes registered at 1657 stations (4), also including recordings of high-magnitude worldwide earthquakes, characterized by strike-slip and thrust faulting mechanisms.

The analysis considers 37 discrete intensity measures, namely the peak ground acceleration (PGA) and the spectral acceleration (SA) recorded at 36 vibration periods $T_j$ in the interval $[0.04\ s, 10\ s]$. By convention, we refer to PGA as the spectral acceleration corresponding to $T_1 = 0\ s$, and to the set of longitudinal observations of spectral acceleration as SA profile. The manual processing of the intensity measures records at the sites results in some cases in incomplete SA profiles over the domain of periods. In the dataset under analysis, the fraction of observed values is close to 1 up to $5\ s$, and then quickly falls to about 0.75 at longer periods. The idea that we propose is to reconstruct the partially observed SA profiles from their last valid record up to $10\ s$ (see Section 4.1 and Section 5.1 of (1) for an overview of alternative reconstruction strategies), and to account later in the analysis for the presence of reconstructed and thus more uncertain values. This prevents the great loss of information that would occur if one discarded from the dataset the profiles with at least one missing value.

The ground motion model proposed in (3), and hereafter referred to as ITA18, estimates the median spectral acceleration separately at the 37 periods, by fitting a scalar linear regression model at every period. At each $T_j$, the model reads

$$\log_{10} \mathrm{SA}_j = a_j + F_M(M_w(T_j), SoF) + F_D(M_w(T_j), R(T_j)) + F_S(V_{S30}) + \epsilon_j, \tag{1}$$

where $a_j$ is the offset, $F_M$, $F_D$, $F_S$ are the source-, path- and site-related terms respectively, and $\epsilon_j$ is the remaining error. Variable $M_w$ denotes the magnitude, $SoF$ the style-of-faulting, $R$ the source-to-site distance corrected by the pseudodepth at the site, and $V_{S30}$ the shear-wave velocity. Notice that parameters $M_h$, $M_{\text{ref}}$ and $h$ in equation (1) are dependent on the periods, implying that we are given with longitudinal observations of the terms $F_M$ and $F_D$. As the aim is to give a fully functional formulation of equation (1), i.e. functional response, coefficients and covariates, we embed this period-dependent terms in a suitable functional space with domain $\mathcal{T}$. The functional covariates are obtained resorting to penalized smoothing (5). Following some *ad hoc* considerations, the functional source term $\mathcal{F}_M$ is defined on a quadratic B-spline basis via a smoothing that penalizes its first derivative. The functional path term $\mathcal{F}_D$, on the other hand, is a cubic B-spline with knots at the sampling points.

Eventually, the embedding of the scalar model into a fully functional framework reads

$$\log_{10} \mathcal{SA} = \alpha + \mathcal{F}_M(M_w, \mathrm{SoF}, \mathcal{M}_h) + \mathcal{F}_D(M_w, d_{JB}, \mathcal{M}_{\text{ref}}, h) + \mathcal{F}_S(V_{S30}) + \mathcal{E}. \tag{2}$$

In (2), $\mathcal{SA}$ is a functional random variable with values in a suitable infinite dimensional space, $\alpha$ is the offset, and $\mathcal{F}_M$, $\mathcal{F}_D$ and $\mathcal{F}_S$ are known functions with domain $\mathcal{T}$. We are assuming $\mathcal{E}$ to be a realization of a zero mean stochastic process.

## 3. Method

Let $y_1, \ldots, y_n$ be the reconstructed curves belonging to the space $L^2(\mathcal{T})$, where $\mathcal{T}$ is an open subset of $\mathbb{R}$. Each reconstructed curve $y_i$ is coupled with a weighting function $w_i : \mathcal{T} \to [0, 1]$. Denote by

DRAFT

$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN})$ the vector of discrete observations, either directly registered or reconstructed, of curve $y_i$ at the sampling instants $t_1, \ldots, t_N$.

The smoothing of each $\boldsymbol{y}_i$ is done by relying on a weighted penalized least square criterion (5), where the weights at the sampling instants are defined as $v_{ij} = w_i(t_j)$ for $j = 1, \ldots, N$. The penalty is imposed on the squared $L^2$-norm of the second derivative of the smoothed function, and the minimum is searched in the space $H^2(\mathcal{T})$ to guarantee the finiteness of this norm. The penalization parameter for each curve is tuned via generalized cross-validation. The resulting smoothed curve is a cubic spline with knots at the sampling points. Differently from the classical weighted penalized smoothing, we propose to make use of weights that vary not only along the sampling instants, but also on a data-by-data basis. This has an impact on the form of the overall smoothing map, that links the matrix of raw observations to the matrix collecting the smoothing coefficients. The overall smoothing map plays a crucial role in the assessment of the uncertainty related to the point coefficients estimates, and is thus relevant to provide an accurate representation of it (refer to Section 3 of (1) for a more detailed discussion on this).

In the second step of the routine, the smoothed functional observations $z_i$ are assumed to be generated by a concurrent regression model, namely

$$z_i = \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

where $x_{i1}, \ldots, x_{iq}$ are independent functional covariates, and $\beta_1, \ldots, \beta_q$ are the functional coefficients defined on $\mathcal{T}$. Errors $\epsilon_1, \ldots, \epsilon_n$ are assumed to be realizations of a zero-mean stochastic process.
We define a penalized fitting criterion which minimizes

$$\sum_{i=1}^{n} \int_{\mathcal{T}} w_i \left(z_i - \beta_j x_{ij}\right)^2 + \sum_{j=1}^{q} \int_{\mathcal{T}} \lambda_j \left(D^2 \beta_j\right)^2. \tag{4}$$

The second sum in (4) is a roughness penalty that regularizes the estimates, and $\lambda_1, \ldots, \lambda_q$ are penalization parameters which are tuned via generalized cross-validation. Each coefficient is associated to a specific penalization parameter, meaning that the estimates of the coefficients are allowed to have diverse levels of smoothness.

The dimensionality of problem (4) is reduced by assuming that each $\beta_j$ belongs to a finite dimensional space spanned by suitable basis functions, and that it is uniquely identified by the coefficients of a specific linear combination of the basis functions. Section 3 in (1) shows that problem (4) is well posed in this setting, and that the explicit solution for $\beta_1, \ldots, \beta_q$ is found by exploiting the dimensionality reduction argument and some algebraic manipulations of the finite dimensional counterpart of (4).

DRAFT

## 4. Results

We refer to Section 5 of (1) for an extensive discussion on the choice of the weights, the calibration of the model, and a full description of the results of analysis.

In this communication we only call attention to the effect that the weighted functional methodology has on the estimates of the regression coefficients, by considering a specific coefficient of ITA18 as paradigmatic example. The site term in equation (1) reads $F_S(V_{S30}) = k \log\left(\frac{V_0}{800}\right)$, where $V_0 = V_{S30}$ if $V_{S30} \leq 1500\ m/s$, $V_0 = 1500\ m/s$ otherwise. Coefficient $k$ accounts for the linear scaling of the shear-wave velocity for values of $V_{S30}$ lower than $1500\ m/s$. The estimate of $k$ resulting from the use of the scalar and functional models is displayed in Figure 1. The fences of the functional boxplot are to be interpreted as a reliable quantification of the variability associated to the functional coefficients point estimates, simultaneously over the entire domain of definition of the reconstructed SA profiles. The argument that legitimizes the adoption of a bootstrap methodology in this context is reported in Section 3.3 of (1). The functional estimate follows the trend of the scalar estimate while displaying a smoother behavior. The main deviation is observed in the right half of the domain, where the weighted analysis impacts the estimates. The fitting of the scalar model at large periods neglects all the information given

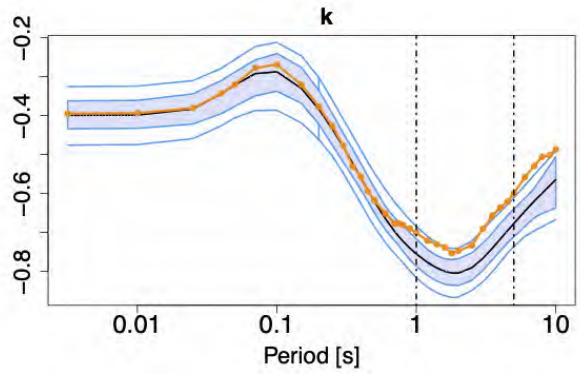



Figure 1: Comparison of the estimates of coefficient $k$, made with the scalar (orange) and the functional (black) models. The light blue bands are the fences of the functional boxplot built for $k$. The vertical lines mark the points of the last instants where 100% and $> 90\%$ of the curves are observed.

by the partially observed SA profiles, and the estimates are built by relying only on the 75% of the data used to get the functional estimates. By exploiting the correlation between the SA ordinates to reconstruct the profiles, and by adopting a weighted functional approach, the proposed method overcomes the loss of information, while providing reliable estimates continuously over the interval of relevant oscillation periods.

## References

[1] Bortolotti, T., Peli, R., Lanzano, G., Sgobba,S., Menafoglio, A.: Weighted functional data analysis for the calibration of ground-motion models in Italy. MOX-report 31/2022, Politecnico di Milano (2022)

[2] de Boor,C.: A Practical Guide to Splines. Revised Edition, Springer, New York (2001)

[3] Lanzano,G., Luzi,L., Pacor,F., Felicetta,C., Puglia,R., Sgobba,S., D'Amico,M.: A Revised Ground Motion Prediction Model for Shallow Crustal Earthquakes in Italy. Bulletin of the Seismological Society of America. **109(2)**, 525–540 (2019) doi: 10.1785/0120180210

[4] Lanzano, G., Ramadan, F., Luzi, L., Sgobba, S., Felicetta, C., Pacor, F., D'Amico, M., Puglia, R., Russo, E.: Parametric table of the ITA18 GMM for PGA, PGV and Spectral Acceleration ordinates. Istituto Nazionale di Geofisica e Vulcanologia (INGV) (2022) https://doi.org/10.13127/ita18/sa_flatfile/

[5] Ramsay, J. O. and Silverman, B. W.: Functional Data Analysis. Springer-Verlag, New York (2005)

[6] Russo, E., Felicetta, C., D'Amico, M. C., Sgobba, S., Lanzano, G., Mascandola, C., Pacor, F., Luzi, L.: ITalian ACcelerometric Archive (ITACA), version 3.2. Istituto Nazionale di Geofisica e Vulcanologia (INGV) (2022) https://itaca.mi.ingv.it/ItacaNet_32/

DRAFT

# Conditional Gaussian Graphical Models for Functional Variables with Partially Separable Operators

Rita Fici[a], Gianluca Sottile[a], and Luigi Augugliaro[a]

[a]Department SEAS, University of Palermo, Italy; rita.fici@unipa.it, gianluca.sottile@unipa.it, luigi.augugliaro@unipa.it

**Abstract**

Functional graphical modeling is gaining increasing attention in recent years. In this paper, we contribute to the literature by extending the notion of conditional Gaussian graphical model to a functional setting. We propose a double-penalized estimator and an efficient algorithm to recover the edge-set encoding both the conditional covariance structure of the response functions and the effects of the predictor functions on the conditional distribution.

***Keywords:*** Graphical models, multivariate functional data, multivariate Gaussian process, partial separability, sparse inference.

DRAFT

## 1. Introduction

In recent years, functional data has become a commonly encountered data type. The first approach aimed to extend graphical models to the functional setting was proposed in (6), where, under the assumption that the random functions follow a multivariate Gaussian process (MGP), the authors introduce the notion of functional Gaussian graphical model (fGGM) and an extension of the graphical lasso (glasso) (9) to estimate the edge-set encoding the conditional dependence structure. A notion of conditional functional graphical model, where the graph links are allowed to vary with the external variables, is introduced in (5). In this paper, we are not interested in constructing a random graph; rather, we are interested in the effect of the explanatory variables on the expected value of the multivariate response process. Recently, in (10) is addressed the general problem of covariance modelling for multivariate functional data, particularly fGGMs. The authors introduce the notion of partial separability for the covariance operator and show that this is particularly useful in functional graphical modelling (FGM) since it allows us to overcome the theoretical problems related to the covariance operator, which is compact and thus not invertible.

In this paper, we contribute to the literature on FGM by extending the notion of conditional Gaussian graphical model (cGGM) (8) and proposing a double-penalized estimator by which to recover the edge-set of the corresponding graph. We complete this section by providing a brief description of the cGGM models.

Let $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ be two random vectors. cGGMs are based on the assumption that $\boldsymbol{Y} \mid \boldsymbol{x} \sim N(\boldsymbol{Bx}, \Sigma)$, where $\boldsymbol{Bx} = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{x})$ and $\Sigma = \mathrm{V}(\boldsymbol{Y} \mid \boldsymbol{x})$, and that exists a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E} = \mathscr{E}_\mu \cup \mathscr{E}_\Theta)$ encoding the effects of $\boldsymbol{X}$ onto the conditional distribution of $\boldsymbol{Y}$. The edge-set $\mathscr{E}$ is defined as the union of two specific sets. The set $\mathscr{E}_\mu$ contains the directed links representing the effects

of $\boldsymbol{X}$ on $\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{x})$, i.e., the directed link $(m,h)$ belongs to $\mathscr{E}_\mu$ iff $X_m$ has an effect on the conditional expected value of $Y_h$, i.e., $\beta_{hm} \neq 0$. The set $\mathscr{E}_\Theta$ contains the undirected links depicting the conditional dependence structure among the response variables, consequently, according to the standard theory on the factorization of the multivariate Gaussian distribution (see (4) for more details), the undirected link $(h,k)$ belongs to $\mathscr{E}_\Theta$ iff the corresponding element of the precision matrix $\Theta = \Sigma^{-1}$ is different from zero. In this class of graphical models, our final goal is to estimate $\boldsymbol{B}$ and $\Theta$ and recover the information encoded in $\mathscr{E}$.

## 2. The functional conditional Gaussian graphical model

**Notation** We use the term multivariate functional data to refer to the realization of a multivariate process. Specifically, we denote the multivariate processes corresponding to the response and predictor functions as $\mathscr{P}_Y = \{\boldsymbol{\mathscr{Y}}(t) \in \mathbb{R}^p : t \in \mathscr{T}\}$ and $\mathscr{P}_X = \{\boldsymbol{\mathscr{X}}(s) \in \mathbb{R}^q : s \in \mathscr{S}\}$, where $\mathscr{T}$ and $\mathscr{S}$ are closed subsets of $\mathbb{R}$. It is assumed that $\mathscr{P}_Y$ and $\mathscr{P}_X$ are MGPs and that $\mathscr{Y}_h$ and $\mathscr{X}_m$ are elements of $\mathscr{L}_2(\mathscr{T})$ and $\mathscr{L}_2(\mathscr{S})$, where $\mathscr{L}_2(\cdot)$ denotes the Hilbert space of square-integrable functions endowed with the standard inner product $\langle g_1, g_2 \rangle = \int_{\mathscr{S}} g_1(s) g_2(s) ds$ and norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$. We also assume that $\boldsymbol{\mathscr{X}}$ has zero mean and a smooth covariance function $G^X(s_1,s_2) = \{G^X_{mn}(s_1,s_2)\}$, where $G^X_{mn}(s_1,s_2) = \mathrm{cov}(\mathscr{X}_m(s_1), \mathscr{X}_n(s_2))$. Similarly, $\boldsymbol{\mathscr{Y}}$ has zero mean and smooth covariance function $G^Y(t_1,t_2) = \{G^Y_{hk}(t_1,t_2)\}$. Finally, for each bivariate function $f \in \mathscr{L}_2(\mathscr{T} \times \mathscr{S})$, by $\|f\| = \{\int\int f(t,s) dt ds\}^{1/2}$ we denote the Hilbert-Schimdt norm.

**The functional conditional Gaussian graphical model** We are interesting in inferring how the predictor process affects the distribution of the conditional process $\mathscr{P}_{Y|X} = \{\boldsymbol{\mathscr{Y}}(t) \mid \boldsymbol{\mathscr{X}} : t \in \mathscr{T}\}$, which is Gaussian and uniquely specified by:

$$\mathrm{E}(\mathscr{Y}_h(t) \mid \boldsymbol{\mathscr{X}}) = \sum_{m=1}^{q} \int_{\mathscr{S}} \beta_{hm}(t,s) \mathscr{X}_m(s) ds, \quad G^{Y|X}(t_1,t_2) = \left\{ G^{Y|X}_{hk}(t_1,t_2) \right\}, \tag{1}$$

where $\beta_{hm}(t,s) \in \mathscr{L}_2(\mathscr{T} \times \mathscr{S})$ are the bivariate regression coefficient functions, and $G^{Y|X}_{hk}(t_1,t_2) = \mathrm{cov}(\mathscr{Y}_h(t_1), \mathscr{Y}_k(t_2) \mid \boldsymbol{\mathscr{X}})$ is the conditional covariance function. To provide a coherent extension of the cGGM, we must define the edge-sets $\mathscr{E}_\mu$ and $\mathscr{E}_\Theta$. While the first set can be easily defined using the left-hand-side in (1), i.e., $\mathscr{E}_\mu = \{(m,h) : \|\beta_{hm}\| \neq 0\}$, a proper definition of $\mathscr{E}_\Theta$ can be obtained only through the notion of conditional cross-covariance function (6):

$$C^{Y|X}_{hk}(t_1,t_2) = \mathrm{cov}(\mathscr{Y}_h(t_1), \mathscr{Y}_k(t_2) \mid \boldsymbol{\mathscr{Y}}_{-(hk)}, \boldsymbol{\mathscr{X}}), \tag{2}$$

which represents the covariance between $\mathscr{Y}_h$ and $\mathscr{Y}_k$ given the processes $\boldsymbol{\mathscr{Y}}_{-(hk)}$ and $\boldsymbol{\mathscr{X}}$. Using (2), we define $\mathscr{E}_\Theta = \{(h,k) : \|C^{Y|X}_{hk}\| \neq 0\}$. In the remaining part of this paper, by functional conditional Gaussian graphical model (fcGGM) we mean the set $\{\mathscr{P}_{Y|X}, \mathscr{G} = \{\mathscr{V}, \mathscr{E} = \mathscr{E}_\mu \cup \mathscr{E}_\Theta\}\}$, and our goal is to recover the edge-set $\mathscr{E}$.

**Partial separability and fcGGM** In principle, we could recover $\mathscr{E}$ using the approach presented in (6), representing each random function by the coefficients of a truncated basis expansion and then estimating $\mathscr{E}$ using a modified glasso estimator. Although this method is an intuitive approach to FGM estimation, the authors show the existence of a theoretical link between precision matrix and true FGM, only under the assumption that each random function takes values in a finite-dimensional space. As elucidated in (10), in an infinite-dimensional setting, the relationship between precision matrix and conditional independence structure is lost because the covariance operator is compact and thus not invertible; therefore, to estimate $\mathscr{E}$ in an fcGGM, we enforce our assumptions by assuming that the covariance operators $G^X$ and $G^Y$ are partially separable. As a consequence, by Theorem 1 in (10), we have the following

DRAFT

multivariate expansions:

$$\mathscr{Y}_h(t) = \sum_{l=1}^{+\infty} Y_{hl}\varphi_l(t), \quad \text{and} \quad \mathscr{X}_m(s) = \sum_{l=1}^{+\infty} X_{ml}\psi_l(s), \tag{3}$$

where $\{\varphi_l\}_{l=1}^{+\infty}$, $\{\psi_l\}_{l=1}^{+\infty}$ are orthonormal bases of $\mathscr{L}_2(\mathscr{T})$ and $\mathscr{L}_2(\mathscr{S})$, whereas $Y_{hl} = \langle \mathscr{Y}_h, \varphi_l \rangle$, $X_{ml} = \langle \mathscr{X}_m, \phi_l \rangle$ are random variables. Since $\mathscr{P}_Y$ and $\mathscr{P}_X$ are Gaussian, the vector $\boldsymbol{Z}_l = (X_{1l}, \ldots, X_{ql}, Y_{1l}, \ldots, Y_{pl})^\top$ is also Gaussian with parameters, $\mathrm{E}(\boldsymbol{Z}_l) = \boldsymbol{0}$ and $V(\boldsymbol{Z}_l) = \Sigma_l$. Moreover, $\boldsymbol{Z}_l \perp\!\!\!\perp \boldsymbol{Z}_{l'}$. Using (3), it is possible to show that:

$$\mathrm{E}(\mathscr{Y}_h(t) \mid \boldsymbol{\mathscr{X}}) = \sum_{m=1}^{q} \sum_{l=1}^{+\infty} \beta_{hml} x_{ml} \varphi_l(t), \tag{4}$$

where $x_{ml}$ denotes a realization of $X_{ml}$ and $\sum_{m=1}^{q} \beta_{hml} x_{ml} = \mathrm{E}(Y_{hl} \mid \boldsymbol{X}_l)$. A direct consequence of the expansion (4) is that $\mathscr{E}_\mu$ can be defined in terms of $\beta_{hml}$, i.e., $(m,h) \in \mathscr{E}_\mu$ iff exists at least an index $l \in \mathbb{N}$ such that $\beta_{hml} \neq 0$.

The main advantage of the expansion (4) is that it allows us to express the conditional cross-covariance function (2) in terms of conditional covariance between $Y_{hl}$ and $Y_{kl}$. First, note that expansion (4) also implies that the residual process admits a multivariate expansion of type (3), thus, according to Theorem 1 of (10), the covariance operator in (1) is also partially separable, consequently, using Theorem 3 in (10) and the standard results on the conditional Gaussian distribution, we have:

$$C_{hk}^{Y|X}(t_1,t_2) = \sum_{l=1}^{+\infty} \mathrm{cov}(Y_{hl}, Y_{kl} \mid \boldsymbol{Y}_{-(hk)}, \boldsymbol{X}_l)\varphi_l(t_1)\varphi_l(t_2) = -\sum_{l=1}^{+\infty} \frac{\theta_{hkl}\varphi_l(t_1)\varphi_l(t_2)}{\theta_{hhl}\theta_{kkl} - \theta_{hkl}^2}, \tag{5}$$

where $\theta_{hkl}$ are the entries of $\Theta_l = \mathrm{V}(\boldsymbol{Y}_l \mid \boldsymbol{X}_l)^{-1}$. Using (5) it follows that an undirected link, say $(h,k)$, belongs to $\mathscr{E}_\Theta$ iff exists at least an index $l \in \mathbb{N}$ such that $\theta_{hkl} \neq 0$.

**The functional joint conditional graphical lasso estimator** In the previous section, we have shown that all the necessary information needed to recover the edge set associated with an fcGGM is contained in the conditional distribution of $\boldsymbol{Y}_l$ given $\boldsymbol{X}_l$. Below, we propose a two-step procedure to estimate $\mathscr{E}$.

Step 1. Suppose we observe $N$ independent realizations from $\mathscr{P}_Y$ and $\mathscr{P}_X$, denoted by $\boldsymbol{\mathscr{Y}}_i = (\mathscr{Y}_{i1}, \ldots, \mathscr{Y}_{ip})^\top$ and $\boldsymbol{\mathscr{X}}_i = (\mathscr{X}_{i1}, \ldots, \mathscr{X}_{ip})^\top$, with $i = 1, \ldots, N$, over $T$ time instants. Expansions (3) allow us to represent each random function as an infinite-dimensional object; thus, it is necessary for some form of dimensionality reduction. First, the mean functions are calculated over the observed time instants as $\bar{\mathscr{Y}}_h(t) = \sum_{i=1}^{N} \mathscr{Y}_{ih}(t)/N$. Then $p$ autocovariance matrices are estimated and each entry is given by $[\widehat{G}_h^Y]_{1,2} = \sum_{i=1}^{N} \{\mathscr{Y}_{ih}(t_1) - \bar{\mathscr{Y}}_h(t_1)\}\{\mathscr{Y}_{ih}(t_2) - \bar{\mathscr{Y}}_h(t_2)\}/N$. All those $T \times T$ matrices are summed and divided by $p$ in order to have $\hat{H}^Y$, which indicates the mean-variance over the responses variables for each couple of time instants. According to Theorem 2 in (10), the basis functions $\varphi_l(t)$ can be estimated performing the eigen-decomposition on $\widehat{H}^Y$. Each $\mathscr{Y}_{ih}$ can be approximated using the first $L$ leading terms, i.e., the function $\mathscr{Y}_{ih}^L(t) = \sum_{l=1}^{L} y_{ihl}\widehat{\varphi}_l(t)$, where the estimated principal component scores are $y_{ihl} = \langle \mathscr{Y}_{ih}, \widehat{\varphi}_l \rangle$. The procedure described above is used to estimate the quantities related to $\mathscr{X}_{im}(s)$, i.e., $\widehat{\psi}_l(s)$ and the corresponding scores $x_{iml} = \langle \mathscr{X}_{im}, \widehat{\psi}_l \rangle$. For ease of notation, we suppose again to use the first $L$ leading terms to approximate the random predictor functions, i.e., $\mathscr{X}_{im}^L(s) = \sum_{l=1}^{L} x_{iml}\widehat{\psi}_l(s)$.

Step 2. Let $\mathbf{Y}_l = (y_{ihl})$ and $\mathbf{X}_l = (x_{iml})$, with $l = 1, \ldots, L$, be the matrices of the estimated scores. Given the assumption underlying the fcGGM, the rows of these matrices are independent realizations from a multiple cGGM, i.e., a collection of cGGMs; therefore, the sets $\mathscr{E}_\mu$ and $\mathscr{E}_\Theta$ can be estimated using a proper extension of the joint glasso (2), such as the one proposed in (3) or, in the context of censored data, in (1) and (7).

Let us denote by $\boldsymbol{B}_l$ and $\Theta_l$ the parameters associated to the $l$th cGGM and let $\{\boldsymbol{B}\} = \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_L\}$ and $\{\Theta\} = \{\Theta_1, \ldots, \Theta_L\}$. As the assumption of partial separability implies that $\boldsymbol{Z}_l \perp\!\!\!\perp \boldsymbol{Z}_{l'}$, for each $l \neq l'$, we

DRAFT

propose to recover $\mathscr{E}$ using the following double-penalized estimator, named functional joint conditional glasso estimator:

$$\{\widehat{\boldsymbol{B}}\},\{\widehat{\Theta}\} = \arg\max \sum_{l=1}^{L} \{\log\det\Theta_l - \mathrm{tr}(\mathbf{S}(\boldsymbol{B}_l)\Theta_l)\} - \lambda P_1(\{\boldsymbol{B}\}) - \rho P_2(\{\Theta\}), \tag{6}$$

where $\mathbf{S}(\boldsymbol{B}_l) = (\mathbf{Y}_l - \mathbf{X}_l\boldsymbol{B}_l)^\top(\mathbf{Y}_l - \mathbf{X}_l\boldsymbol{B}_l)/N$. The penalty functions in (6) selects convex functions that encourage sparsity in each matrix and specific forms of similarity across the regression coefficient matrices and the precision matrices. In this paper, we propose to use the group lasso penalty functions: $P_1(\{\boldsymbol{B}\}) = \sum_{h=1}^{p}\sum_{m=1}^{q}\left(\sum_{l=1}^{L}\beta_{hml}^2\right)^{1/2}$ and $P_2(\{\Theta\}) = \sum_{h\neq k}\left(\sum_{l=1}^{L}\theta_{hkl}^2\right)^{1/2}$, thus, the desired edge-sets can be estimated by $\widehat{\mathscr{E}}_\mu = \left\{(m,h) : \sum_{l=1}^{L}\hat{\beta}_{hml}^2 > 0\right\}$ and $\widehat{\mathscr{E}}_\Theta = \left\{(h,k) : \sum_{l=1}^{L}\hat{\theta}_{hkl}^2 > 0\right\}$.

## 3. A simulation study

To simulate a sample of $N = 600$ independent observations from an fcGGM, we use the following model:

$$y_{ihr}^L = \sum_{l=1}^{L} y_{ihl}\varphi_l(t_r) + \varepsilon_{ihr}, \qquad x_{imr}^L = \sum_{l=1}^{L} x_{iml}\psi_l(s_r) + \varepsilon_{imr},$$

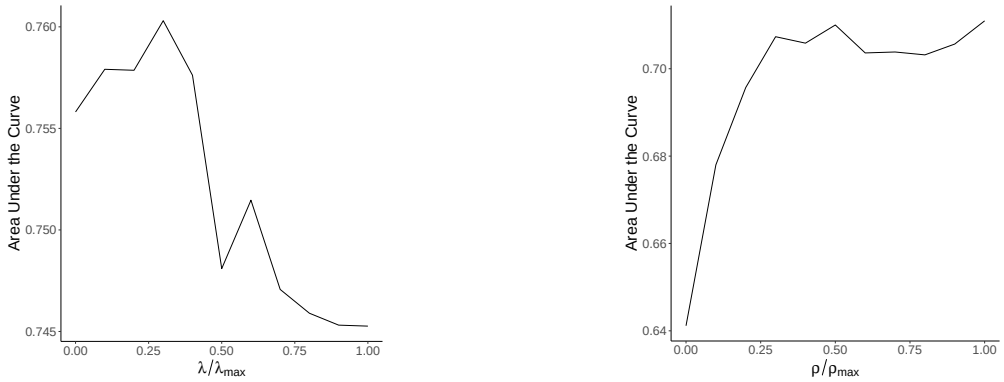
where $\varepsilon_{ihr}$ and $\varepsilon_{imr}$ are independent random errors drawn from $N(0,10)$, $\{t_r\}_{r=1}^{30}$ and $\{s_r\}_{r=1}^{30}$ are evenly spaced sequences with $t_1 = s_1 = 0$ and $t_{30} = s_{30} = 1$, and $\{\varphi_l\},\{\psi_l\}$ are Fourier bases. In our study, we set $L = 3$, $p = 24$ and $q = 7$. According to the assumptions underlying the proposed fcGGM, for each $l$, the vectors $\boldsymbol{z}_{il} = (\boldsymbol{x}_{il}^\top, \boldsymbol{y}_{il}^\top)^\top$ are independent realizations from a multivariate Gaussian distribution with zero expected value and covariance matrix $\Sigma_l^z$ structured as follows:

$$\Sigma_l^z = 3l^{-1.8} \times \begin{bmatrix} \sigma_x^2\boldsymbol{I} + \boldsymbol{B}_l\Theta_l\boldsymbol{B}_l^\top & -\boldsymbol{B}_l\Theta_l \\ -\Theta_l\boldsymbol{B}_l^\top & \Theta_l \end{bmatrix}^{-1}. \tag{7}$$

where, as in (10), the decreasing factor $3l^{1.8}$ guarantees that $\mathrm{tr}(\Sigma_l^z)$ decreases monotonically in $l$, whereas the quantities in the matrix in (7) are related to the marginal distribution of $\boldsymbol{X}_{il}$ and to the conditional distribution of $\boldsymbol{Y}_{il}$ given $\boldsymbol{x}_{il}$ be the identities: $\mathrm{V}(\boldsymbol{X}_{il}) = \sigma_x^2\boldsymbol{I}$, with $\sigma_x^2 = 20$, $\mathrm{E}(\boldsymbol{Y}_{il} \mid \boldsymbol{x}_{il}) = \boldsymbol{B}_l\boldsymbol{x}_{il}$ and, finally, $\{\mathrm{V}(\boldsymbol{Y}_{il} \mid \boldsymbol{x}_{il})\}^{-1} = \Theta_l$. To generate a sparse fcGGM, for each $l$, each row of $\boldsymbol{B}_l$ has only two non-zero regression coefficients sampled from $U([-0.8,-0.5]\cup[+0.5,+0.8])$ whereas the conditional precision matrix $\Theta_l$ is structured in such a way that the associated graph is the union of a common and a specific star. Formally, the non-zero entries of each $\Theta_l$ are sampled by the model: $\theta_{hkl} \sim U([-0.15,-0.10]\cup[+0.10,+0.15])$, with $h \in \{1, 6l+1\}$ and $k = (h+1),\ldots,(h+5)$.

To compute the estimator (6), we use the algorithm proposed in (7). The behaviours of $\mathscr{E}_{\hat{\Theta}}$ and $\mathscr{E}_{\hat{\mu}}$ are studied under different combinations of $\lambda$ and $\rho$; to analyze the coefficient path for $\{\widehat{\Theta}\}$, we first control the amount of shrinkage on $\{\widehat{\boldsymbol{B}}\}$ by keeping fixed the ratio $\lambda/\lambda_{max}$, and then, for this fixed $\lambda$ value, the path for $\{\widehat{\Theta}\}$ is computed across a decreasing sequence of eleven evenly spaced values of $\rho$, from $\rho_{max}$ to 0 with steps equal to $0.1\times\rho_{max}$. We use the median area under ROC curves to evaluate the resulting path in network recovery. Figure 1 shows the results. The left side of figure 1 shows the values of the AUC given by the eleven values of $\rho$ when $\lambda$ is fixed. On the $X-$axes, there is the value of $\lambda$ expressed in percentage of its maximum value; the first value, $0\times\lambda_{max}$, corresponds to the case in which $\{\widehat{\boldsymbol{B}}\}$ is not penalized, the last value corresponds to the case where the predictor variables do not affect the conditional expected value of the response variables. On the $Y-$axes, the AUC median value over 50 simulations with $\lambda$ fixed is reported. The left side of figure 1 shows that the level of shrinkage of $\{\widehat{\boldsymbol{B}}\}$ affects the ratio between TPR and FPR of $\mathscr{E}_{\hat{\Theta}}$, indeed, when the penalization for $\{\widehat{\boldsymbol{B}}\}$ is small, the resulting AUC of the network recovery for $\{\widehat{\Theta}\}$ is high. The higher the penalization parameter for $\{\widehat{\boldsymbol{B}}\}$, the more difficult for the model to detect the correct set of edges, suggesting that the explanatory variables are needed for an accurate evaluation of $\mathscr{E}_{\hat{\Theta}}$ and that the regression model is working well. The comparison of $\{\widehat{\boldsymbol{B}}\}$ paths is made using the same strategy as for $\{\widehat{\Theta}\}$, but inverting the role. Unlike the

DRAFT




(a) AUC of $\{\widehat{\Theta}\}$ (b) AUC of $\{\widehat{\boldsymbol{B}}\}$

Figure 1: AUC for different percentages of $\lambda_{max}$ and $\rho_{max}$, and different values of $\sigma_{\boldsymbol{x}}$.

plot on the left of figure 1, the plot on the right shows a light effect of $\rho$ on the ROC curve for $\{\widehat{\boldsymbol{B}}\}$, which confirms what is known in the literature.

## 4. Conclusion

Our model employs the multivariate expansion proposed by (10), to decompose the variables into two components: the stochastic component, which does not depend on the continuous domain, and the deterministic component, which depends on the domain of the variables. To the best of our knowledge, the proposed model is the first in which the analysis of the stochastic component is used to infer the structure of the dependencies among variables on both the side of the conditional independence relations of the variables of a functional response process and their dependencies on the functional covariates. The simulation study shows good performance of the cfGGM in terms of adjacency regression and precision matrices.

**Acknowledgements.** Luigi Augugliaro and Gianluca Sottile gratefully acknowledge financial support from the University of Palermo (FFR2021-22).

DRAFT

## References

[1] Augugliaro, L., Sottile, G., and Vinciotti, V. The conditional censored graphical lasso estimator. *Statistics and Computing 30* (2020), 1273–1289.

[2] Danaher, P., Wang, P., and Witten, D. M. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B 76*, 2 (2014), 373–397.

[3] Huang, F., Songcan, and Huang, S.-J. Joint estimation of multiple conditional Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems 29*, 7 (2018), 3034–3046.

[4] Lauritzen, S. L. *Graphical Models*. Oxford University Press, Oxford, 1996.

[5] Lee, K.-Y., Ji, D., Li, L., Constable, T., and Zhao, H. Conditional functional graphical models. *Journal of the American Statistical Association 118*, 541 (2023), 257–271.

[6] Qiao, X., Guo, S., and James, G. M. Functional graphical models. *Journal of the American Statistical Association 114*, 525 (2019), 211–222.

[7] Sottile, G., Augugliaro, L., Vinciotti, V., Arancio, W., and Coronnello, C. Sparse inference of the human hematopoietic system from heterogeneous and partially observed genomic data. https://arxiv.org/abs/2206.09863, 2022.

[8] YIN, J., AND LI, H. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*, 4 (2011), 2630–2650.
[9] YUAN, M., AND LIN, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, 1 (2007), 19–35.
[10] ZAPATA, J., OH, S. Y., AND PETERSEN, A. Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika 109*, 3 (2022), 665–681.

DRAFT

# Does the Inflation Factor need tuning? Simulation-based adjustment for Outlier Detection via the Functional Boxplot

Annachiara Rossi[a], Andrea Cappozzo[a], and Francesca Ieva[a,b]

[a]MOX, Department of Mathematics, Politecnico di Milano;
annachiara.rossi@mail.polimi.it, andrea.cappozzo@polimi.it,
francesca.ieva@polimi.it

[b]Health Data Science Center, Human Technopole

DRAFT

**Abstract**

The detection of outliers in functional data analysis (FDA) is crucial as unusual curves can lead to incorrect inference and biased parameters estimation. However, the identification of such anomalous samples may prove difficult in the infinite-dimensional space containing such data. To address this issue, we propose a simulation-based adjustment of the fence inflation factor in the functional boxplot, a widely employed tool in FDA. The adjustment is performed by controlling the proportion of observations considered anomalous in a purified population based on the original one. To do so, robust estimators of location and scatter are required: we compare the performance of multivariate procedures, suitable for the *small N, large P* problems, and functional operators for implementing the tuning. A simulation study highlights the benefit of the proposed method.

***Keywords:*** Functional data analysis, Functional boxplot, Outlier detection, Robust estimation

## 1. Introduction and motivation

Functional data analysis (FDA) has become increasingly popular in recent years. The infinite-dimensional nature of these observations makes traditional multivariate techniques inappropriate for this complex data objects. Particularly, one field affected by this limitation is functional outlier detection, which is a crucial step in data exploration for ensuring the correctness of any subsequent statistical analysis. Outliers are recognized as curves that diverge from the common pattern of the data and should therefore be examined to determine if they are the result of errors or noise, or if they contain important information about the phenomenon under study. In this context, a variety of outlying behaviors can be observed: a comprehensive taxonomy of functional outliers is provided in (3).

The main focus of the present paper will be *amplitude outliers*, i.e., the direct generalization of multivariate outliers to the functional setting. The functional boxplot, introduced in (10), is a powerful tool for identifying and visualizing amplitude outliers: it is based on the classical boxplot, in which the central region of the data is represented by a box, while the fences are determined by

$$[Q1 - F \cdot IQR, Q3 + F \cdot IQR],$$

where $F$ is a factor that inflates the range, $Q1$ and $Q3$ are the first and third quartiles, respectively, and $IQR$ represents the interquartile range. To calculate quantiles, a center-outward ranking must be

defined: we make use of modified band depths (MBD, 6), which provides a method for determining the centrality of a functional signal within its sample. Observations whose paths cross the fences are flagged as outliers. Likewise for univariate data, an inflation factor $F = 1.5$ is generally employed for the purpose. The reason for this is due to the fact that probability of standing above the fences for a univariate Gaussian population can be computed as $2P(Z > Q3 + F \times IQR) = 2\Phi(4z_{0.25})$ which equals to a probability of $0.7\%$. Nonetheless, the presence of spatio-temporal dependence within curves may impact the performance of outlier detection. Specifically, using a constant factor of $1.5$ may be too high when correlation exists, as, typically, spatially correlated curves are more concentrated than those that are independent. To overcome this issue, we extend the original simulation-based adjustment for the inflation factor, initially developed in (11), considering state-of-the-art robust multivariate estimators and functional operators. In details, we leverage the contribution of recent literature in robust statistics for both high-dimensional and infinite-dimensional objects, adopting novel methods for the generation of clean synthetic populations via Gaussian processes.

The remainder of the paper is organized as follows. In Section 2.the tuning procedure for the inflation factor is described, and the considered robust estimators briefly presented. Section 3. reports the result of a simulation study, in which the beneficial effect of the proposed procedure is emphasized. Section 4. concludes the paper.

## 2. Tuning $F^*$ via robust estimators

As discussed in the previous section, a constant value of the inflation factor $F = 1.5$ is too restrictive when it comes to outlier detection in a functional setting. To make the functional boxplot sample-specific and data-driven, a simulation-based adjustment of $F$ is proposed. Building upon the original idea of (11), we aim at generating outlier-free samples possessing the main characteristics of the original dataset at hand. Clearly, in order to do so the parameters of the process must not be impacted by extreme curves, and thus robust procedures are needed. In details, the algorithmic pipeline can be summarized as follows:

1. Given a dataset of $N$ curves, estimate the mean function via a *robust* estimation of location and the covariance function by means of a *robust* estimate of dispersion,
2. Simulate a new dataset $\tilde{X}_b$ of dimension $N$ sampling from a Gaussian process having as mean and covariance functions the estimates computed in step 1,
3. Compute $C_{0.5}$ - the $50\%$ deepest region - and the adjusted inflation factor for the current replication $b$, referred to as $F_b$, given by the value $F_b$ minimizing the probability of spotting no outliers in $\tilde{X}_b$, that is
$$F_b = argmin_F P\left(\tilde{X}_b \notin F \cdot C_{0.5}\right) - 2\Phi\left(4z_{0.25}\right)$$
4. Repeat 2. and 3. $B$ times, collecting $F_b$, $b = 1, \ldots, B$,
5. Build the functional boxplot for the original data, using as inflation factor
$$F^* = \frac{1}{B}\sum_{b=1}^{B} F_b.$$

Undoubtedly, the most delicate task is performed in step 1, where suitable robust estimators must be selected. While for the mean function the median curve (i.e., the sample with highest MBD) can readily be adopted, for the covariance structure several options are at our disposal. We distinguish between robust multivariate estimators and functional operators. Within the former class of approaches, we consider Ledoit-Wolf (as a non robust benckmark, 5), OGK (7), Minimum Regularized Covariance Determinant (MRCD, 1) and Kernel MRCD (9) estimators. For the latter, we include Spherical Covariance Operator (2), Median Covariation operator (4) and Kendall's $\tau$ function (12).

The simulation phase outlined in step 2 slightly differs when either multivariate estimators or functional operators are considered. For the first set of estimators, we employ a finite-dimensional approximation and we directly sample from a multivariate Gaussian distribution with robustly estimated parameters.

DRAFT

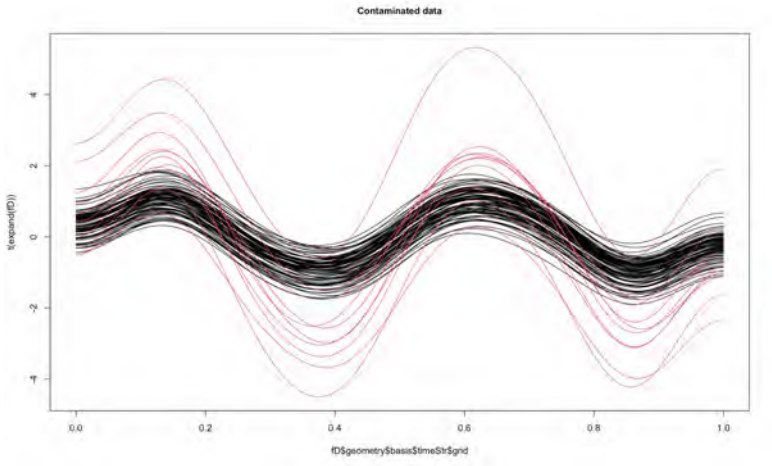



Figure 1: Example of simulated dataset according to the DGP outlined in Section 3. Uncontaminated curves are depicted in black, while red samples identify magnitude outliers.

When functional operators are considered, synthetic samples are built exploiting the Karhunen-Loève decomposition

$$X = \mu + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\zeta_i\phi_i,$$

where $\zeta_i \sim \mathcal{N}(0,1)$, $\{\lambda_i, \phi i\}_{i=1,\ldots,\infty}$ are the eigencouples of the covariance function, and the eigenvalues $\{\lambda_i\}_{i=1,\ldots,\infty}$ are in decreasing order of magnitude. This expansion can be truncated at $L$ components capturing most of the data variability: in our analysis, $L$ is set equal to 10.

A promising simulation study, showcasing the effectiveness of both approaches for tuning $F$ in the functional boxplot for outlier detection is presented in the next section.

DRAFT

## 3. Simulation study

The considered data-generating process (DGP) for the uncontaminated data is

$$X_i(t) = \sin(4\pi t) + \epsilon_i(t),$$

where $\epsilon_i(t)$ denotes a centered Gaussian process with Exponential Covariance function

$$C(s,t) = \alpha e^{-\beta|s-t|},$$

with $s$ and $t$ elements of the interval $I = [0,1]$. The parameters $\alpha$ and $\beta$ are set to $0.12$ and $0.4$, respectively, which results in a low level of variability and high degree of autocorrelation. For the purpose of contamination, a portion of the observations in the dataset is replaced with outliers. The fraction of corrupted data is set to $15\%$: the contamination process involves inflating such a proportion of curves by multiplying the mean function by a randomly generated number $u \sim U(2,3)$, that is, for a contaminated curve the mean function is equal to $u \times \sin(4\pi t)$, $t \in I = [0,1]$. A sample obtained by the considered DGP is displayed in Figure 1. The adjustment procedure based on the estimators listed in the previous section are compared in terms of False Positive (FPR) and True Positive (TPR) outlier detection rate. A no-adjustment option, with $F$ kept fixed at $1.5$ is also included in the comparison.

The empirical distribution of the $F^*$ values obtained by tuning the functional boxplot over $100$ repetitions of the simulated experiment is displayed in Figure 2. All estimators but $kMRCD$ induce $F^*$s that are lower than the default value of $1.5$, corroborating the need of an adjustment procedure. For the functional estimators, the resulting distribution of $F^*$ is similar and they seem to have the most stable and robust behavior. Among the multivariate alternatives, the MRCD estimator (with hyper-parameter $\alpha$ set equal to either $0.5$ or $0.75$) appears to agree with the performance of the functional ones, whilst

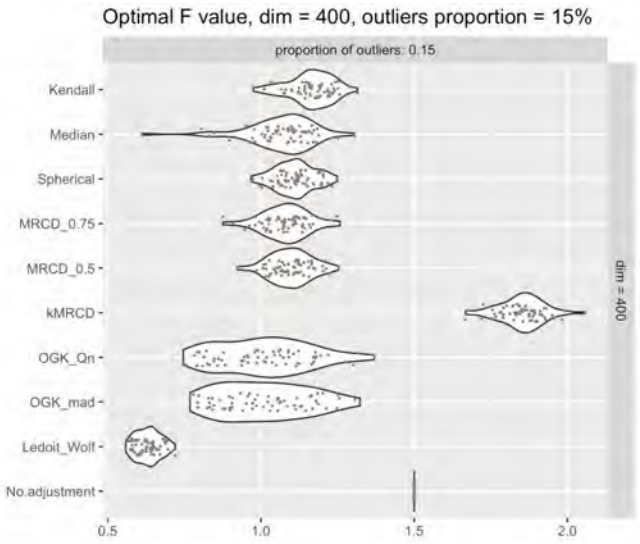



Figure 2: Violin plots of the optimal F values tuned with the adjustment procedures listed in Section 2. over 100 repetitions of the simulated experiment.

Ledoit-Wolf is biased towards smaller values of $F^*$ being it sensitive to the presence of outliers. This behavior will lead to many False Positives, i.e., curves wrongly flagged as outliers. On the other hand, kMRCD is very conservative, resulting in very few False Positives but missing many truly outlying curves. This behavior is apparent when looking at Figure 3, where a scatterplot of average TPR against average FPR for each estimator is showcased. OGK results (obtained considered both $Q_n$ and $MAD$ estimates of scale) display slightly lower TPR and higher FPR with respect to MRCD, Spherical, Median, and Kendall's $\tau$ estimates. Overall, the Median Covariation brings the best trade-off between identifying the true anomalous curves whilst not producing too many false positives for this data contamination scenario.

All in all, given the poor performance of the no-adjustment setting, it is clear that tuning procedures are needed in order to improve the functional boxplot as an effective tool for outlier detection.

DRAFT

## 4. Conclusion and discussion

The paper has shown the importance of developing ad-hoc procedures when dealing with outliers in infinite-dimensional spaces. Specifically, being the functional boxplot one of the most-widely employed tools for the purpose, the importance of adjusting its inflation factor $F$ in a distribution-free manner has been discussed. To improve the robustness of the methodology, the focus has been on using either multivariate or functional estimators for the covariance function. A taxonomy of the up-to-date choices for performing the tuning has been described and subseqeuntly employed in a simulated setting, showcasing better results than keeping $F$ fixed.

Future research may involve extending the proposed procedure to the multivariate functional scenario. Although functional boxplots for multivariate curves have recently been introduced in the literature (8), the adaptation of the tuning procedure to this context remains a challenge that requires attention. Future effort will focus on exploring multiple possibilities that are currently under consideration.

## References

[1] K. Boudt, P. J. Rousseeuw, S. Vanduffel, and T. Verdonck. The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128, feb 2020.

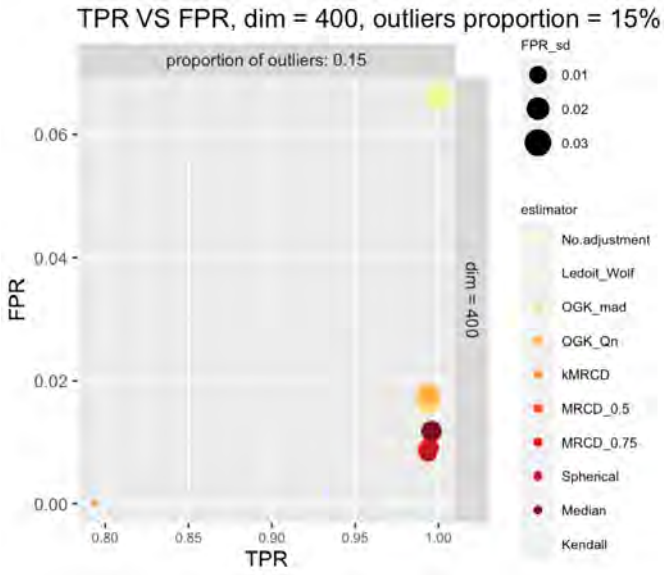



Figure 3: Scatterplot of True Positive Rate vs False Positive Rate averaged over 100 repetitions of the simulated experiment. The size of the points depends on the empirical variability of the FPR metric.

DRAFT


[2] D. Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

[3] M. Hubert, P. J. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods and Applications*, 24(2):177–202, 2015.

[4] D. Kraus and V. M. Panaretos. Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99(4):813–832, dec 2012.

[5] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, feb 2004.

[6] S. López-Pintado and J. Romo. On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, 104(486):718–734, jun 2009.

[7] Y. Ma and M. G. Genton. Highly Robust Estimation of Dispersion Matrices. *Journal of Multivariate Analysis*, 78(1):11–36, jul 2001.

[8] Z. Qu and M. G. Genton. Sparse Functional Boxplots for Multivariate Curves. *Journal of Computational and Graphical Statistics*, 31(4):976–989, oct 2022.

[9] J. Schreurs, I. Vranckx, M. Hubert, J. A. Suykens, and P. J. Rousseeuw. Outlier detection in non-elliptical data by kernel MRCD. *Statistics and Computing*, 31(5):1–18, 2021.

[10] Y. Sun and M. G. Genton. Functional Boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, jan 2011.

[11] Y. Sun and M. G. Genton. Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, 23(1):54–64, 2012.

[12] R. Zhong, S. Liu, H. Li, and J. Zhang. Robust functional principal component analysis for non-Gaussian longitudinal data. *Journal of Multivariate Analysis*, 189:104864, 2022.

# Functional Graphical Models to map Brexit debate on Twitter

Nicola Pronello[a], Emiliano del Gobbo[b], Lara Fontanella[a], Rosaria Ignaccolo[c], Luigi Ippoliti[a], and Sara Fontanella[d]

[a]Università degli Studi "G. d'Annunzio" Chieti - Pescara; nicola.pronello@unich.it, lara.fontanella@unich.it, luigi.ippoliti@unich.it
[b]Università degli Studi di Foggia; emiliano.delgobbo@unifg.it
[c]Università degli Studi di Torino; rosaria.ignaccolo@unito.it
[d]Imperial College London; s.fontanella@imperial.ac.uk

DRAFT

**Abstract**

In recent years a literature on multivariate functional graph models has been developed. The graphical representation of the conditional dependence among a finite number of random variables is indeed appealing in different applications, such as e.g. the analysis of the brain connectivity. We want to investigate a novel extension of this methodology, considering random functions spatially and temporally correlated. A motivating case study is the analysis of the semantic network that tracks the change of the Brexit debate on Twitter across UK during a particular time frame. By considering the change in time of a word usage as a functional realization, a semantic network on the topic of interest is defined by a graphical representation of the conditional dependence among functional variables.

***Keywords:*** Functional graphical models, Functional data analysis, Kernel Smoothing, Semantic network

## 1. Introduction

In recent years, literature on graphical models for functional data has been developed (6; 4). Indeed visualizing an estimated graph representing relationships between functional variables can be very effective to represent conditional dependence among them. In this framework, we contribute to extend this methodology by considering functional graphs that are spatially and temporally correlated; in particular they are supposed to vary on a spatio-temporal lattice, and for their estimation only limited measurements are available. This setting is motivated by data representing daily word usage on the Brexit debate on Twitter during 13 months (temporal units) and 41 districts in UK (spatial units) without replicates. Our main goal is to estimate a semantic network representing the connections of words used in such a debate. By means of functional graphical models it is possible to represent the connection between words from their monthly usage trends in Twitter. So doing, we offer a different insight on a public debate, moving beyond classical semantic networks built from co-occurrences of words in a sentence/tweet.

## 2. Motivating case study: Twitter conversations about Brexit

We dispose of data extracted from Twitter regarding the Brexit debate, as collected and pre-processed by del Gobbo et al. (1), by combining around one million of tweets with hashtag *Brexit* spanning in

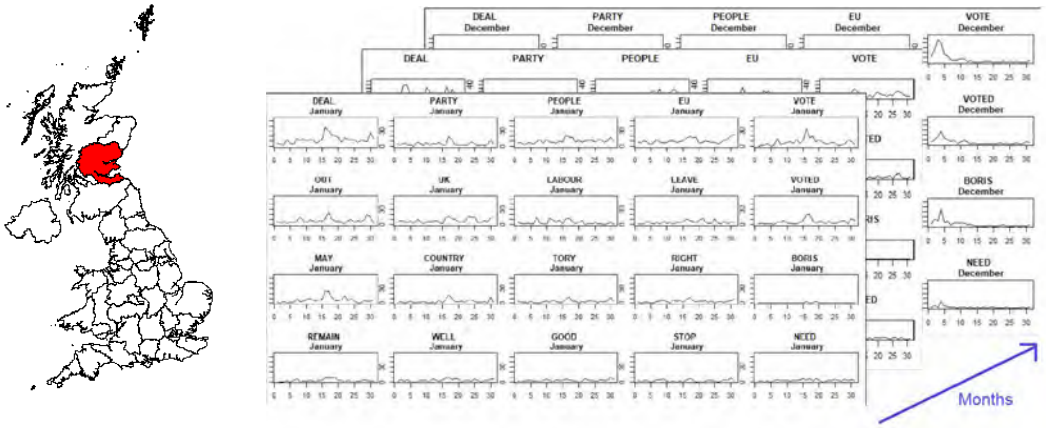



Figure 1: Example for a specific district (Eastern Scotland)

a range of 13 months (from 31 December 2018 to 9 February 2020) and geolocalized in the United Kingdom territory with 41 NUTS-2 (Nomenclature of territorial units for statistics) districts. We consider the daily usage of the first $p$ most common words in each district in a time window of a month. Thus we have $p$ daily time series (one for the daily usage of each word) for each month in each district, as shown in Figure 1 for the particular case of Eastern Scotland, and each time series can be seen as a noisy realization of a functional random variable. Hence the domain of the considered multiple functional data is a spatio-temporal irregular lattice $\mathcal{L} = \mathcal{S} \times \mathcal{M}$ where $\mathcal{S}$ represents the set of 41 districts and $\mathcal{M}$ the set of 13 months.

DRAFT

## 3. Functional graphical models

Let us consider a spatio-temporal irregular lattice $\mathcal{L} = \mathcal{S} \times \mathcal{M}$ where $\mathcal{S}$ represents a set of spatial units (districts) and $\mathcal{M}$ a set of time frames. Let us denote with $\mathbf{c} = (s, t)$ the spatio-temporal coordinate of a generic element of $\mathcal{L}$ and let $\mathbf{Y^c} = \{Y_1^{\mathbf{c}}, ..., Y_p^{\mathbf{c}}\}$ be a $p$-dimensional functional random variable taking values in $(L_2(\mathcal{T}))^p$, with $\mathcal{T} = [a, b]$, defined at a specific site (district) and time frame (month). The vector of mean functions is $E[\mathbf{Y^c}(\tau)] = \mu^{\mathbf{c}}(\tau) = \{\mu_j^{\mathbf{c}}(\tau)\}_{j=1,..,p}$, while the matrix of functional covariances is written as $\mathbf{C}(\mathbf{Y^c}(\tau), \mathbf{Y^c}(\tau')) = \mathbf{C}_{\mathbf{Y}}^{\mathbf{c}}(\tau, \tau') = \{C_{j,r}^{\mathbf{c}}(\tau, \tau')\}_{(j,r)=1,..,p}$.

By considering a continuous orthonormal basis functions, $\phi_{j,1}, \phi_{j,2}, \ldots$, the well-known Karhunen-Loéve expansion (KLE, see e.g. (2)) allows us to represent each functional variable $Y_j(\tau)^{\mathbf{c}}$ as

$$Y_j^{\mathbf{c}}(\tau) = \sum_{d=1}^{\infty} a_{j,d}^{\mathbf{c}} \phi_{j,d}(\tau),$$

where the expansion coefficients $a_{j,d}^{\mathbf{c}} = \int_{\mathcal{T}} Y_j^{\mathbf{c}}(\tau)\phi_{j,d}(\tau)d\tau$ are uncorrelated random variables, and then we consider with $D < \infty$ its truncated version

$$\tilde{Y}_j^{\mathbf{c}}(\tau) = \sum_{d=1}^{D} a_{j,d}^{\mathbf{c}} \phi_{j,d}(\tau).$$

Let $\mathbf{A^c} = \{\mathbf{a}_1^{\mathbf{c}}, ..., \mathbf{a}_p^{\mathbf{c}}\} \in \mathcal{R}^{pD}$ be the $(D \times p)$ matrix collecting the expansion coefficients associated with the $p$ functions and let

$$\Sigma^{\mathbf{c}} = \{Cov(\mathbf{a}_j^{\mathbf{c}}, \mathbf{a}_r^{\mathbf{c}})\}_{(j,r)=1,...,p} = \{\Sigma_{j,r}^{\mathbf{c}}\}_{(j,r)=1,...,p}$$

be the $(Dp \times Dp)$ matrix collecting all covariances between expansion coefficients of all $p$ functional variables. For the functional covariances we have the approximation:

$$C_{j,r}^{\mathbf{c}}(\tau, \tau') = C(Y_j^{\mathbf{c}}(\tau), Y_r^{\mathbf{c}}(\tau')) \approx C(\tilde{Y}_j^{\mathbf{c}}(\tau), \tilde{Y}_r^{\mathbf{c}}(\tau')),$$

and since vectors $\mathbf{a}_j^{\mathbf{c}}$ share the same information with $\tilde{Y}_j^{\mathbf{c}}(\tau)$ we can work with $\Sigma_{j,r}^{\mathbf{c}}$ and the matrix of partial correlations $\Theta^{\mathbf{c}} = (\Sigma^{\mathbf{c}})^{-1} = \{\Theta_{j,r}^{\mathbf{c}}\}_{(j,r)=1,\dots,p}$.

The matrix of partial correlation just defined is a key ingredient to evaluate relationships among functional variables in the framework of the graphical models. To this goal, for each coordinate $\mathbf{c}$, let $G^{\mathbf{c}} = \{V, E^{\mathbf{c}}\}$ be a undirected graph where $V = \{1, ..., p\}$ represents the set of vertices corresponding to the $p$ random functions and $E^{\mathbf{c}} \subseteq \{(j,r) \ \in V \times V, \ j \neq r\}$ represents the set of edges specified by means of

$$(j,r) \notin E^{\mathbf{c}} \ \text{ if } \ ||\Theta_{j,r}^{\mathbf{c}}||_F = 0$$

where $||\cdot||_F$ denotes the Frobenius norm. The condition below is equivalent to conditional independence between the $j$-th and $r$-th variables in the Gaussian case and, in analogy with Qiao et al. (2019, (6)), by considering the $p$ functional random variables as vertices of the graph $G^{\mathbf{c}}$ we can say that $\mathbf{Y}^{\mathbf{c}}$ follows a *Functional Graphical Model (FGM)*.

To retrieve a network, for every $\mathbf{c}$, among $p$ functional random variables in a *FGM* we need to estimate $\Theta^{\mathbf{c}}$. Moreover, since we are interested in underlying just the most important connections among words, we consider a sparse estimator for $\Theta^{\mathbf{c}}$ . To this goal, we consider the Functional graphical lasso criterion (`fglasso`) introduced by (6) as a block extension of the classical `glasso` algorithm (8; 3). The sparsity in the precision matrix is achieved by imposing a group lasso penalty, so that the estimated $\widehat{\Theta}^{\mathbf{c}}$ at each $\mathbf{c}$ is obtained by solving an optimization problem:

$$\widehat{\Theta}^{\mathbf{c}} = \underset{\Theta^{\mathbf{c}}}{\operatorname{argmax}} \left( \log\det \Theta^{\mathbf{c}} - \operatorname{trace}(\hat{\Sigma}^{\mathbf{c}}\Theta^{\mathbf{c}}) - \lambda \sum_{j \neq r} ||\Theta_{j,r}^{\mathbf{c}}||_F \right), \tag{1}$$

where $\hat{\Sigma}^{\mathbf{c}}$ is an estimate of $\Sigma^{\mathbf{c}}$ (that needs to be obtained) and $\lambda$ is a nonnegative tuning parameter. The group lasso penalty $\lambda \sum_{j \neq r} ||\Theta_{j,r}^{\mathbf{c}}||_F$ shrinks all the elements in $\Theta_{j,r}^{\mathbf{c}}$ towards zero (or all nonzero, that is the case of an estimate edge between $Y_j^{\mathbf{c}}$ and $Y_r^{\mathbf{c}}$ ) leading to a sparser $\widehat{\Theta}^{\mathbf{c}}$ in a blockwise way, and consequently to a sparser graph $\hat{G}^{\mathbf{c}}$, when $\lambda$ increases. Obviously, if $\lambda = 0$ there is no penalty and the choice of this regularization parameter is crucial: there exist suitable AIC indexes used in (6), but instead we adopt an heuristic choice to ease the interpretability of the resulting semantic network by considering an overall percentage of nonzero links below 10%.

DRAFT

## 4. Nonparametric estimator of $\Sigma^{\mathbf{c}}$

The reconstruction of sparse network structures by means of Equation 1 is possible if we have an estimation of $\Sigma^{\mathbf{c}}$ varying on the lattice $\mathcal{L}$, that is one for spatio-temporal coordinate $\mathbf{c} = (s,t)$ (indicating district and month in our case study). However, constructing an estimator of $\Sigma^{\mathbf{c}}$ represents a challenge because we deal with the case of extremely sparse data since in each $\mathbf{c}$ we observe only one realization of $\mathbf{Y}^{\mathbf{c}}$. By considering mean-corrected functional data for the sake of simplicity, a naive estimate of $\Sigma^{\mathbf{c}}$ is given by the raw covariance

$$\Sigma^{*\mathbf{c}} = \left\{\mathbf{\Sigma}_{j,r}^{*\mathbf{c}}\right\}_{j,r=1,\dots,p} = \left\{\mathbf{a}^{\mathbf{c}}{}_j \mathbf{a}^{\mathbf{c}}{}_r^T\right\}_{j,r=1,\dots,p},$$

that exploits only the information in one datum and is not full rank since by construction it holds $\operatorname{rank}(\Sigma_{j,r}^{*\mathbf{c}}) = 1, \forall j,r$. Then to obtain a reliable estimator of $\Sigma^{\mathbf{c}}$ we propose to borrow information from the neighbours of the unit with coordinate $\mathbf{c}$ in the lattice by means of linear smoothing.

Given $n$ units in the spatio-temporal irregular lattice $\mathcal{L}$ with coordinates $\mathbf{c}_i$, with $i = 1, \dots, n$, we consider the class of linear smoother estimators defined by

$$\widehat{\Sigma}^{\mathbf{c}} = \sum_{i=1}^{n} \omega_i(\mathbf{c}) \Sigma^{*\mathbf{c}_i}$$

where the coefficients $\omega_i(\mathbf{c})$ of the linear combination need to be determined; this class includes the Gaussian process regression estimator as well as the Kernel smoother and local polynomial estimator.

Finding $\omega(\mathbf{c}) = (\omega_1(\mathbf{c}), ..., \omega_n(\mathbf{c}))$ is equivalent to solve an optimization problem:

$$\widehat{\Sigma}^{\mathbf{c}} = \text{argmin}_{\Sigma^{\mathbf{c}}} \sum_{i=1}^{n} K(\mathbf{c}_i, \mathbf{c}) d(\Sigma^{\mathbf{c}}, \Sigma^{*\mathbf{c}_i}),$$

where and $K(\cdot, \cdot) : \mathcal{L} \times \mathcal{L} \to R$ is a kernel function and $d(\cdot, \cdot)$ is a suitable distance between covariance matrices. The choice of the distance $d$ is not uniquely identified and different choices may or may not incorporate the constraints of the space of the positive definite matrices.

**Determining $K$ from space-time contiguity** In this work, we propose to construct kernel weights that are suitable for a spatio-temporal lattice domain as in our case study. Usually a Kernel function is defined as a function of distances, and so in a spatial lattice distances among areas would be among their barycenters while in a temporal lattice distances would be lags.

To consider properly the nature of the spatio-temporal domain where graphs live, we instead take a contiguity point of view both in space and time and define the kernel by means of a Laplacian $\mathbf{L}$. Let $\mathbf{W}^{\text{Space}}$ be the adjacency matrix for the areal units such that entries are equal to 1 if two areal units are neighbours (contiguous regions), and 0 otherwise; and let $\mathbf{W}^{\text{Time}}$ be the adjacency matrix for the time frames (months) with element equal to 1 when a month before or after is considered (like a 3 months moving average window). In order to have a global adjacency matrix for the units in the spatio-temporal lattice $\mathcal{L}$ we take the Kronecker product and define

$$\mathbf{W} = \mathbf{W}^{\text{Space}} \otimes \mathbf{W}^{\text{Time}},$$

and then we consider the Laplacian

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

with $\mathbf{D} = diag\{\mathbf{W}\mathbf{1}_n\}$.

Finally, as kernel values in the vector $\omega(\mathbf{c})$ for the linear smoother we use the elements of the matrix

$$\mathbf{K} = (\mathbf{I} + \gamma\mathbf{L})^{-1},$$

where $\gamma$ is a smoothing parameter (to be fixed) that behaves as a kernel bandwidth. In fact, $\gamma$ tunes the values of the weights and the extension of the neighborhood of each unit in $\mathcal{L}$ and for $\gamma \to \infty$ one has $\omega_i \to 1/n$.

DRAFT

## 5. Results about Brexit

The proposed methodology allows to estimate semantic networks, tracking the change of the Brexit debate on Twitter, for each of the 41 districts in UK along months from January 2019 to January 2020, having borrowed information by a spatio-temporal neighborhood. We take the first $p = 20$ most common words in the dataset and consider the daily usage of each word for each month in each district as (spatially and temporally correlated) functional data. Looking at the sequence of estimated semantic networks in a specific district along time (as e.g. those in Figure 2) we can detect which word is really central in the debate month by month. The estimated networks show an interesting pattern: during the first months of 2019 the word `MAY` (the former prime minister surname) is connected with other words in the debate; this holds until the summer when she resigned in favor of Boris Johnson, and then the word `BORIS` becomes connected with others in August and October. Figure 2 shows the estimated networks in Inner West London for the months January, February and March 2019, with two different values of $\gamma$, namely $\gamma = 0.2$ and $\gamma = 2$. Even with only three months we can note that the estimated network changes enough with a smaller value of $\gamma$ (first row in Figure 2) while it persists with a larger value (second row in Figure 2): $\gamma$ controls the 'width' of the kernel used in the smoothing process, i.e. the magnitude of elements of the matrix $\mathbf{K}$ . In the first case, with smaller $\gamma$, the only weights different from zero are those related to the spatial neighbours in the same month; while a larger value acts as a large kernel bandwidth and leads to an estimate close to an average.

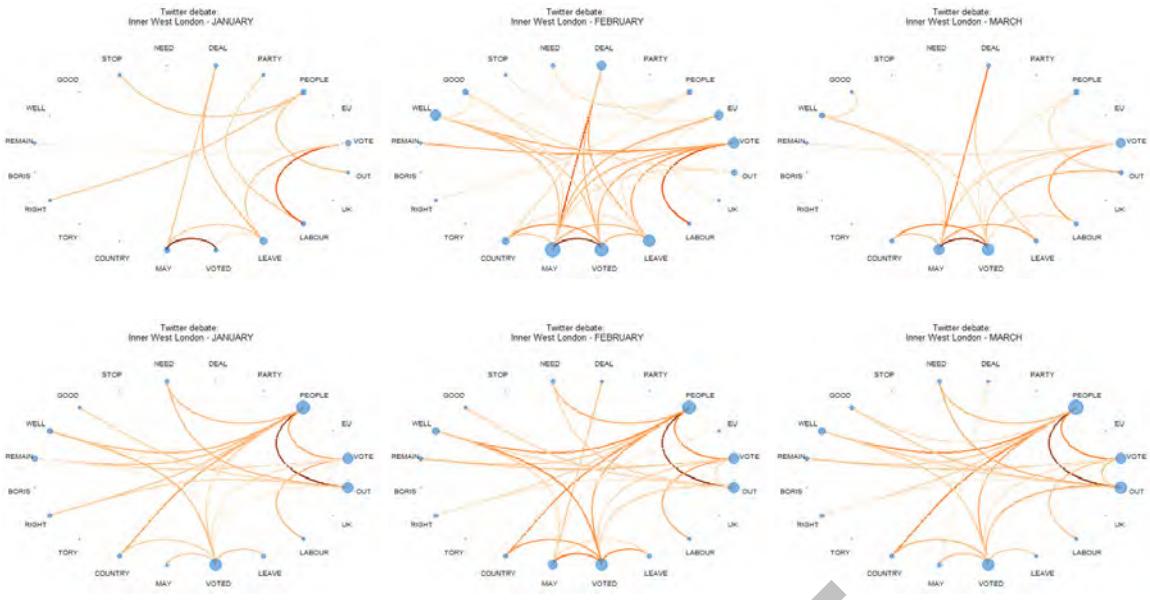



Figure 2: Estimated networks in Inner West London for the months January, February and March 2019, with $\gamma = 0.2$ (*first row*) and with $\gamma = 2$ (*second row*). Each vertex corresponds to a functional variable representing the usage of a word, the area of the blue dots is proportional to how many times that word is connected to the other words. The color and the thickness of the estimated edge change with the norm of $\hat{\Theta}_{j,r}$ for two words $j$ and $r$.

## 6. Discussion

By considering the change in time of a word usage as a functional realization, we investigate how to estimate semantic networks via functional graphical models. Actually, by penalizing, we obtain sparse solutions that uncover interesting connections among variables/words.
In future developments, the choice of the smoothing parameter $\gamma$ needs to be addressed and new strategies to preserve positive definitiveness in the estimates investigated. As for the Brexit debate analysis, the number of words $p$ considered will be substantially increased.

DRAFT

## References

[1] del Gobbo, E., Fontanella, S., Sarra, A., Fontanella, L.: Emerging Topics in Brexit Debate on Twitter Around the Deadlines. Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, **156(2)**, 669-688 (2021)

[2] Daw, R., Simpson, M., Wikle, C.K., Holan, S.H., Bradley. J.R.: An Overview of Univariate and Multivariate Karhunen Loéve Expansions in Statistics. J Indian Soc Probab Stat, **23**, 285-326 (2022)

[3] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso, Biostatistics, **9(3)**, 432-441 (2008)

[4] Lee, K.Y., Ji, D., Li, L., Constable, T., Zhao, H: Conditional Functional Graphical Models, Journal of the American Statistical Association (2021) doi: 10.1080/01621459.2021.1924178

[5] Petersen, A., Deoni, S., Müller H.G.: Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional co-evolution of myelination in the developing brain. The Annals of Applied Statistics, **13(1)**, 393-419 (2019)

[6] Qiao, X., Guo, S., James, G.M.: Functional graphical models. Journal of the American Statistical Association, **114(525)**, 211-222 (2019)

[7] Yin, J., Geng, Z., Li, R., Wang, H.: Nonparametric covariance model. Statistica Sinica, **20(1)**,

469-479 (2010)
[8] Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94(1)**, 19-35 (2007)

DRAFT

# Measuring Dependence in Multivariate Functional Datasets

Francesca Ieva[a], Michael Ronzulli[a], and Anna Maria Paganoni[a]

[a]MOX Lab, Department of Mathematics, Politecnico di Milano, Milan 20133, Italy;
francesca.ieva@polimi.it, michael.ronzulli@mail.polimi.it,
anna.paganoni@polimi.it

**Abstract**

We generalize the notion of Spearman correlation coefficient to situations where the observations are curves generated by a stochastic processes. In particular, we propose a consistent estimator of the Spearman index and we use this notion to define the Spearman matrix, a mathematical object expressing the pattern of dependence among the components of a multivariate functional dataset. Finally, the notion of Spearman matrix is exploit to analyze two different populations of multivariate curves (specifically, Electrocardiographic signals of healthy and unhealthy people), in order to test if the pattern of dependence between the components is statistically different in the two cases.

***Keywords:*** Spearman correlation coefficient, multivariate functional data, ECG signals.

DRAFT

## 1. Methods

The Spearman index is initially presented in a multivariate framework as a non-parametric measure of association between two random variables X and Y defined as the Pearson correlation coefficient $\rho_p$ between the ranks of X and Y without requiring any assumptions on the distribution of the variables.
Let $X_i$ and $Y_i$ be i.i.d from the law of X and Y, respectively. Through the notion of grades, interpreted as the relative position of the observation $x_i$ (resp. $y_i$) in the set **x** (resp. **y**), the sample version of the Spearman index is defined as the sample Pearson correlation coefficient of **u** and **v**:

$$\hat{\rho}_s(\mathbf{x},\mathbf{y}) = \hat{\rho}_p(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{(\sum_{i=1}^{n}(u_i-\bar{u})^2\sum_{i=1}^{n}(v_i-\bar{v})^2)^{\frac{1}{2}}},$$

where **u** and **v** are the estimated grades defined evaluating each observation in the empirical cumulative distribution function of the corresponding sample and $\bar{u}$, $\bar{v}$ stand for the sample means of **u** and **v**.
After defining grades for functional data the Spearman index $\rho_s$ defined for random variables has be extended to the case of two stochastic processes $X_t$ and $Y_t$, generalizing to the infinite dimensional framework.
To define the notion of grade for functions two concepts called the Inferior Length and Superior Length of a curve are defined as the foundation of a depth, where the sample version of both *IL-grade* and *SL-grade* of any fixed curve $x = x(t)$ quantify the relative position of $x$ with respect to the other curves of the

sample (see (5)).
The Spearman index for ($X_t$,$Y_t$) is defined as

$$\rho_s(X_t, Y_t) = \rho_p((IL - grade(X_t), IL - grade(Y_t)),$$

where $\rho_p$ denotes the Pearson correlation coefficient and *IL-grade*($\cdot$) is the grade associated to a stochastic process. The corresponding sample version is denoted by $\hat{\rho}_s$(**x**, **y**) and it is defined as

$$\hat{\rho}_s(\mathbf{x}, \mathbf{y}) = \hat{\rho}_p(IL_n - grade(\mathbf{x}), IL_n - grade(\mathbf{y})).$$

Subsequently the consistency of the sample version of the Spearman coefficient is proved using the fact that it can be expressed as a U-statistic. From the definition of UB-statistic (1) and the definition of a preorder between functions (*functional order based on grades* $\prec$) the Spearman correlation coefficient $\hat{\rho}_s$ can be extended to the functional case and it can be expressed as a UB-statistic,

$$U_n = \binom{n}{3}^{-1} \sum_{1 \le i_1 < i_2 < i_3 \le n} \Phi\{(X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2}), (X_{i_3}, Y_{i_3})\},$$

where

$$\Phi[(x_i, y_i), (x_j, y_j), (x_z, y_z)] = 6I(x_i \prec x_j, y_i \prec y_z) + 6I(x_j \prec x_i, y_z \prec y_i) - 3,$$

where *I* denotes the indicator function.
In analogy with Kendall's $\tau$ correlation coefficient, expressing the functional $\hat{\rho}_s$ as a UB-statistic, the consistency of functional $\hat{\rho}_s$ is proved applying Theorem 2 in (7), obtaining an important asymptotic result in the functional field also for the Spearman's coefficient.
**Theorem** (*Consistency of* $\hat{\rho}_s$) Let $(x_1, y_1), ..., (x_n, y_n)$ be a sample of independent and identical functional observations from (X,Y). Then,

$$\hat{\rho}_s \to \rho_s \text{ a.s. as } n \to \infty.$$

We construct a new mathematical object, the Spearman Matrix, in order to express the pattern of dependence among the components of a multivariate functional dataset. The sample Spearman Matrix $\widehat{SM}(\mathbf{X})$ is given by

$$\begin{bmatrix} \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_h) \\ \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_h) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_h) \end{bmatrix}_{t \in I},$$

where $\hat{\rho}_s(\mathbf{x}_i, \mathbf{x}_j)$ is the sample Spearman index computed on the bivariate functional dataset [$\mathbf{x}_i$,$\mathbf{x}_j$]. The great advantage with respect to the variance-covariance operator is the fact that the dependence among components is described through scalar indexes that may be tested in a suitable inferential context.
In (6) some simulation studies are presented with different functional bivariate datasets built ad hoc in order to test the performance of the Spearman index in correctly detecting the pattern of dependence.

## 2. Case study

Our data consist in a multivariate functional dataset containing the ECG traces of a population of healthy people and one composed by individuals affected by an heart disease called Left Bundle Brunch Block (LBBB). Each statistical unit (patient) is characterized by the 8-variate functional datum of his/her electrocardiogram, which describes his/her heart dynamics on the eight leads I, II, V1, V2, V3, V4, V5 and V6. The data are from PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) database. PROMETEO project has been started in 2008 with the aim of spreading the intensive use of ECGs as prehospital diagnostic tool. Each file contained in PROMETEO

DRAFT

| | $I$ | $II$ | $V1$ | $V2$ | $V3$ | $V4$ | $V5$ | $V6$ |
|---|---|---|---|---|---|---|---|---|
| $I$ | 1 | 0.382 | -0.069 | 0.303 | 0.327 | 0.386 | 0.439 | 0.456 |
| $II$ | | 1 | 0.036 | 0.202 | 0.500 | 0.596 | 0.605 | 0.611 |
| $V1$ | | | 1 | 0.674 | 0.372 | 0.146 | -0.001 | -0.046 |
| $V2$ | | | | 1 | 0.635 | 0.475 | 0.376 | 0.300 |
| $V3$ | | | | | 1 | 0.830 | 0.678 | 0.496 |
| $V4$ | | | | | | 1 | 0.869 | 0.662 |
| $V5$ | | | | | | | 1 | 0.811 |
| $V6$ | | | | | | | | 1 |

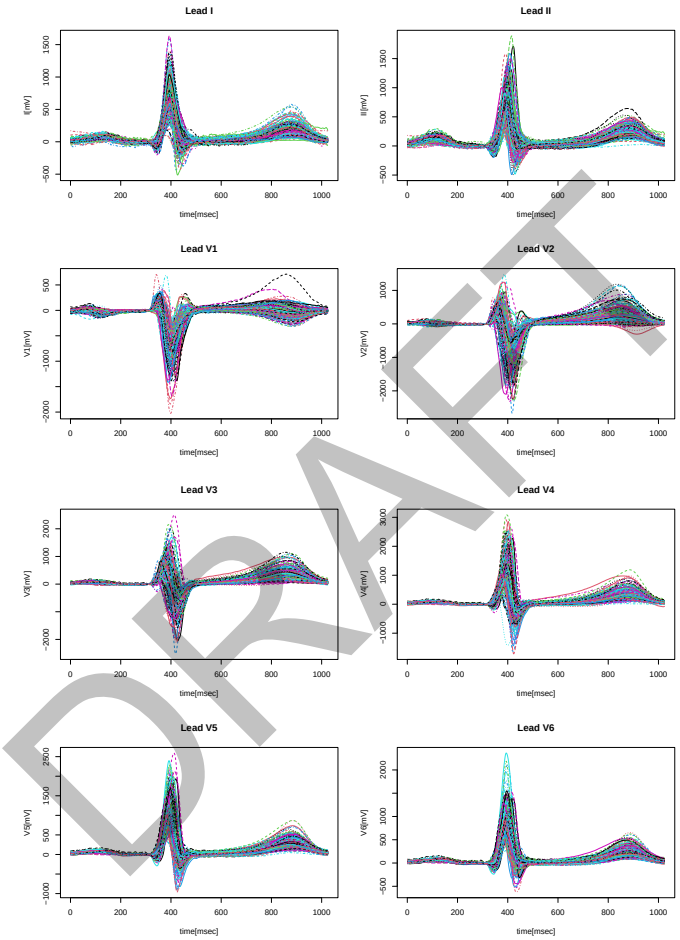
Table 1: Spearman Matrix $\widehat{SM}(\mathbf{X})$ for the physiological signals.

| | $I$ | $II$ | $V1$ | $V2$ | $V3$ | $V4$ | $V5$ | $V6$ |
|---|---|---|---|---|---|---|---|---|
| $I$ | 1 | 0.459 | −0.392 | -0.052 | -0.016 | 0.346 | 0.607 | 0.653 |
| $II$ | | 1 | -0.095 | 0.036 | 0.198 | 0.471 | 0.599 | 0.582 |
| $V1$ | | | 1 | 0.750 | 0.560 | 0.123 | −0.220 | −0.370 |
| $V2$ | | | | 1 | 0.734 | 0.363 | 0.010 | -0.150 |
| $V3$ | | | | | 1 | 0.688 | 0.246 | -0.036 |
| $V4$ | | | | | | 1 | 0.727 | 0.451 |
| $V5$ | | | | | | | 1 | 0.843 |
| $V6$ | | | | | | | | 1 |

Table 2: Spearman Matrix $\widehat{SM}(\mathbf{Y})$ for the LBBB signals.

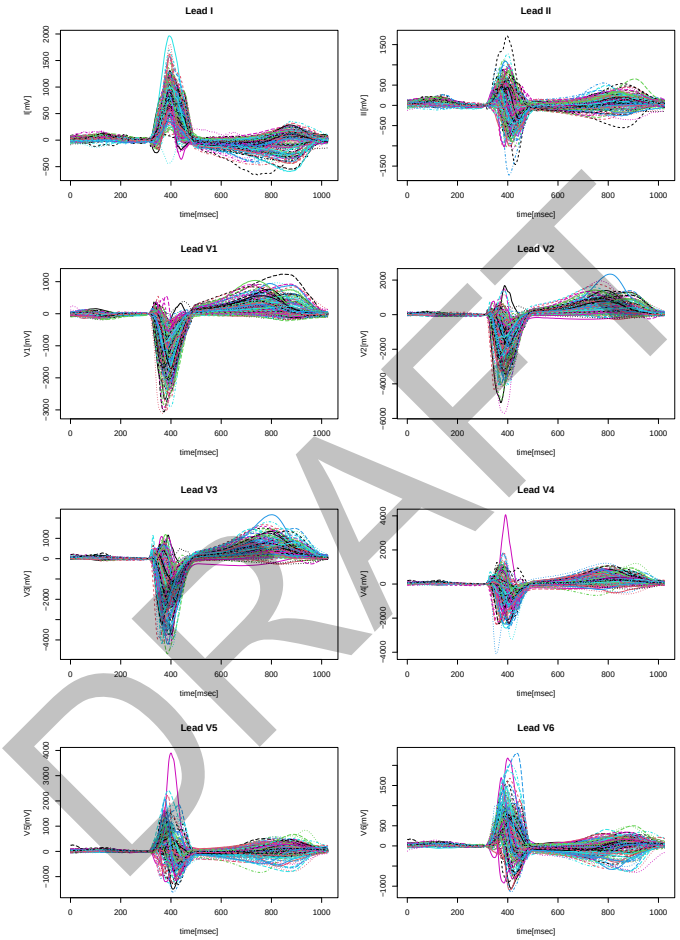
DRAFT

database can be associated to three sub-files, called *Details*, *Rhythm* and *Median*. For the aims of the analysis, only the last one is necessary. The *Median* file depicts a reference beat (obtained throught an authomatic filtering procedure applied to the Rhythm file) lasting 1.2 seconds on a grid of 1200 points. It provides, among others, 8 curves (one for each ECG lead) for each patient, representing patient's *Median* beat for that lead. This representative heartbeat is a trace of a single cardiac cycle (heartbeat), i.e., of a P wave, a QRS complex, a T wave, and a U wave. Actually PROMETEO database contains 6734 curves; among these, 1633 are healthy (i.e., not affected by cardiovascular diseases detectable through the ECG), whereas 5101 are affected by different heart diseases. See (2; 3; 4) for further details on the dataset and its use for statistical applications. In what follows we will focus just on one of the most common disease, that is easily detectable observing the ECG signal. It is a kind of Myocardial Infarction named Left Bundle Brunch Block (LBBB). In the PROMETEO dataset, 314 people are affected by this pathology. After suitable preprocessing and robustification (see (4) for more details) of the dataset, the final sample available for the analyses is composed by 1564 Physiological curves and 205 LBBB curves, discretized on a uniformly time grid T of 1024 points. Each patient is represented by his/her discretized multivariate signal, i.e., for $i = 1, ..., n$, $\Phi_i(t)$: $T \subset R \rightarrow R^8$. All the curves of the available sample are registered and denoised (see (2) for further details on wavelet denoising and landmarks registration adopted for preprocessing data). To fix the notation, we assume that the ECG signals of physiological and pathological patients are realizations of two different multivariate stochastic processes, $X_t = (X_t^1, X_t^2, ..., X_t^8)$ and $Y_t = (Y_t^1, Y_t^2, ..., Y_t^8)$, respectively. For the analysis, we construct two different multivariate functional dataset from the available samples: the first is denoted with $\mathbf{X}$ and collects $n_x = 200$ randomly chosen ECG signals from the population of the physiological (healthy) patients. In other words, $\mathbf{X}$ is a dataset $200 \times 8$ discretized functions, where the *i-th* row contains the multivariate curve (ECG) associated to the *i-th* selected patient. The second functional dataset is denoted with $\mathbf{Y}$ and contains the multivariate curves of $n_y = 200$ randomly chosen patients affected by LBBB. Notice that, without loss of generality, we are considering in order to ease the computations two populations of data with the same number of realizations. Figures 1 and 2 show the ECG signals selected in the datasets $\mathbf{X}$ and $\mathbf{Y}$, respectively. The first step of our analysis is to calculate the Spearman matrices for physiological ($\widehat{SM}(\mathbf{X})$) and pathological ($\widehat{SM}(\mathbf{Y})$) ECGs, respectively. The entries coloured in yellow represent those for which there isn't statistical evidence of being different from zero indicating that the corresponding pairs of leads can be assumed independent. Their detection is performed observing the confidence intervals computed using 1000 bootstrap iterations: if an interval contains zero, the hypothesis of independence between the corresponding pair of leads is not rejected and so the component of the Spearman Matrix is highlighted to indicate a non significant dependence. The two matrices provide an effective insight on the way in which the leads of the ECG signals depend on each other and give us the possibility to compare the pattern of




DRAFT

Figure 1: Registered and denoised ECG signals of the $n_x = 200$ physiological patients used for the analysis.




DRAFT

Figure 2: Registered and denoised ECG signals of the $n_y = 200$ LBBB patients used for the analysis.

dependence in the two populations. We can notice a similarity in the upper diagonals of the matrices, a part from one case, with high and significantly different from zero entries. We also notice remarkable differences: the pattern of dependence of physiological signals is more connected, whereas the one of LBBBs is more sparse, due to the presence of several pairs of independent leads. Moreover, it seems that V2 is particularly affected by the disease and in the case of physiological signals, the entries that are significatively different from zero are positive, indicating that there is agreement between the grades of the leads, while the same does not happen for the LBBB signals where we notice some negative entries. What we observe can be interpreted in terms of heart dynamics: in physiological patients, the heart dynamics is more regular and expresses coordinated behaviours in all the components of the ECGs, whereas it becomes more chaotic and characterized by independent behaviours when the pathology is present. Hence, it seems that the disease is able to change the natural relation of dependence among some leads of the ECG. In order to have a statistical evidence of our statement we introduce a test of hypothesis verifying the equality between the two matrices. We want to perform the test:

$$H_0 : SM(\mathbf{X}_t) = SM(\mathbf{Y}_t) \quad vs$$
$$H_1 : SM(\mathbf{X}_t) \neq SM(\mathbf{Y}_t).$$

with a non parametric approach presenting a permutation-based testing procedure that it does not require any distributional assumption on data that may restrict its application. This test is called $K$-sample Anderson-Darling test ($K \geq 2$) and the test statistic associated is based on the notion of the generalized cosine measure between two symmetric matrices (8). In our application we have $K = 2$ and the test statistic for the equality of the two Spearman matrices is $1 - cos(\mathbf{M}_1, \mathbf{M}_2)$, where the cosine is computed according to

$$cos(\mathbf{M}_1, \mathbf{M}_2) = \frac{vech^*(\mathbf{M}_1)^T vech^*(\mathbf{M}_2)}{\|vech^*(\mathbf{M}_1)\| \|vech^*(\mathbf{M}_2)\|}.$$

The modified half-vectorization operator $vech^*(\cdot)$ completely remove the redundancy in symmetric matrices and are very easy to compute. This newly proposed test statistic exploits a simple geometric property of symmetric matrices and require no distributional assumptions of data. We further adopt a permutation approach to obtain the distribution under the null hypothesis of the test statistic and compute p-values of the test. Applying the two-sample Anderson-Darling test to the case of the Spearman matrices of physiological and LBBB patients and computing 1000 permutational replication of our test statistic T under $H_0$ we observe that the observed value of $T_0$ is not likely under the null hypothesis (the p-value of the test is close to 0). Hence, the test gives strong evidence to reject $H_0$ and to state that the Spearman matrices of physiological and pathological signals are statistically different.

DRAFT

## 3. Conclusion

In this work, we presented the notion of Spearman index in the infinite dimensional framework to quantify the dependence among two families of functional data. Starting from this definition, we built the Spearman Matrix, a new mathematical object that mimics the covariance matrix of multivariate statistics and that provides an effective insight of the pattern of dependence among the components of a multivariate functional dataset. In the applicative part of our work, we moved to the analysis of a real dataset. We compared the Spearman Matrix arising from the 8-variate electrocardiographic signals of a population of healthy people with the one arising from signals of people affected by Left Bundle Brunch Block (LBBB). The aim was to verify if the pattern of dependence in the two cases are different due to the presence of the disease. We tested the hypothesis that physiological and pathological signals have the same Spearman Matrix, adapting to our framework a non-parametric test which check the equality of Spearman correlation matrices arising from different populations of multivariate data.

## References

[1] Boroskikh, Yu.: U-statistics in Banach space. VSP BV, Oud-Beijerland (1996)

[2] Ieva, F., Paganoni, A.M., Pigoli, D. and Vitelli, V.: Multivariate functional clustering for the analysis of ECG curves morphology. Journal of the Royal Statistical Society, Series C. **62**, 401–418 (2013)
[3] Ieva, F., Paganoni, A.M.: Risk Prediction for Myocardial Infarction via Generalized Functional Regression Models. Statistical Methods in Medical Research. **25**, 1648–1660 (2013)
[4] Ieva, F., Paganoni, A.M.: Component-wise outlier detection methods for robustifying multivariate functional samples. Statistical Papers. **61**, 595–614 (2020)
[5] Lopez-Pintado, S. and Romo, J.: A half-region depth for functional data. Computational Statistics and Data Analysis. **55**, 1679–1695 (2011)
[6] Ronzulli, M.: A Non-Parametric-based Inferential Framework for Multivariate Functional Data: an application to ECG signals. Politecnico di Milano, School of Industrial and Information Engineering, Master of Science in Mathematical Engineering, Department of Mathematics, Italy (2022)
[7] Valencia, D., Lillo, R. and Romo, J.: A Kendall correlation coefficient between functional data. Advances in Data Analysis and Classification. **13**, 1083–1103 (2019)
[8] Wu, L., Weng, C., Wang, X., Wang, K. and Liu, X.: Test of Covariance and Correlation Matrices. https://arxiv.org/abs/1812.01172, (2018)

DRAFT

# Robust Statistical Process Monitoring of Multivariate Functional Data

Christian Capezza[a], Fabio Centofanti[a], Antonio Lepore[a], and Biagio Palumbo[a]

[a]Department of Industrial Engineering, University of Naples Federico II, Italy;
christian.capezza@unina.it, fabio.centofanti@unina.it,
antonio.lepore@unina.it, biagio.palumbo@unina.it

**Abstract**

Modern data sets are often contaminated with anomalous observations that can seriously decrease the performance of control charting procedures, especially in complex and high dimensional settings. To mitigate this issue, this paper presents a new control chart for multivariate functional data that is robust to functional casewise and cellwise outliers.

***Keywords:*** Statistical Process Monitoring, Profile Monitoring, Casewise and Cellwise Outliers

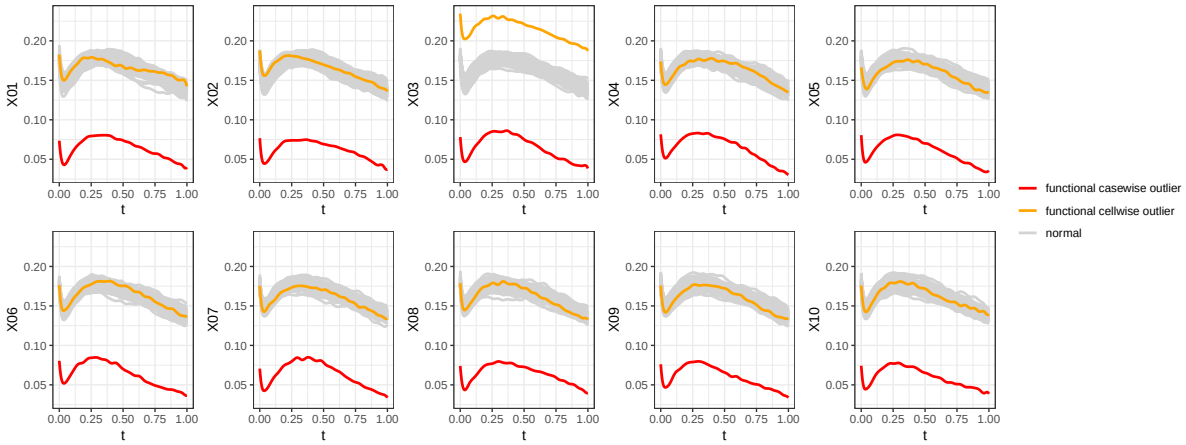
DRAFT

## 1. Introduction

In modern industrial processes, data acquisition systems allow the collection of massive amounts of high-frequency data. In statistical process monitoring (SPM) applications, the experimental measurements of the quality characteristics, i.e., the features that can be used to assess the quality of the process, are often characterized by complex and high dimensional formats that can be well represented as functional data, also referred to as profiles (17; 12). This justifies the growing interest in *profile monitoring* (16), which is the monitoring of a process through quality characteristic in the form of one or multiple profiles. Control charts are known as the main tools for SPM and are commonly implemented in two phases. Phase I is concerned with the identification of a clean data set to be assumed as representative of the in-control state of the process. This data set is then used to quantify the expected variation of a future observation to be used for the prospective process monitoring in Phase II. However, the identification of Phase I observations in high-dimensional contexts is not an easy task due to the presence of several outliers in one or more components of the quality characteristic observations. Indeed, control charts are very sensitive to the presence of outlying observations in the Phase I sample that can lead to inflated control limits and reduced power to detect process changes in Phase II.

Several robust approaches, i.e., able to deal with outliers, have been proposed in the literature for the multivariate SPM of scalar quality characteristics (2; 4). However, to the best of the authors' knowledge, a robust approach able to successfully capture the functional nature of a multivariate functional quality characteristic has not been devised so far. Traditional multivariate robust estimators assume a contamination model with a mixture distribution where a large fraction $(1-\varepsilon)$ of the data is generated from a distribution that is free of contamination, while a small fraction $\varepsilon$ of the data, denoted as *casewise* outliers, comes from an unspecified outlier generating distribution that may affect all the variables. In presence of many variables, a more plausible contamination model considers outlying observations where only a small fraction of the functional variables may be contaminated, hereinafter denoted as functional *cellwise* outliers. An toy example that graphically highlights the difference between casewise and

Figure 1: Example of functional casewise and cellwise outliers.




cellwise outliers is shown in Figure 1, where 100 simulated observations from a vector of 10 functional variables are plotted. Note that big shifts have been applied to clearly show the two types of outliers. The uncontaminated observations are plotted in gray, then a casewise functional outlier is added as a vector of ten red curves, while a cellwise functional outlier is plotted in orange. It is evident that functional casewise outliers come from a different distribution that affects all functional variables, which has been obtained by applying the same mean shift to the ten profiles, and may be relatively easy to identify with appropriate outlier detection methods. Whereas, in the functional cellwise outlying observation, only the third functional variable is independently contaminated, while the remaining nine ones seem to come from the same distribution as the uncontaminated data. Moreover, instead of discarding the entire observation because of only one anomalous variable, the information from the other nine components could still be useful to better estimate the distribution of the data. The main problem with cellwise outliers is that, if contamination is assumed to independently arise in each functional variable, then the fraction of perfectly observed cases can be rather small, therefore traditional multivariate robust estimators are known to fail and to be affected by the outlier propagation problem (3). Nonetheless, as pointed out by Agostinelli et al. (1), both types of data contamination, casewise and cellwise, may occur together.

In the face of these problems, we propose a new framework, referred to as robust multivariate functional control chart (RoMFCC), for the SPM of multivariate functional data in presence of both casewise and cellwise outliers. By means of a Monte Carlo simulation study, the ability of the RoMFCC in identifying mean shifts in the functional variables in presence of casewise and cellwise outliers is compared with other control charts already appeared in the literature. Finally, the proposed framework is demonstrated through a real-case study.

DRAFT

## 2. The Robust Multivariate Functional Control Chart Framework

The RoMFCC is proposed as a new general framework for the SPM of multivariate functional data and relies on the following four main elements. For additional methodological details, the reader is referred to (6).

(I) The functional univariate filter, referred to as FUF, identifies functional cellwise outliers and is an extension of the filtering proposed in (1). Specifically, the proposed FUF replaces with a missing component each component observation of the multivariate functional quality characteristic that achieves an overly large functional distance measure from the center of the data, which is specifically chosen to be robust to the presence of outliers. The corresponding multivariate functional quality characteristic observations (i.e., those achieving a large functional distance in at least one component) are then denoted as a functional cellwise outlier.

(II) The robust functional data imputation (RoFDI) method extends the robust imputation approach (4) to the functional setting. Summarily, the RoFDI method replaces the missing components of each functional cellwise outlier by sequentially imputing the observation that achieves the minimum distance

from the space generated by the complete realization, obtained through the RoMFPCA."

(III) Robust multivariate functional principal component analysis (RoMFPCA) is then used in place of the well-known standard multivariate functional principal component analysis (14) to reduce the infinite dimensionality of the multivariate functional data while taking into account the casewise outliers. The RoMFPCA extends the approach (18) to multivariate functional data by applying the ROBPCA to a specific transformation of the coefficients obtained through the basis function expansion method (17).

(IV) The consolidated monitoring strategy is performed on a multivariate functional quality characteristic based on the Hotelling's $T^2$ and *SPE* control charts (16; 11; 8; 7).

The proposed RoMFCC framework collects the elements I-IV in a Phase II monitoring strategy where the Phase I sample can be contaminated with functional casewise and cellwise outliers.

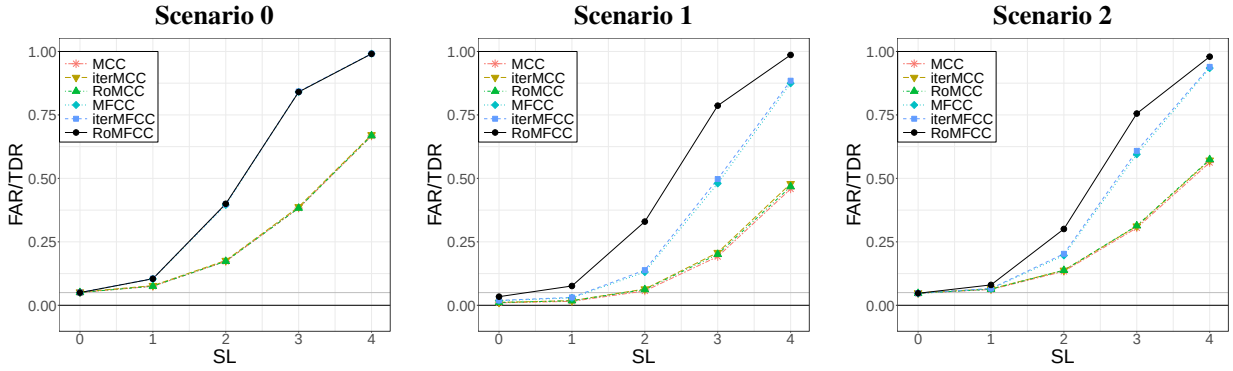
## 3. Simulation Study

The performance of the RoMFCC in identifying mean shifts of the multivariate functional quality characteristic is assessed through a Monte Carlo simulation of two scenarios with different Phase I sample contamination and data generated to mimic typical behaviours of the dynamic resistance curves (DRCs) shown in the real-case study of Section 4. Specifically, in Scenario 1 and Scenario 2, the Phase I sample is contaminated by functional cellwise and casewise outliers, respectively. For each scenario, the contamination model mimics typical anomalous DRCs observed in presence of splash welds, caused by excessive welding current (19). In addition, Scenario 0 simulates a Phase I sample that is not contaminated by any type of outliers. The Phase II sample is generated analogously to the contamination model at 4 different severity levels $SL = \{1,2,3,4\}$.

The proposed RoMFCC framework is compared with several natural competing approaches grouped as follows into control charts for multivariate non-functional and functional data, respectively. Specifically, the first group consists of control charts for multivariate data built on the vectors of time average (scalar) values of each component of multivariate functional data observations. In this group, we consider (*i*) the classical *multivariate* classical Hotelling's $T^2$ control chart, referred to as MCC; (*ii*) its *iterative* variant, referred to as iterMCC, where outliers detected by the control chart in Phase I are iteratively removed and control limits are revised until all data are assumed to be in control (IC); (*iii*) the *multivariate robust* control chart proposed in (9), referred to as RoMCC. In the second group, we consider two approaches recently appeared in the profile monitoring literature, namely (*iv*) the *multivariate functional control charts* and referred to as MFCC, proposed in (7); (*v*) its *iterative* variant, referred to as iterMFCC, where, as before, outliers detected in Phase I are iteratively removed until all data are assumed to be IC. For each scenario and severity level, 50 simulation runs are performed. The RoMFCC and the competing methods performance is assessed by means of the true detection rate (TDR) and the false alarm rate (FAR), which are defined as the proportion of points that fall outside the control limits whilst the process is, respectively out of control (OC) or IC. Figure 2 displays the mean FAR ($SL = 0$) and TDR ($SL \neq 0$) as a function of the severity level *SL* for Scenario 0, Scenario 1 and Scenario 2. When the Phase I sample is not contaminated by outliers (Scenario 0) Figure 2 shows that all the approaches able to account for the functional nature of the data (namely, MFCC, iterMFCC, RoMFCC) achieve the same performance. Although this scenario should be not favourable for robust approaches, the iterMFCC and RoMFCC perform equal to the MFCC. The non-functional approaches (namely MCC, iterMCC, RoMCC) show worse performance than the functional counterparts, and there is no significant performance difference among them as well. In Scenario 1, where the Phase I sample is contaminated by cellwise outliers, the proposed RoMFCC largely outperforms the competing methods. The iterMFCC, which is representative of the baseline method in this setting, does not greatly improve the MFCC performance. This is probably due to the masking effect that prevents the iterMFCC from iteratively identifying functional cellwise outliers in the Phase I sample. The performance of the non-functional methods is very unsatisfactory because they are not able to both capture the functional nature of the data and successfully deal with outliers. Overall, the MCC is the worst method, readily followed by iterMCC and RoMCC. Scenario 2 shows similar results as in Scenario 1, where the non-robust approaches perform slightly better than in Scenario 1 because cellwise contamination is more difficult to manage with respect to casewise outliers.

DRAFT

Figure 2: Mean FAR ($SL = 0$) or TDR ($SL \neq 0$) achieved by MCC, iterMCC, RoMCC, MFCC, iterMFCC and RoMFCC as a function of the severity level $SL$ in Scenario 0, Scenario 1 and Scenario 2.




## 4. Real-Case Study

To demonstrate the potential of the proposed RoMFCC in practical situations, a real-case study in the automotive industry is presented hereafter. It addresses the issue of monitoring the quality of the resistance spot-welding process, which guarantees the structural integrity and solidity of welded assemblies in each vehicle (15). The modern automotive Industry 4.0 framework allows the automatic acquisition of a large volume of RSW process parameters, and, in particular, of the DRC, which is considered the most informative proxy of the RSW process quality and a low-cost relative to the alternative destructive tests. Further details on how the typical behaviour of a DRC is related to the physical and metallurgical development of a spot weld are provided in (5).

Data analyzed in this study are courtesy of Centro Ricerche Fiat (Italy) and are recorded at the Mirafiori Factory during lab tests on the body of the Fiat 500BEV. An automobile body-in-white stage is in general characterized by a large number of spot welds with different characteristics, e.g, the thickness and material of the sheets to be joined together and the welding time. To be specific, in this real-case study, we focus on the monitoring of spot welds made by only one of the welding machines. In particular, we consider the multivariate functional quality characteristic represented by the vector of ten DRCs corresponding to the same ten spot weld locations on each sub-assembly normalized on the time domain $[0, 1]$. The data set contains a total number of 1839 sub-assemblies.

The RSW process quality is directly affected by electrode wear that leads to changes in electrical, thermal and mechanical contact conditions at electrode and metal sheet interfaces (13). Thus, to take into account the wear issue, electrodes go through periodical renovations. In this setting, a paramount issue refers to the swift identification of DRCs mean shifts caused by electrode wear, which could be considered as a criterion for electrode life termination and guide the electrode renovation strategy.

In the light of this, the 919 multivariate profiles corresponding to spot welds made immediately before electrode renewal are used to form the Phase I sample, whereas, the remaining 920 observations are used in Phase II to evaluate the in-line monitoring performance in detecting the mean shift of the Phase II DRCs caused by electrode wear. The RoMFCC is implemented as in Section 3. In Phase II, the RoMFCC signals 72.3% of the observations as OC, which reflects the proposed method performance in tracking mean shifts caused by large electrode wear. The proposed method is compared with the competing methods presented in the simulation study of Section 3. through the estimated TDR, denoted as $\widehat{TDR}$, on the Phase II sample, as shown in Table 1. Similarly to (8), the uncertainty of $\widehat{TDR}$ is quantified through a bootstrap analysis (10). Table 1 reports the mean of the empirical bootstrap distribution of $\widehat{TDR}$, denoted by $\overline{TDR}$, and the corresponding bootstrap 95% confidence interval (CI) for each monitoring method. Bootstrap 95% confidence intervals achieved by the RoMFCC are strictly above those of all considered monitoring approaches. In general, as in Section 3, the considered non-functional approaches (namely, MCC, iterMCC, RoMCC) show worse performance than the functional counterparts because they are not able to satisfactorily capture the functional nature of the data and robust approaches always improve the non-robust ones on this study. Therefore, the proposed RoMFCC stands out as the best

DRAFT

| | $\widehat{TDR}$ | $\overline{TDR}$ | CI |
|---|---|---|---|
| MCC | 0.336 | 0.335 | [0.305,0.368] |
| iterMCC | 0.462 | 0.461 | [0.428,0.496] |
| RoMCC | 0.513 | 0.512 | [0.481,0.547] |
| MFCC | 0.541 | 0.541 | [0.511,0.574] |
| iterMFCC | 0.632 | 0.632 | [0.595,0.664] |
| RoMFCC | 0.723 | 0.723 | [0.695,0.753] |

Table 1: Estimated TDR values $\widehat{TDR}$ on the Phase II sample, mean $\overline{TDR}$ of the empirical bootstrap distribution of $\widehat{TDR}$, and the corresponding bootstrap 95% confidence interval (CI) for each monitoring method.

method to promptly identify OC conditions in the considered RSW process characterized by a Phase I sample contaminated by functional outliers.

## 5. Conclusions

In this paper, we propose the RoMFCC framework for the statistical process monitoring of multivariate functional data that is robust to functional casewise and cellwise outliers. The proposed framework is suitable to monitor industrial processes as the RSW one that motivated this research where multivariate functional quality characteristics such as DRCs are available and occasionally contaminated by outliers.

The performance of the RoMFCC framework is assessed through a Monte Carlo simulation study where it is compared with several competing monitoring methods. The ability of the proposed method to estimate the distribution of the data without removing observations while being robust to functional casewise and cellwise outliers allows the RoMFCC to outperform the competitors in all the considered scenarios and to be the only alternative in high-dimensional scenarios where the most competing methods may even fail.

DRAFT

## Acknowledgments

This work has been done in the framework of the R&D project of the multiregional investment programme "REINForce: REsearch to INspire the Future" (CDS000609) with Hitachi Rail STS, supported by the Italian Ministry for Economic Development (MISE) through the Invitalia agency.

## References

[1] Agostinelli, C., Leung, A., Yohai, V.J., Zamar, R.H.: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test **24**(3), 441–461 (2015)

[2] Alfaro, J., Ortega, J.F.: A comparison of robust alternatives to Hotelling's $T^2$ control chart. Journal of Applied Statistics **36**(12), 1385–1396 (2009)

[3] Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H.: Propagation of outliers in multivariate data. The Annals of Statistics pp. 311–331 (2009)

[4] Cabana, E., Lillo, R.E.: Robust multivariate control chart based on shrinkage for individual observations. Journal of Quality Technology pp. 1–26 (2021)

[5] Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: Functional clustering methods for resistance spot welding process data in the automotive industry. Applied Stochastic Models in Business and Industry **37**(5), 908–925 (2021)

[6] Capezza, C., Centofanti, F., Lepore, A., Palumbo, B.: Robust multivariate functional control charts. arXiv preprint arXiv:2207.07978 (2022)
[7] Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Control charts for monitoring ship operating conditions and $CO_2$ emissions based on scalar-on-function regression. Applied Stochastic Models in Business and Industry **36**(3), 477–500 (2020)
[8] Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S.: Functional regression control chart. Technometrics **63**(3), 281–294 (2021)
[9] Chenouri, S., Steiner, S.H., Variyath, A.M.: A multivariate robust control chart for individual observations. Journal of Quality Technology **41**(3), 259–271 (2009)
[10] Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
[11] Grasso, M., Menafoglio, A., Colosimo, B.M., Secchi, P.: Using curve-registration information for profile monitoring. Journal of Quality Technology **48**(2), 99–127 (2016)
[12] Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. Chapman and Hall/CRC (2017)
[13] Manladan, S., Yusof, F., Ramesh, S., Fadzil, M., Luo, Z., Ao, S.: A review on resistance spot welding of aluminum alloys. The International Journal of Advanced Manufacturing Technology **90**(1), 605–634 (2017)
[14] Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust statistics: theory and methods (with R). John Wiley & Sons (2019)
[15] Martín, Ó., Pereda, M., Santos, J.I., Galán, J.M.: Assessment of resistance spot welding quality based on ultrasonic testing and tree-based techniques. Journal of Materials Processing Technology **214**(11), 2478–2487 (2014)
[16] Noorossana, R., Saghaei, A., Amiri, A.: Statistical analysis of profile monitoring, vol. 865. John Wiley & Sons (2011)
[17] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer (2005). DOI 10.1007/b98888. URL https://doi.org/10.1007/b98888
[18] Sawant, P., Billor, N., Shin, H.: Functional outlier detection with robust functional principal component analysis. Computational Statistics **27**(1), 83–102 (2012)
[19] Xia, Y.J., Su, Z.W., Li, Y.B., Zhou, L., Shen, Y.: Online quantitative evaluation of expulsion in resistance spot welding. Journal of Manufacturing Processes **46**, 34–43 (2019)

DRAFT

# The effects of mobility restrictions on public health: a functional data analysis for Italy over the years 2020 and 2021

Veronica Mazzola[a], Giovanni Bonaccorsi[b], Piercesare Secchi[a], and Francesca Ieva[a,c]

[a]MOX lab, Department of Mathematics, Politecnico di Milano;
veronica.mazzola@mail.polimi.it, piercesare.secchi@polimi.it, francesca.ieva@polimi.it
[b]Department of Management, Economics and Industrial Engineering, Politecnico di Milano;
giovanni.bonaccorsi@polimi.it
[c]Health Data Science Center, Human Technopole

DRAFT

**Abstract**

In response to the COVID-19 pandemic, Italy adopted restrictive mobility measures to contain the spread of the coronavirus. In this study we develop novel strategies for assessing the relationship between mobility and overall mortality. In particular, we first identify diversified behaviors over different geographical areas in terms of mobility, then evaluate the relationship between restrictions and mortality through a functional approach combined with a Spearman correlation analysis. This allows for a stratified assessment, from a precision policy perspective. Specifically, we analysed mobility data provided by Facebook between Italian provinces from March 2020 to December 2021. Measuring the Spearman's correlation coefficient between epidemiological and mobility curves, we developed two criteria for quantifying the effect of policies restrictions stratified over different areas.

***Keywords:*** COVID-19, mobility networks, functional data analysis, Spearman index

## 1. Introduction

The year 2020 will be remembered worldwide for the sanitary emergency of COVID-19 pandemic. Despite the numerous interventions adopted by the country to cope with the situation and limit infections (8), the virus spread fast. Two cases of coronavirus were identified in Rome at the end of January, but only on February 21 the first case of SARS-CoV-2 infection was diagnosed in Italy. Soon, the epidemic spread across Northern regions, as well as in other states. On March 9, Italy was the first European country to apply a national lockdown. As of May 1, 2020, the Italian health authorities reported 28.238 deaths nationally (14). The Italian Government responded with Non-Pharmaceutical Interventions (NPI), such as school and university closures, social distancing and full lockdown, aimed at reducing the mobility of citizens to decrease the rate of contagion. Among NPIs, mobility restrictions resulted to play a central role in decreasing social contacts (4). In particular, travel restrictions proved to be particularly useful in the early stage of the outbreak, when contagion was confined to certain areas (8), but they were less effective once the virus became more widespread during the so-called second wave of infections. In October 2020, new restrictive measures were adopted to contain the emergency COVID-19. On November

6, 2020, the Italian government adopted new restrictive measures to contain the COVID-19 emergency, dividing the country into three colored areas according to the severity of contagion at the local level (7). The countrywide curfew, introduced during this period, was abolished only in June 2021.

In order to characterize the effect of mobility restrictions, we analyzed a large-scale collection of near real time data provided by the Facebook platform (6). We first identified clusters of areas with similar mobility density behaviors, applying a functional K-means informed by a Wasserstein distance. We then spotted areas with peculiar mortality density behavior with respect to the surroundings with LISA maps, exploiting the Local Moran's I statistics whose weights account for spatial correlations. Then the Spearman correlation of mortality and mobility data is computed to assess the lag of significant association between the two phenomena.

## 2. Data

We analyzed mobility between municipalities based on *Disease Prevention Maps* provided by Facebook through its *Data for Good* program[1]. Measurements are collected with an 8-h frequency by aggregating individual trips of Facebook users that have agreed to share their location. Mobility flows are constructed to ensure privacy and anonymity (4), and are related to movements between Italian municipalities from March 1, 2020 to December 31, 2021.

Such data may be represented as weighted networks (2), with provinces acting as nodes. For this reason, the strength node centrality may be used as an indicator of mobility. Indeed, we decided to consider as an indicator of mobility for each province the weekly strength node centrality. Thus, for each province, we calculated 96 strength values, one for each week within the period March 2020 - December 2021, meaning by strength not only the sum of people who have left or entered a given province, but also considering those who have moved internally within the province. Smoothing such curves through natural cubic splines, we end up with functional data representing the mobility of all the Italian provinces over the period of interest.

DRAFT

Figure 1: Total flow of the Italy mobility network. Restriction periods related to the first (in red) and second (in orange) epidemic waves are highlighted.

In Fig. 1 the weekly total flow of the Italy mobility network in the period March 2020 - December 2021 is reported, obtained by considering each province as a node. A pronounced decrease in mobility corresponding to lockdown restrictions can be noticed, confirming results already found in the literature (3) (4), and a weaker but prolonged effect of the second wave on mobility in late 2020 and early 2021.

In order to correctly interpret this functional data, we normalized them as in (12), therefore not studying directly the mobility strengths, but rather the corresponding mobility densities.

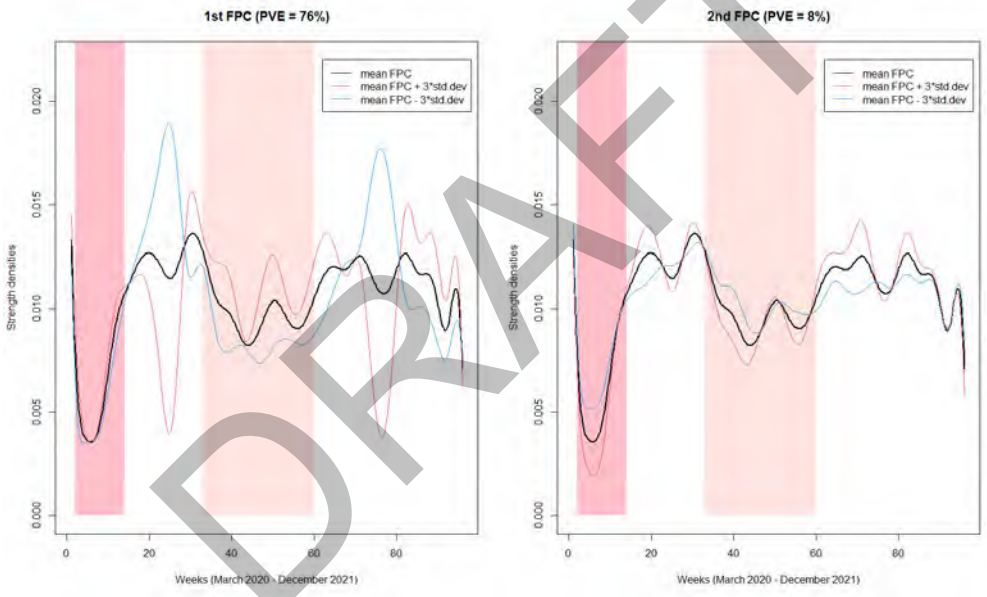
[1]For what concerns Facebook human mobility, all data are provided under an academic license agreement with Facebook through its *Data for Good* program. Facebook releases data upon request to nonprofit organization and academics.

Regarding the epidemiological data, we analyzed the daily number of deaths for all causes in each municipality provided by Istat (9), aggregating them both at the temporal level (on a weekly basis) and at the spatial level (by province). Following the same normalization procedure performed on mobility data, we studied the provincial mortality densities. We considered the same time interval analyzed for mobility, i.e., from March 2020 to December 2021.

## 3. Methods and Results

The variability structure of the mobility densities was explored by decomposing them in their main modes of variability through the Functional Principal Component Analysis (FPCA) (11).

In Fig. 2 the first and second functional principal components of mobility density functions retrieved by mobility networks are reported. We observe that Italian provinces may be characterized mainly in terms of the concentration of mobility they experienced in the summer periods and, secondly, for the mobility they recorded in periods of lockdown compared to that in the so-called recovery periods. Indeed, the first FPC represents a contrast between provinces that, with respect to the mean, had a higher or lower concentration of mobility in the summer, while the second principal component captures the concentration of mobility during lockdown, restrictions and recovery periods.




DRAFT

Figure 2: First and second functional principal components of mobility density functions retrieved by mobility networks. The mean of the mobility densities $d$ (black curve) is perturbed in the direction of the principal components. Specifically, the blue curve is $d - 3\sqrt{\lambda_j}\phi_j$ and the red curve is $d + 3\sqrt{\lambda_j}\phi_j$ for $j = 1, 2$, with $\lambda_j$ being the eigenvalue associated to the j-th FPC and $\phi_j$ the eigenfunction.

To keep also spatial dependency into account, spotting provinces with very high (low) scores on the two first FPCs which are also surrounded by provinces characterized by the same kind of mobility, we computed the Local Moran's I statistic (1) for the FPC scores. Figure 3 shows the LISA maps obtained with this procedure, which enables the identification of provinces acting as hot-spot with respect to the categorization induced by the FPCs.

To better investigate (spatial) patterns in the mobility density functions through a single clustering procedure, we then decided to cluster the mobility density functions using a Wasserstein distance based functional K-means. This approach returned the provincial mobility clusters reported in Fig. 4.

It is possible to characterize them as follows: Cluster 1 (blue) contains provinces strongly affected by the pandemic in both waves, and especially during the first one. Cluster 2 (green) contains provinces

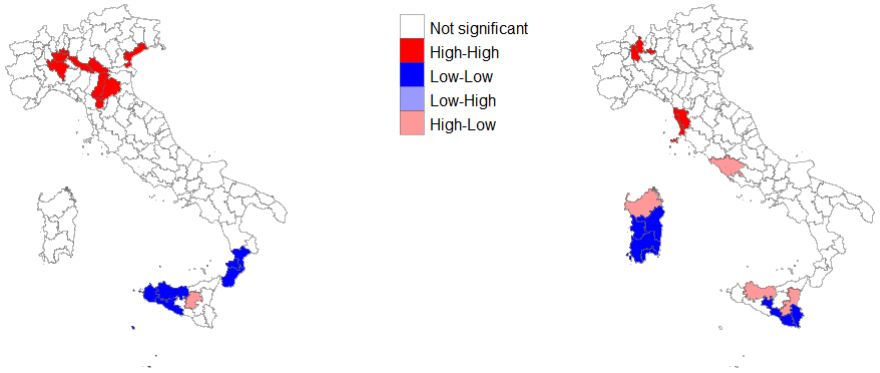



Figure 3: LISA maps on the scores of the first (left) and of the second (right) FPC according to the Local Moran's I statistic.

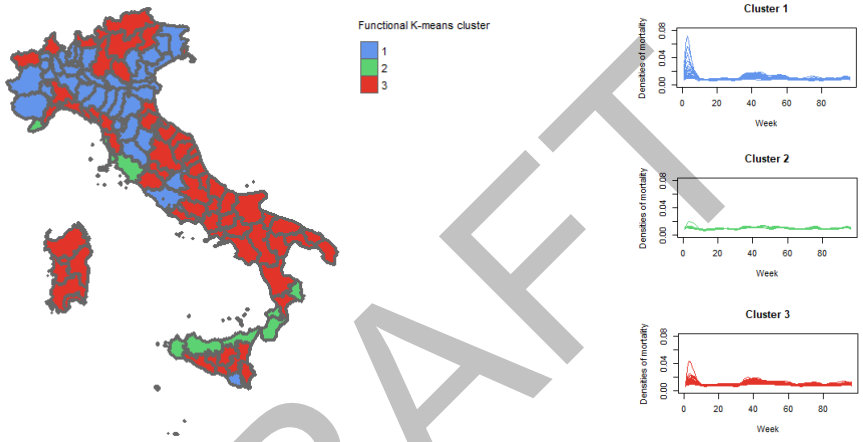



Figure 4: Clusters of provinces (left panel) deriving from the application of a Wasserstein distance based functional K-means to mobility densities, and corresponding mortality densities (right panel).

DRAFT

not deeply affected by the pandemic. Finally, cluster 3 (red) contains provinces affected by the pandemic in both waves, but mainly in the second.

The curves belonging to the groups defined by the Wasserstein-based functional K-means have been then related with the epidemiological curves described in Sect. 2. through a quantitative procedure. Specifically, we calculated the correlation between epidemiological and mobility curves through the Spearman's index (13). Notice that, in doing so, despite the groups being defined using the entire time domain, it is possible to set two different criteria, to be applied to time domain subset of interest, for determining the time required by mobility restriction measures to mitigate the effects of the epidemic. The first criterion relies on the point estimates of the Spearman's index: the mitigation of the epidemic induced by mobility measures ends when the correlation between mobility at the beginning of the restrictions and mortality in the following weeks changes in sign. The second concerns the p-value associated with the Spearman's estimates: when it indicates that correlation between mortality and mobility densities is no longer significant, the mobility restriction measures are no longer necessary.

Notice that we decided to take as a reference the mobility densities recorded during the first week in which the restrictive measures were activated and to calculate the Spearman's correlation between these set values and the mortality densities of that week and the following ones.

In Table 1 the Spearman's Index between mobility and mortality densities during the first and second waves, along with the corresponding p-values, are reported.

The greatest (negative) correlation is at lag=0, i.e., during the first week of restrictions, then it grad-

Table 1: Spearman's correlations (and associated p-values) between the mobility densities in the first week of restriction of the first wave and second wave, and the mortality densities of the corresponding and following weeks, for the Italian provinces belonging to cluster 1.

| Lag | First wave | Second wave |
|---|---|---|
| 0 | -0.64 ($< 0.0001$) | -0.47 ($< 0.01$) |
| 1 | -0.70 ($< 0.0001$) | -0.48 ($< 0.01$) |
| 2 | -0.69 ($< 0.0001$) | -0.45 ($< 0.01$) |
| 3 | -0.70 ($< 0.0001$) | -0.35 ($< 0.05$) |
| 4 | -0.67 ($< 0.0001$) | -0.21 ($0.22$) |
| 5 | -0.59 ($< 0.0001$) | -0.06 ($0.73$) |
| 6 | -0.37 ($< 0.05$) | 0.08 ($0.63$) |
| 7 | -0.30 ($0.08$) | - |
| 8 | -0.25 ($0.13$) | - |
| 9 | -0.14 ($0.40$) | - |
| 10 | 0.009 (0.09) | - |

ually rises to zero. This means that when the first lockdown was imposed and mobility suffered a sharp reduction, deaths within Italian provinces were still numerous. Then the correlation decreases in absolute value, becoming almost null between the sixth and the seventh week after the beginning of restrictions. This may be intended as a suggestion for the mobility measures to become effective in about 6 weeks in reducing the mortality densities. This result is coherent with those found at regional level in Boschi et al. (5), where a lag of about 30 days between mobility and mortality was discovered for the first wave of COVID-19.

Concerning the second wave, we obtained shorter time lags probably due to the fact that other precautionary measures, such as extensive testing, contact tracing and social distancing, were in action after the first phase of the pandemic (10; 14), weakening the relationship between epidemic and mobility.

## 4. Conclusions

This work was motivated by the interest in analysing mobility data and in providing decision makers with actionable tools for assessing mobility restriction policies. We exploited Facebook data to measure mobility in the country, with province granularity, over the weeks, and defined two criteria to determine when mobility restriction measures can be considered effective in containing the COVID-19 spread in terms of overall mortality. Both criteria are based on Spearman's correlation. Although correlation does not imply causation, we believe that these dynamic criteria can quantify the effects of mobility restrictions induced by the spread of COVID-19, in mitigating the epidemic.

In general, we found negative correlations between mortality and mobility densities, consistent with the fact that when high mortality concentrations were detected, the Italian Government adopted more restrictive measures and therefore allowed less in terms of mobility, and vice versa. This is particularly evident for the provinces most affected by the epidemic, especially concerning the first wave, and in cases where also the spatial component during clustering was considered. Moreover, the time lags between mobility reduction and actual results in limiting and controlling the epidemic were found to be longer in the provinces that suffered most from the COVID-19 pandemic in the period under consideration.

**Acknowledgments** The work is motivated as a part of the *CHANCE project*, specifically of the working group devoted to the evaluation of the effects of health strategies for the reduction of SARS-CoV-2 risk, with the aim of predicting the implications on possible future epidemics.

DRAFT

## References


[1] Anselin, L.: Local indicators of spatial association - LISA. Geographical Analysis. (1995) doi: 10.1111/j.1538-4632.1995.tb00338.x

[2] Barabási, A.L.: Network Science. Cambridge University Press. (2016)

[3] Bonaccorsi, G. et al.: Economic and social consequences of human mobility restrictions under COVID-19. PNAS. (2020) doi: 10.1073/pnas.2007658117

[4] Bonaccorsi, G., Pierri, F., Scotti, F., Flori, A., Manaresi, F., Ceri, S., Pammolli, F.: Socioeconomic differences and persistent segregation of Italian territories during COVID-19 pandemic. Scientific Reports. (2021)

[5] Boschi, T., Di Iorio, J., Testa, L., Cremona, M.A., Chiaromonte, F.: Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy. Nature. **11**, (2021) https://www.nature.com/articles/s41598-021-95866-y

[6] Meta. Data for Good program. Accessed March 2022 at https://dataforgood.facebook.com/dfg/tools/movement-maps

[7] Presidenza del Consiglio dei Ministri, le misure adottate dal Governo. Accessed March 2022

[8] Kraemer, M. U. G. et al.: The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. (2020) doi: 10.1126/science.abb4218

[9] Istat. Accessed March 2022 at https://www.istat.it/

[10] Nouvellet, P., Bhatia, S., Cori, A. et al.: Reduction in mobility and COVID-19 transmission. Nat Comm. **12**, (2021)

[11] Ramsay, J.O., Silverman, B. W.: Functional Data Analysis. Springer (2005)

[12] Scimone, R., Menafoglio, A., Sangalli, L.M., Secchi, P.: A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities. Spatial Statistics. (2021) doi: 10.1016/j.spasta.2021.100541

[13] Valencia, D., Lillo, R., Romo, J.:Spearman coefficient for functions. Universidad Carlos III de Madrid. (2013)

[14] Vollmer, Michaela A. C. et al.: Using mobility to estimate the trasmission intensity of COVID-19 in Italy: A subnational analysis with future scenarios. MedR$\chi$iv. (2020) doi: 10.1101/2020.05.05.20089359


DRAFT

# A vocabulary-based approach for risk detection in textual annotations of contracts of public procurement

Giulio Giacomo Cantone[a], Simone Del Sarto[b], and Michela Gnaldi[b]

[a]Business School, University of Sussex; prgcan@gmail.com
[b]Department of Political Science, University of Perugia; simone.delsarto@unipg.it, michela.gnaldi@unipg.it

DRAFT

**Abstract**

A text-mining approach is proposed to identify inconsistencies between the textual description of a public contract object and its pre-assigned classification using the Common Procurement Vocabulary (CPV). After the construction of suitable vocabularies for each macro CPV class (MCPV), two tests are proposed for (i.) detecting contracts whose object description contains terms/elements incoherent with their MCPV classification and (ii.) accordingly, red flagging them as at-risk contracts. The tests are applied to short textual annotations of the object of public contracts awarded within the Italian procurement system. It is demonstrated that in the adopted sample, one test performs much better than the other. This result is commented and explained. As a future development, it is discussed how vocabulary methods can be expanded for model-based classification. Si dimostra che nel campione adottato un test performa molto meglio dell'altro. Questo risultato è commentato e spiegato. Tra gli sviluppi futuri, si discute come i metodi basati sui vocabolari possono essere espansi per classificazioni dei contratti che siano basate su modelli statistici.

***Keywords:*** Text mining, Outlier detection, Public procurement, Short texts, TF-IDF

## 1. Introduction

This study aims at identifying inconsistencies between the textual description of contracts of public procurement and their classification through the Common Procurement Vocabulary (CPV), an international convention that codifies the object of the contract through a standard and internationally recognised ontology (5). The implicit assumption underlying the proposed procedure is that contract at-risk of corruption are associated with discrepancies between the textual elements enclosed in the object description and the contract CPV codification. The rationale of the proposed procedure is to provide tools to further support the process of corruption risk assessment in public procurement by informing policymakers and public authorities in charge of control activities. Moreover, as the best potentials of the proposed procedure are for prior screening of high volumes of contracts in public procurement, it can be conveniently employed to reduce the number of contracts to scrutinise in procedures of qualitative *audit*.

There are relatively few other attempts in the international literature to identify potentially risky circumstances in the public procurement process by the means of text mining (3; 4; 1; 2). This gap in the literature may be due to the semantic complexity of detecting outliers across relatively short textual strings, which brings computational burdens when applied to big data and high statistical uncertainty. Specifically, the identification of incoherent elements in the textual description of a contract object with

respect to its CPV codification implies content-specific knowledge and a sound understanding of text-mining techniques, whose application requires high calculation efforts due to sparse matrices.

Using text-mining techniques, this work proposes two basic tests for identifying inconsistencies between the object description of a public contract and its CPV classification. The procedures are tested on a sample of the Italian National Database of Public Contracts (BDNCP, *Banca Dati Nazionale dei Contratti Pubblici*), an open-access database about Italian public procurement. The proposed tests, when positive, identify contracts at risk of corruption, which are then red-flagged. Therefore, the procedure fits in that part of the scientific literature proposing red flag indicators for detecting corruption risk in public procurement (6; 9; 8).

The paper is organised as follows. Section 2, together with its subsections, describes the materials and methods employed for this work, while comments to preliminary results are reported in Section 3. Finally, concluding remarks are drawn in Section 4, by also highlighting limitations of the presented approach and proposing some future developments.

## 2. Materials and methods

This section is devoted to a description of the data at hand (Sect. 2.1), the proposed procedure for constructing a vocabulary (Sect. 2.2) and the tests for identifying inconsistencies between contract textual descriptions and contract objects, as identified by their CPV codes (Sect. 2.3).

DRAFT

### 2.1 Data

We consider data about the Italian public procurement process drawn from the BDNCP, a huge database handled by the Italian Anticorruption Authority, and containing detailed information of every public contract (around 61 million of contracts since 2007) managed by Italian contracting authorities (around 39 thousand).

Textual descriptions of the contract object are available in the BDNCP, as well as a CPV code assigned to any contract during the preparation of the call for tenders. The person in charge for the procurement phases within the contracting authority should select a CPV code, which should be as close as possible to the actual contract object.

We work on a sample drawn from the BDNCP, which contains 64,279 contracts issued from 2016 to 2022 by 9,811 contracting authorities. For this work, we consider the first two digits of the CPV, namely the CPV division (45 different codes), which corresponds to a macro-category of the contract object (MCPV in the following, standing for Macro CPV).

Contracts do not distribute uniformly across MCPV. MCPV 33 ("Medicine, drugs and personal care products") and MCPV 45 ("Construction works") are the most represented macro-categories of contract objects (18,241 and 17,405 contracts, respectively, corresponding to around 55% overall). The third most frequent is the MCPV 71 ("Architecture and engineering services"), with only $2,820$ contracts (4.4%), while the least frequent is the MCPV 76 ("Gas and oil services") with only 20 contracts. Moreover, 11 out of 45 MCPVs collect more than $1,000$ contracts, and 9 out of 45 include less than 100.

### 2.2 Costruction of a vocabulary

A *toy* vocabulary is constructed from the sample with software `tidytext` in R language (10). The command `unnest_tokens` converts the textual description of the contract object into a list of "tokens" (or terms), which consist of strings of lowercase alphanumerical symbols located between two blank spaces. In addition, using the SnowballC vocabularies dedicated to Italian language (11), tokens are reduced to their thematic roots ("stemming") and stopwords (e.g., articles, connectives, etc.) are removed.

The lists of all tokens (with associated MCPV) extracted from the sample are then recomposed into a dataframe and summarised as counts of the occurrence of each token in each MCPV. This is a tidy approach that falls under the methodology known as "Bag-of-words", since it does not account for the order of the words (7).

Let $f_{ik}$ be the (conditional) relative frequency of token $i$ given MCPV $k$, with $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$ ($K = 45$ in our case study). These frequencies are weighted through a weighting scheme analogue to the Term Frequency - Inverse Document Frequency metric, or TF-IDF (10; 7). Specifically, the weighted version of $f_{ik}$ is obtained as follows:

$$f_{ik}^{w} = f_{ik} \times \log_{10} \frac{K}{K_i}, \tag{1}$$

where $K_i = \mid \{k : i \in k\} \mid$ and represents the cardinality of the set including all the MCPVs that contain token $i$ at least once. As such, $f_{ik}^{w}$ is a measure of the relevance of the token $i$ for MCPV $k$, so the token with a higher value is more semantically salient for that MCPV. Of course, this value exists for each token that is counted at least once for that MCPV ($f_{ik} \neq 0$).

In order to extract a vocabulary $\mathcal{V}_k$ of relevant tokens for each MCPV, the use of a fixed number of tokens (i.e., the first ten with the greatest $f_{ik}^{w}$) is not advisable, since the number of contracts is unbalanced across MCPVs and the most frequent ones (i.e., MCPV 33 and MCPV 45 in our case) would have the same number of tokens as the least frequent. In order to let variate the cardinality of $\mathcal{V}_k$, we propose to consider a cut-off value between 0 and 1, denoted by $\vartheta$, and to build the vocabulary for each MCPV by including only those tokens whose cumulative weighted frequencies exceed this threshold, excluding the tokens that appear only once in that MCPV. In a formula, we have:

$$\mathcal{V}_k = \{i : F_{ik|c_{ik}>1}^{w} > \vartheta\}, \tag{2}$$

where $F_{ik|c_{ik}>1}^{w}$ is the cumulative sum of ordered $f_{ik}^{w}$ (in descending way), after considering the tokens with $c_{ik} > 1$ (more than one occurrence in MCPV $k$).

In this way, the number of tokens in each $\mathcal{V}_k$ depends on the frequencies of contracts in the MCPV. Furthermore, independently from $\vartheta$, in our case study the vocabulary of MCPV 33 and MCPV 45 will always contain a wide list of tokens, whereas those of the least frequent MCPVs (i.e., MCPV 76 or MCPV 16) will consist in very short lists.

## 2.3 Tests for risk detection

Two tests are proposed for identifying inconsistencies between contract textual descriptions and their MCPVs and, consequently, for red-flagging risky contracts. Given a generic contract $j$ assigned to a specific MCPV (say, $k^*$), let us denote its tokenised object by $\mathbf{x} = [x_1, \ldots, x_m]^\top$ (reference to contract $j$ is removed for ease of notation), which is the vector of the $m$ tokens extracted from the textual description of the contract object (after stemming and stopwords removal).

Test I is positive for the contract at issue (hence, a red flag is raised for that contract) if in vector $\mathbf{x}$ there is not even one token contained in the list of relevant tokens for the MCPV assigned to that contract ($k^*$), that is,

$$x_z \notin \mathcal{V}_{k^*}, \ \forall z = 1, \ldots, m. \tag{3}$$

Alternatively, Test II detects as contracts at risk those contracts whose object description contains at least one token not included in the list of relevant tokens for its MCPV, but contained in the list of relevant tokens of a different MCPV:

$$\exists \ z : x_z \notin \mathcal{V}_{k^*} \text{ and } x_z \in \mathcal{V}_{\bar{k}}, \bar{k} \neq k^*. \tag{4}$$

# 3. Preliminary results

The two proposed tests are applied to the data at hand. Figure 1 shows that Test II is very uninformative, independently of $\vartheta$, as it would label almost all contracts with a red flag. On the other hand, Test I is sensitive to the choice of $\vartheta$[1], although not monotonically. The marginal reduction of "positive" contracts by conjoining the two Tests is abysmal, especially for higher $\vartheta$.

[1] $\vartheta = 0.3$ has been elicited as the minimum value for the parameter because all relevant tokens for MCPV 16 would be filtered out under this value.

DRAFT

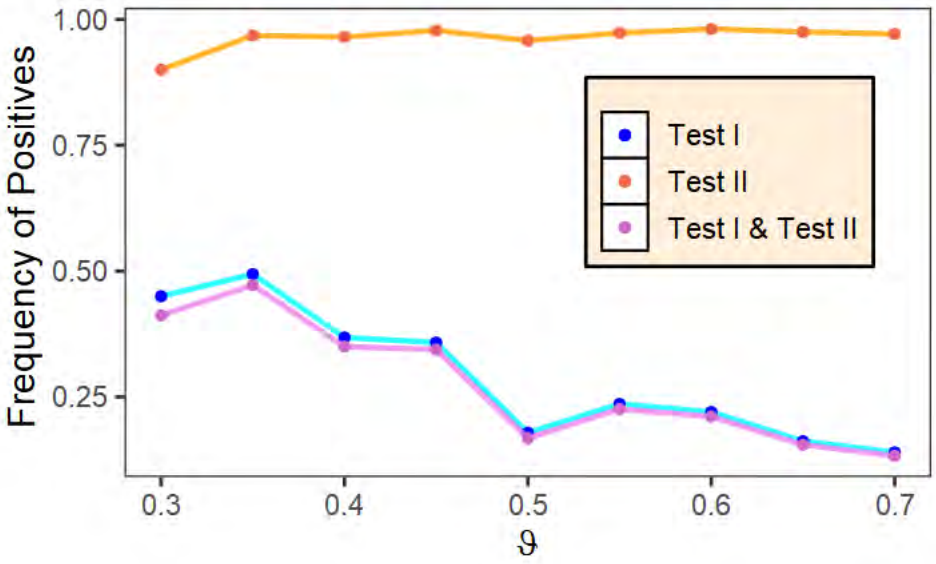



Figure 1: Relative frequency of contracts with a raised red flag according to the two proposed tests (or both).

DRAFT

These results are very dependent on the vocabulary, built on a relatively small sample. This affects $f_{ik}^{w}$ since Eq. 1 considers the MCPV where token $i$ occurs at least once, and the probability of a token $i$ to be counted at least once depends positively on the number of contracts. Moreover, in Eq. 2 tokens that are counted only once in the MCPV are excluded from the cumulative frequency; again, the probability to occur more than once depends positively on the number of contracts in the sample.

Furthermore, the reason for Test II being so uninformative is that the probability that vector $\mathbf{x}$ includes a token associated with a MCPV different from that assigned to the contract is higher when the vocabularies contains many terms. In fact, when $\vartheta$ is small, the frequency of red flags for Test II is slightly lower (although still high).

The parametric tuning of $\vartheta$ shows that Test II must be avoided in absence of a proper large vocabulary. On the other hand, Test I is a more solid pick and its output can be optimised through the parameterisation of $\vartheta$. Given the nature of the data at hand, it is expected that the effective number of risky contracts should be small, so Test I with $\vartheta > 0.7$ is still expected to be a powerful but not very specific test to detect contracts at risk. Especially with a larger and better vocabulary, the conjunction of Test I and Test II (i.e., red flags activated when both tests are positive) could be a better device to detect potentially suspicious contracts.

The best employment of these tests is for prior screening of high volumes of contracts in public procurement. Correctly parameterised, Test I can reduce up to 1/6 the number of contracts to scrutinise in procedures of qualitative *audit*. This performance is expected to improve with a complete vocabulary from all the contracts from the BDNCP.

## 4. Future developments

In this work we present a text-mining approach for identifying dissimilarities between the textual description of a public contract object and its pre-assigned classification using CPV. After the construction of suitable vocabularies for each macro class (MCPV), two tests are proposed for detecting contracts whose object description contains terms/elements incoherent with their classification.

Results are strongly conditioned for the vocabulary built on a relatively small sample, but also from the heterogeneity in the population of public contracts. Heterogeneity is problematic because it induces tokens to join vocabularies of several MCPVs, despite the use of a particularly robust weighting method

for the relative frequency of each token (see Eq. 1). In fact, some MCPVs are associated with subject categories that can be clustered according to semantic affinities. These semantic clusters are not caught by CPV classification. Moreover, weighting schemes cannot correctly assign each token to one and only one MCPV, because all the MCPVs in the cluster are too semantically close. For example, for $\vartheta = .3$, the token "server" joins the vocabulary of two MCPVs, that is, MCPV 48 ("Software and information technologies") and MCPV 72 ("Software consultancy and development"). While the objects of the contracts associated with these two MCPVs may differ in the phenomenology (e.g., selling a license, then a good, against giving a service of consultancy), the semantics may basically be the same, e.g. the act of "installing a server" could fall within both the MCPVs.

In practice, the presence of tokens belonging to different MCPVs reduces the specificity of the test and induces a high quota of false positives of the red flag. However, the low specificity offers a high power of the test to detect anomalies. Considering the previous example, an *auditor* could be interested in checking if there were reasons for the installation of a server to fall in MCPV 48 or MCPV 72, so in general, even in the current state, the method has clearly usability as a filter for a qualitative inquire of a set of contracts.

To overcome this issue, the proposal can be improved in two ways. The first is to adopt an advanced system of object classification, minimising heterogeneity in the number of contracts and merging classes that are semantically clustered.

The second approach considers the adoption of a model-based classification instead of a logical one. It is still based on vocabularies and "bag-of-words" structures, but also adopts a system of large parametric equations on a unique vocabulary of non-trivial tokens instead of listing relevant terms for each MCPV. For example, assuming $K$ classes (MCPVs), there exists a matrix that assigns a coefficient, $\beta_{ik}$, for each pair of token ($i$) and MCPV ($k$), so that the greater the coefficient, the higher the pertinence of token $i$ with MCPV $k$. Given vector $\mathbf{x}$ of tokens from a contract object, an overall coefficient, say $\beta_k$, scores the pertinence of the contract to the class $k$ and can be derived from the coefficients related to the tokens in $\mathbf{x}$, for example, by taking their arithmetic mean.

Then, for each contract we have a distribution of $\beta_k$, with $k = 1, \ldots, K$. As outlined above, the contract has a pre-assigned MCPV ($k^*$), so the method would check if the assigned class is "coherent" with the distribution of $\beta_k$; for example, in the best scenario, the assigned MCPV coincides with the class associated with the highest classification score:

$$k^{(p)} = \underset{k}{\operatorname{argmax}}(\beta_k).$$

This procedure is technically analogous to "topic modelling", with classes as fixed topics (12; 13). As a detection method, it is flexible because it allows for setting richer and more complex rules for the activation of the red flag, accounting for expected semantic similarities between classes and the $\beta_{k^*}$, compared to other $\beta_k$.

DRAFT

## References


[1] Almeida, G., Revoredo, K., Cappelli, C., Maciel, C.: Improvement of transparency through mining techniques for reclassification of texts: The case of Brazilian transparency portal. In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, ACM Press, New York (2018) doi:10.1145/3209281.3209332

[2] Torres-Berru, Y., López Batista, V.F.: Data and Text Mining for the Detection of Fraud in Public Contracts: A Case Study of Ecuador's Official Public Procurement System. In: S. Berrezueta, K. Abad (eds.), Doctoral Symposium on Information and Communication Technologies, pp. 116–127. Springer (2022). doi: 10.1007/978-3-030-93718-8_10

[3] Romano, M.F., Baldassarini, A., Pavone, P.: Text Mining of Public Administration Documents: Preliminary Results on Judgments. In: D.F. Iezzi, D. Mayaffre, M. Misuraca (eds.), Text Analytics: Advances and Challenges, pp. 117–126. Springer (2020). doi: 10.1007/978-3-030-52680-1_10

[4] Rabuzin, K., Modrusan, N.: Prediction of Public Procurement Corruption Indices using Machine Learning Methods, In Proceedings of the 11th International Joint Conference on Knowledge Dis-

covery, Knowledge Engineering and Knowledge Management, pp. 333–340. SCITEPRESS, Vienna (2019). doi: org/10.5220/0008353603330340
[5] Cosinex GmbH.: Consultancy services for Common Procurement Vocabulary (CPV) expert group – Revision of CPV (2017). Available at https://ec.europa.eu/docsroom/documents/27821
[6] Fazekas, M., Cingolani, L., Tóth, B.: A comprehensive review of objective corruption proxies in public procurement: Risky actors, transactions, and vehicles of rent extraction (2017). Government Transparency Institute Working Paper Series No. GTI-WP/2016:03, Budapest.
[7] Gentzkow, M., Kelly, B., Taddy, M.: Text as data. J Econ Lit, **57**(3), 535–574. (2019) doi: 10.1257/jel.20181020
[8] Gnaldi, M., Del Sarto, S., Falcone, M., Troía, M.: Measuring corruption. In: Carloni E, Gnaldi M (eds.): Understanding and fighting corruption in Europe - From repression to prevention, pp. 43–71. Springer (2021)
[9] OECD: Analytics for integrity. Data-driven approaches for enhancing corruption and fraud risk assessments (2019). Available at www.oecd.org/gov/ethics/analytics-for-integrity.pdf
[10] Silge, J., Robinson, D.: Text Mining with R: A Tidy Approach. O'Reilly, Boston (2017). Available at www.tidytextmining.com
[11] Porter M.F.: Snowball: A language for stemming algorithms (2001). Available at https://snowballstem.org/texts/introduction.html
[12] Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456. ACM Press, New York (2013). doi: 10.1145/2488388.2488514
[13] Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H.: Topic Modeling of Short Texts: A Pseudo-Document View. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2105–2114, ACM Press, New York (2016). doi: 10.1145/2939672.2939880

DRAFT

# Explainable Machine Learning based on Group Equivariant Non-Expansive Operators (GENEOs). Protein pocket detection: a case study

Giovanni Bocchi[a], Alessandra Micheletti[a], Patrizio Frosini[b], Alessandro Pedretti[c], Andrea R. Beccari[d], Filippo Lunghini[d], Carmine Talarico[d], and Carmen Gratteri[e]

[a]Dept. of Environmental Science and Policy, University of Milan;
giovanni.bocchi1@unimi.it, alessandra.micheletti@unimi.it
[b]Dept. of Mathematics, University of Bologna; patrizio.frosini@unibo.it
[c]Dept. of Pharmaceutical Sciences, University of Milan; alessandro.pedretti@unimi.it
[d]DompÃ¨ Farmaceutici S.p.A.; andrea.beccari@dompe.com,
filippo.lunghini@dompe.com, carmine.talarico@dompe.com
[e]Dept. of Health Sciences, UniversitÃ  degli Studi "Magna GrÃ¦cia";
carmen.gratteri.ext@exscalate.eu

DRAFT

**Abstract**

Artificial intelligence (AI) is now widely diffused in everyday life, and it is quite often based on machine learning (ML) techniques. ML, in spite of being quite effective in many applications, is very often lacking of transparency. Here we propose a methodology for explainable ML, based on a geometric approach and we apply it to the case study of protein pocket detection. Indeed our approach may lead to a faster and more sustainable process for drug design.

***Keywords:*** XAI, XML, GENEOs, pocket detection.

## 1. Introduction

Equivariant operators are proving to be increasingly important in deep learning, in order to make neural networks more transparent and interpretable (2; 7). The use of such operators corresponds to the rising interest in the so called "explainable artificial intelligence" (6; 14), which looks for methods and techniques whose functioning can be understood by humans. In accordance with this line of research, Group Equivariant Non-Expansive Operators (GENEOs) have been recently proposed as elementary components for building new kinds of networks (3; 4; 8). Their use is grounded in Topological Data Analysis (TDA) and guarantees good mathematical properties to the involved spaces, such as compactness, convexity, and finite approximability, under suitable assumptions on the space of data and by choosing appropriate topologies.

A GENEO is a functional operator that transforms data into other data. By definition, it is assumed to commute with the action of given groups of transformations (equivariance) and to make the distance between data decrease (non-expansivity). The groups contain the transformations that preserve the "meaning" of our data, while the non-expansivity condition ensures that the operator simplifies the data metric structure. Both equivariance and non-expansivity are important: while equivariance reduces the

computational complexity by exploiting symmetries of data, non-expansivity guarantees that the space of GENEOs can be finitely approximated.

In this paper we will introduce GENEOs and we will show promising results obtained in an industrial application, namely protein pocket detection.

## 2. Basic definitions and properties of GENEO spaces

Let us now formalize the concept of GENEO, as was introduced in (3). We assume that a space $\Phi$ of functions from a set $X$ to $\mathbb{R}^k$ is given, together with a group $G$ of transformations of $X$, such that if $\varphi \in \Phi$ and $g \in G$ then $\varphi \circ g \in \Phi$. We call the couple $(\Phi, G)$ *perception pair*. We also assume that $\Phi$ is endowed with the topology induced by the $L_\infty$-norm $D_\Phi(\varphi_1, \varphi_2) = ||\varphi_1 - \varphi_2||_\infty,\ \varphi_1, \varphi_2 \in \Phi$. Let us assume that another perception pair $(\Psi, H)$ is given, with $\Psi$ endowed with the topology induced by the analogous $L_\infty$-norm distance $D_\Psi$, and let's fix a homomorphism $T : G \to H$.

**Definition 1.** *A map $F : \Phi \to \Psi$ is called a* group equivariant non-expansive operator (GENEO) *if the following conditions hold:*

1. $F(\varphi \circ g) = F(\varphi) \circ T(g)$ *for every $\varphi \in \Phi$, $g \in G$ (equivariance);*
2. $\|F(\varphi) - F(\varphi')\|_\infty \leq \|\varphi - \varphi'\|_\infty$ *for every $\varphi, \varphi' \in \Phi$ (non-expansivity).*

If we denote by $F_{all}$ the space of all GENEOs between $(\Phi, G)$ and $(\Psi, H)$ and we introduce the metric

$$D_{\text{GENEO}}(F_1, F_2) = \sup_{\varphi \in \Phi} ||F_1(\varphi) - F_2(\varphi)||_\infty, \qquad \forall\, F_1, F_2 \in F_{all}$$

the following main properties of $F_{all}$ can be proven (see (3) for the proofs).

**Theorem 1.** *If $\Phi$ and $\Psi$ are compact, then $F_{all}$ is compact with respect to the topology induced by $D_{GENEO}$.*

**Theorem 2.** *If $\Psi$ is convex, then $F_{all}$ is convex.*

Theorem 1 guarantees that if the spaces of data are compact, then also the space of GENEOs is compact, thus it can be well approximated by a finite number of representatives, reducing thus the complexity of the problem. Theorem 2 implies that if the space of data is also convex, then any convex combination of GENEOs is still a GENEO. Thus when both properties hold we have an easy instrument to obtain new GENEOs starting from a finite number of them.

DRAFT

## 3. GENEOnet

We used GENEOs to build *GENEOnet* (5), a geometrical explainable machine learning method to detect pockets on the surface of proteins which are likely to host ligands (where ligands are usually drugs). Protein pockets detection is a core problem in the context of drug development, since the ability to restrict the search only to a finite number of good sites, allows a scientist to speed up virtual screening procedures, saving both computational resources and time.

This problem is particularly suitable to be treated with GENEOs. In fact there is some important empirical chemical-physical knowledge that can not be directly embedded in the usual machine learning techniques, but can be injected in a GENEO architecture. Additionally the problem shows a natural equivariance property, since, if we rotate or translate a protein, its pockets will be coherently transformed in the same way. This suggests that pocket detection is equivariant with respect to the group of spatial isometries.

To use GENEOs, input data have been processed by surrounding each protein with a bounded region divided into a 3D grid of voxels. In this way we could represent the data as bounded functions from the

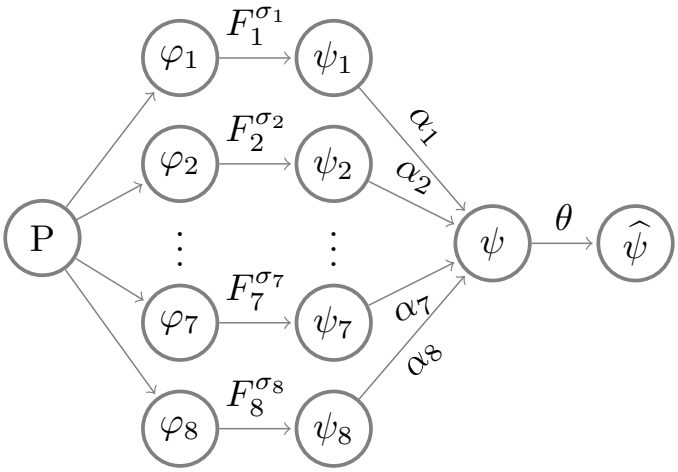



Figure 1: Model workflow: input channels $\varphi_1, \ldots, \varphi_8$ are fed to the GENEOs $F_1, \ldots, F_8$ dependent on the shape parameters $\sigma_1, \ldots, \sigma_8$. The intermediate outputs $\psi_1, \ldots, \psi_8$ are combined through convex combination with weights $\alpha_1, \ldots, \alpha_8$ to get the final result $\psi$. To get predictions a thresholding operation with a parameter $\theta$ is applied obtaining the binary function $\widehat{\psi}$

Euclidean space $\mathbb{R}^3$ to $\mathbb{R}^d$. Indeed, we considered $d = 8$ distinct geometrical, chemical and physical potential fields, called *channels*, that were computed for each molecule [1].

Such functions are fed to a layer of GENEOs chosen from a set of parametric families of operators, each one parametrized by a shape parameter $\sigma_i, i = 1, \ldots, d$. These families were designed in order to include the a priori knowledge of the experts of medicinal chemistry. We opted for convolutional operators with $L^1$ normalized kernels: if $\varphi$ is one of the considered channel measured on a molecule, we compute

$$F_k(\varphi) = \int_{\mathbb{R}^3} \varphi(x) k(x - y)\, dy,$$

where $k$ is a kernel. The behavior of such operators is determined by their kernels, thus by making the $i$-th kernel dependent only on one shape parameter $\sigma_i$, we have direct control on the action of each operator. We mainly used Gaussian kernels[2], or kernels having shapes of spheres, or of spherical crowns, assuming alternatively positive and negative values in different parts of the interior of the sphere or crown, and zero outside (see (5) for further details about the kernels). Nonetheless all the kernels are rotationally invariant functions. This fact, together with the properties of convolution, guarantees that the corresponding operators satisfy the key requirement to be equivariant with respect to the group of isometries of $\mathbb{R}^3$.

In the second layer the $d$ operators are combined through a convex combination, with weights $\alpha_1, \ldots, \alpha_d$, with $\alpha_i \in [0, 1], \forall i$ and $\sum_{i=1}^{d} \alpha_i = 1$. The output of the convex combination is normalized to a function $\psi$ from $\mathbb{R}^3$ to $[0, 1]$. Here $\psi(x)$ can be read as the probability that a point $x \in \mathbb{R}^3$ belongs to a pocket. Finally, considered a probability threshold $\theta \in [0, 1]$, we obtain the different predicted pockets by taking the connected components of the superlevel set $\{\psi \geq \theta\} \subseteq \mathbb{R}^3$. The entire model pipeline is depicted in Figure 1.

The model described until now relies just on 17 free parameters. The fact that the model only employs convolutional operators, and their linear combinations, allowed to set up a training pipeline quite similar to that of a 3D Convolutional Neural Network (CNN), but with two fundamental differences. First of all GENEOnet has a really tiny set of parameters[3]. Additionally the convolutional kernels of the GENEOs are not learned entry by entry as in classical CNNs (in this way equivariance would not be

[1] See (5) for further details on the specific channels.

[2] that is with general expression $k(x) = C \exp\left(-\frac{||x||^2}{2\sigma_i}\right)$

[3] For comparison DeepPocket (1), a recent approach that uses a 3D CNN, has 665 122 parameters.

DRAFT

preserved), instead the kernels are continuously generated from the values of the shape parameters that are updated during the optimization.

## 4. Model training and comparison with other methods

In order to identify the unknown parameters, we chose to optimize a cost function that evaluates the goodness of our predictions, in terms of volume fraction of the cavity which contains the ligand that has been correctly identified. Eventually, after training, pockets are found as the connected components of the thresholded output of the model, resulting in a set of unordered pockets. Actually this representation is not much informative, since it is usually desirable to compute also the "druggability" of the identified cavities, that is a ranking score for the pockets, on the basis of their fitness to host a ligand. Thus we used the 'not thresholded' output $\psi$ of the model to score the pockets, so that the final output consists in a list of pockets, ranked by their corresponding scores (5).

In order to identify the optimal model, we opted for a two-step optimization procedure: in the first step we generated $m = 200$ models $(\mathcal{M})_{k=1}^{m}$ optimized from $(T_k, IC_k)_{k=1}^{m}$, where $T_k$ is a training set of size 200, subsampled from the whole dataset and $IC_k$ are the randomly generated initial values of the parameters. In the second step each model was evaluated for its scoring capabilities, by computing $H_1$ (see (1) for the definition) on a validation set in order to select the model maximizing $H_1$. This final model was evaluated on an independent (both from the training and the validation sets) test set to produce the results of next paragraph.[4]

We compared the results of GENEOnet with other recent methods for protein pocket detection, some of which based on ML techniques. We based our comparison on the scores assigned by the different methods to the pockets. In this way we compared the models in the assignment of highest scores to pockets matching the true ones. Given a dataset of proteins having only one ligand, and thus one "true pocket" each, we computed the following quantities

$$H_j = \frac{\#(\text{proteins whose true pocket is hit by the jth top ranked})}{\#(\text{proteins})} \tag{1}$$

and the corresponding cumulative quantities

$$T_j = \frac{\#(\text{proteins whose true pocket is hit within the jth top ranked})}{\#(\text{proteins})} = \sum_{i=1}^{j} H_j.$$

In this way different methods can be compared directly: if a model provides $T_j$ greater than all the others for all $j$, then that model is definitely better.

The results reported in Table 1 show that GENEOnet is better than all the other considered methods in detecting the true pocket within the first 4 top ranked pockets, and is only slightly less performing than DeepPocket and Fpocket in the general detection of the true pockets, whatever the ranking provided by the methods. Anyway in virtual screening procedures, only the first few top ranked pockets on each protein are considered for testing their ability to host a ligand (druggability), in order to speed up the process, that usually is applied to tens of thousands of different molecules. Thus the performance of a method must be judged on the ability of the first few top ranked pockets to detect the true one. Furthermore the computational efficiency and transparency of GENEOnet are much better than the other methods, since it depends only on 17 parameters, which can be trained very rapidly even with a small training set.

Additionally, the values of the estimated parameters $\sigma_i, \alpha_i,\ i = 1, \dots, d$ can be used to provide further interpretation to the results, and thus to increase the explainability of GENEOnet: the relative magnitude of the shape parameters $\sigma_i$ reflects their influence on the convolution kernels, while the values of the convex combination parameters $\alpha_j$ can be regarded as the corresponding channel's importance.

[4]Dataset size: 12995, Training set size: 200, Validation set size 3073 (48 proteins in the intersection with the training set), Test set size: 9070 (totally disjoint from the other sets). Protein data retrieved from PDBbind v2020 dataset (11)

DRAFT

| Method | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $\sum_{j\geq 1} H_j$ |
|---|---|---|---|---|---|
| GENEOnet (5) | **0.792** | **0.905** | **0.941** | **0.955** | 0.975 |
| P2Rank (10) | 0.728 | 0.847 | 0.892 | 0.917 | 0.952 |
| DeepPocket (1) | 0.652 | 0.798 | 0.860 | 0.896 | **0.978** |
| CAVIAR (13) | 0.616 | 0.739 | 0.783 | 0.806 | 0.837 |
| SiteMap (9) | 0.424 | 0.502 | 0.529 | 0.542 | 0.558 |
| Fpocket (12) | 0.331 | 0.462 | 0.534 | 0.585 | **0.978** |
| CavVis (15) | 0.224 | 0.376 | 0.483 | 0.567 | 0.842 |

Table 1: $T_j$ values for the methods in the comparison. In the last column the fraction of molecules whose correct pocket has been identified by at least one of the ranked cavities is reported. See (5) for the full analysis.

## 5. Discussion

We presented GENEOnet, a geometric machine learning model based on a new explainable artificial intelligence paradigm provided by the theory of GENEOs. As described, a GENEO-based approach needs some prior knowledge to be set up, but, if available, such prior knowledge can be effectively combined with the equivariance property of GENEOs in order to reduce the number of learnable parameters and the hunger for training examples of deep learning models.

A GENEO-based model possesses multiple sources of explainability: equivariance allows the model to embed in itself prior knowledge about specific geometric transformations of the data, so that this knowledge must not be learned from the training set; the low number of free parameters gives the possibility of assigning clear and understandable meaning to them; and, finally, the non-expansiveness property ensures that a GENEO-based model is stable, in the sense that close data will be mapped to close outputs.

The application to the detection of druggable pockets on the surface of proteins considered here, allowed us to show many of the stated properties of GENEOs architectures; moreover, in this case study, the use of GENEOs did not cause a reduction in model performance. On the contrary, GENEOnet achieves better or comparable results than all other models in use, but with huge advantages in terms of training time and transparency. For this reason, GENEOnet may help to improve the efficiency and speed of biochemical virtual screening procedures and lead to faster more reliable processes for drug design.

DRAFT

## References

[1] Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. V., and Priyakumar, U. D.: DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. Journal of Chemical Information and Modeling (2021) doi: 10.1021/acs.jcim.1c00799

[2] Anselmi, F., Evangelopoulos, G., Rosasco, L., and Poggio, T. : Symmetry-adapted representation learning. Pattern Recognition (2019) doi: 10.1016/j.patcog.2018.07.025

[3] Bergomi, M. G., Frosini, P., Giorgi, D., and Quercioli, N.: Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0087-3

[4] Bocchi, G., Botteghi, S., Brasini, M., Frosini, P., and Quercioli, N.: On the finite representation of group equivariant operators via permutant measures. Annals of Mathematics and Artificial Intelligence (in press) (2023) doi: 10.1007/s10472-022-09830-1

[5] Bocchi, G., Frosini, P., Micheletti, A., Pedretti, A. *et al.*: GENEOnet: A new machine learning paradigm based on Group Equivariant Non-Expansive Operators. An application to protein pocket detection. (2022) preprint at arXiv:2202.00451.

[6] Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.J. *et al.*: Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Scientific Reports

(2021) doi: 10.1038/s41598-021-83922-6
[7] Cohen, T. and Welling, M.: Group equivariant convolutional networks. In proceedings of the International Conference on Machine Learning (2016).
[8] Conti, F., Frosini, P., and Quercioli, N.: On the Construction of Group Equivariant Non-Expansive Operators via Permutants and Symmetric Functions. Frontiers in Artificial Intelligence (2022) doi: 10.3389/frai.2022.786091
[9] Halgren, T.: New method for fast and accurate binding-site identification and analysis. Chemical Biology & Drug Design (2007) doi: 10.1111/j.1747-0285.2007.00483.x
[10] Krivak, R. and Hoksza, D.: P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of Cheminformatics (2018) doi: 10.1186/s13321-018-0285-8
[11] Liu, Z., Su, M., Han, L., Liu, J. *et al.*: Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Accounts of Chemical Research (2017) doi: 10.1021/acs.accounts.6b00491
[12] Le Guilloux, V., Schmidtke, P., and Tuffery, P.: Fpocket: An open source platform for ligand pocket detection. BMC Bionformatics (2019) doi: 10.1186/1471-2105-10-168
[13] Marchand, J.R., Pirard, B., Ertl, P., and Sirockin, F.: CAVIAR: a method for automatic cavity detection, description and decomposition into subcavities. Journal of Computer-Aided Molecular Design (2021) doi: 10.1007/s10822-021-00390-w
[14] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019) doi: 10.1038/s42256-019-0048-x
[15] Simoes, T. M. C., and Gomes, A. J. P.: CavVis-A Field-of-View Geometric Algorithm for Protein Cavity Detection. Journal of Chemical Information and Modeling (2019) doi: 10.1021/acs.jcim.8b00572

DRAFT

# Hedging global currency risk with factorial machine learning models

Paolo Pagnottoni and Alessandro Spelta

Department of Economics and Management, University of Pavia.
Corresponding author: Paolo Pagnottoni, paolo.pagnottoni@unipv.it,
via San Felice 5, 27100 Pavia, Italy .

**Abstract**

We propose a dynamic method to hedge foreign exchange risk of international equity portfolios. The method is based on the currency returns predictions obtained from a set of alternative machine learning models, built on the main factorial components of the time series currency returns. The analysis of several model performance indicators allows to conclude that accurate predictions of global factor returns, such as those obtained with non linear machine learning models, can improve currency risk hedging.

***Keywords:*** Currency returns, Currency risk factors, Mean-variance optimization, Time series machine learning.

DRAFT

## 1. Introduction

Global investment positions are nowadays substantial and have grown quickly in recent decades. When redenominated in the investors' home currency, the currency exposure resulting from international investments can significantly alter the underlying assets' overall risk-return profile, highlighting the key question of how investors should manage their foreign exchange exposure.

The main strategy for controlling foreign exchange (FX) exposure is mean-variance optimization, which establishes the best currency hedging positions. The strategy is theoretically interesting and includes requirements for both speculative hedging and risk management naturally. However, due to the well-known difficulties in forecasting exchange rates, this strategy, when used out-of-sample, suffers from severe estimating error in currency returns - see (3) -, leading to poor overall currency hedging performance - see (1; 2).

To solve the problem, in the paper we propose a novel methodology to dynamically determine currency hedge positions, that we denominate dynamic currency factor (DCF) hedging. The approach exploits the predictability of currency returns desumed from the predictability of the corresponding risk factors. Recent breakthroughs in international macro-finance by Lustig et al. (2011) and Verdelhan (2018) have found that the cross-section of currency returns can be explained as compensation for risk in a linear factor model that includes two global currency risk factors: dollar and carry. The dollar factor corresponds to the average return of a basket of currencies against the US dollar, while the carry factor reflects the returns from currency carry trade. These factors also account for a large proportion of the time-series exchange rate behavior in contemporaneous regressions, and thus, if their returns are predictable, currency returns are also partially predictable. We show that exploiting a predictable component in both global factors helps to mitigate the estimation error that typically hinders traditional mean-variance currency hedging, thereby delivering important investment gains.

Specifically, we build time series machine learning models to accurately forecast factor returns and improve the predictability of single currency returns.

To illustrate our proposed methodology from an empirical viewpoint, in the empirical part of the paper we take the perspective of a US investor who invests in a portfolio of G10 developed economies. The investor is assumed to have a predetermined long position in either foreign equities or bonds and desires to manage the Foreign Exchange exposure by forming optimal hedge positions within a mean-variance optimisation framework. To construct optimal hedge positions, monthly estimates of currency returns are initially formed by estimating currenciesâ exposures to dollar and carry factors (that is, factor betas). Factor returns are then forecasted using variables that have been theoretically motivated to drive either one or both factor returns, including Foreign exchange volatility (Merton, 1973; Menkhoff et al., 2012a; Cenedese et al., 2014), the average forward discount (Lustig et al., 2014), the TED spread (Brunnermeier and Pedersen, 2009; Brunnermeier et al.,2009), and commodity returns (Ready et al., 2017). These predicted global factor returns are combined with the estimated factor betas to form out-of-sample (OOS) forecasts of currency returns. The predicted currency returns are finally incorporated within the mean-variance optimizer to produce optimal, currency-specific, hedge positions.

Our empirical results show that the estimation error which frequently affects mean-variance currency hedging can be reduced and that significant economic investment gains can be achieved by taking advantage of a non-linear forecasting of components in both global factors.

# 2. Methodology

## 2.1 Framework for deriving optimal currency hedge positions

The main goal is to dynamically create the best currency hedge positions, from the viewpoint of a US investor. FX forward contracts are added to an existing portfolio of reference international equity indices in order to hedge FX exposure. Currency hedges are chosen to optimize a mean-variance investor's utility, whose objective function is used to determine hedge positions, and is specified as

$$\mu_{p,t} - \frac{\gamma}{2}\sigma^2_{p,t} \tag{1}$$

where $\mu_{p,t} = w_t'\mu_t$ represents the expected portfolio return for the following period, $\sigma^2_{p,t} = w_t'\Sigma_t w_t$ represents the portfolio risk, $\mu_t$ is the vector of expected excess returns in US dollars with variance-covariance matrix $\Sigma_t$. The vector $w_t$ encompasses portfolio weights, and $\gamma$ is the investor's level of risk aversion.

Portfolio weights in an unconstrained setting are given by $w_t^* = \frac{1}{\gamma}\Sigma_t^{-1}\mu_t$. In our setup, it is adequate to assume the underlying asset weights are predetermined by the portfolio manager, therefore determining the weights given to FX forward contracts is the ultimate goal (i.e., the optimal currency hedge positions). By first partitioning the expected return vector and related covariance matrix across the underlying equity indices and FX future contracts, the optimization issue can be rephrased as

$$\mu_t = \begin{pmatrix} \mu_{x,t} \\ \mu_{f,t} \end{pmatrix}, \quad \Sigma_t = \begin{pmatrix} \Sigma_{xx,t} & \Sigma_{xf,t} \\ \Sigma_{fx,t} & \Sigma_{ff,t} \end{pmatrix} \tag{2}$$

where the underlying indices and FX forwards are represented by $x$ and $f$, respectively. Hence, the optimal weights in foreign exchange forwards are then given by

$$w^*_{f,t}(f \mid x) = \frac{1}{\gamma}\left(\Sigma^{-1}_{ff,t}\mu_{f,t}\right) - \delta_t w_{x,t} \tag{3}$$

where $w_{x,t}$ is the vector of weights and $\delta_t$ is the regression coefficient obtained from regressing underlying indices returns on long FX forward contracts returns, namely $\delta_t = \Sigma^{-1}_{ff,t}\Sigma_{fx,t}$.

Each element of $w^*_{f,t}$ is constrained to be between $-w_{x,t}$ (fully hedged) and zero (unhedged). This restriction mirrors the practice in currency overlay management that prohibits managers from taking speculative FX forward holdings above the position in the underlying security - i.e., a position cannot

DRAFT

be overhedged or leveraged by FX forwards. We estimate $\delta_t$ and $\Sigma_t$ each month using a 5-year rolling window and set the underlying portfolio weights to be equally weighted across indices.

## 2.2 Building currency factors

We firstly define the currency excess return. We use the symbols $S$ and $f$ to represent the log of the spot exchange rate and the forward exchange rate, respectively, both expressed in terms of foreign currencies per US dollars. The value of the domestic currency rises when the dollar value does. Simply put, the net log currency excess return for an investor who goes long in foreign currency $i$ is

$$r^l_{i,t+1} = f_{i,t} - s_{i,t+1} \tag{4}$$

The investor purchases foreign currency or, equivalently, sells the dollar forward at the price $(f_t)$ at time $t$ and purchases dollars at the price $(s_{t+1})$ in the spot market in $t+1$. Similarly, for an investor who is short the foreign currency - and therefore long in the dollar - the net log currency excess return is given by:

$$r^s_{i,t+1} = -f_{i,t} + s_{i,t+1} \tag{5}$$

We divide the sample of currencies into four portfolios at the end of each period $t$ based on the forward discounts

$$d_{i,t} = f_{i,t} - s_{i,t} \tag{6}$$

that observed at that time. Portfolio 1 comprises the currencies with the lowest interest rates or smallest forward discounts, and portfolio 4 contains the currencies with the highest interest rates or largest future discounts. They are ranked from low to high interest rates. By averaging the log currency excess returns for each portfolio $j$, we can get the log currency excess return $r^j_{i,t+1}$ for portfolio $j$. We make the assumption that investors will short all of the foreign currencies in the first portfolio in order to calculate returns.

According to predictions made by linear factor models, the average returns on a variety of assets can be attributed to risk premia related to their exposure to a limited number of risk factors. Our currency portfolios' principal component analyses show that two factors account for more than 65% of the difference in returns across these four portfolios. Since all portfolios load equally on the first principal component, which we labeled $\lambda^{dol}$, it can be said that it acts as a level factor and accounts for more than 65% of the common variation in portfolio returns. Given that portfolio loadings rise monotonically across portfolios, the second primary component, denoted $\lambda^{car}$, which accounts for about 22% of common variance, can be seen as a slope factor. The systematic component of the dollar factor appears to correlate with low frequency global business cycle conditions and can therefore be interpreted as capturing the level of global macroeconomic risk - see (4). The carry factor is a zero-cost portfolio, constructed by investing in high-yielding currencies while funding the position in low yielding currencies.

## 2.3 Expected currency returns and common risk factors

The expected currency return vector $\mu_{f,t}$ is a key input in eq. (3), given that the return at time $t+1$, to a US investor who enters a long forward contract on foreign currency $i$ at time $t$ is defined by eq. (4).

We model the cross-section of expected currency returns as function of two common risk factors, named the dollar and carry factors:

$$Er_{i,t} = \zeta(\lambda_t^{dol,car}) \tag{7}$$

Factor return predictability is then exploited for forecasting expected returns. In particular, we run predictive models of factor returns on a set of time $t-1$ predictor variables $X_{t-1}$ described by:

$$\lambda^j_t = \psi_j(X_{t-1}); \quad j = \{dol, car\} \tag{8}$$

to derive multi-step-ahead conditional expected factor returns $E_t\lambda^j_{t+h} = \hat{\psi}_j X_t$ where $h = t+1, ..., t+H$ where $H$ is the forecast horizon.

DRAFT

Given the estimated factor betas from eq. (7) and the expected factor returns (premiums) from eq. (8), conditional expected currency return over the following H periods for each currency pair $i$ can be found as:

$$E_t R_{i,t+h} = \hat{\zeta}(E_t \lambda^j_{t+h}) \tag{9}$$

In our setting, we propose to estimate the relationships between $Er_i$ and $\lambda^{dol,car}$, which we label as $\zeta$, and those among $\lambda^{dol,car}$ and $X$ ($\psi_j$) by means of time series machine learning models. Many of the machine learning algorithms are well known for delivering good forecasting performances and can be exploited to accurately forecast the dollar and carry factors and the conditional expected currency returns. In particular, we test different families of machine learning forecasting models for time series, namely Discriminant Analysis for time series (DA), Regression Trees (RT) and Support Vector Machine (SVM).

## 2.4 Currency factor predictors

As currency factor predictors we first exploit the forward currency discount $d_{i,t}$ represented in eq. (6). Secondly, since the carry factor returns exhibit a strong negative relationship with FX volatility we make use of exchange rates volatility measured as the daily squared returns currency pairs $i$ against the US dollar. Namely, we exploit the change in FX volatility:

$$\Delta\sigma^i_t = \log\left(\frac{\sigma^i_t}{\sigma^i_{t-1}}\right)$$

where $\sigma^i_t = r^2_{i,t}$. Thirdly, given that tighter funding liquidity, as proxied by increases in the spread among the LIBOR and the Tbills, can also forecast carry returns, we employ changes in the LIBOR ($LIBOR_t$) and in the Tbills ($Tbill_t$) as exogenous predictors:

$$\Delta LIBOR_t = \log\left(\frac{LIBOR_t}{LIBOR_{t-1}}\right)$$

$$\Delta Tbill_t = \log\left(\frac{Tbill_t}{Tbill_{t-1}}\right)$$

Fourthly, it is established that higher commodity prices predict higher carry trade returns. Therefore, we employ two commodity price indices as currency factor predictors. The first one is the Invesco DB Commodity Index Tracking ($CI^1$) which tracks changes in the level of the DBIQ Optimum Yield Diversified Commodity Index Excess Return plus the interest income from the Fund's holdings of primarily US Treasury securities and money market income. The second one is the iShares S&P GSCI Commodity-Indexed Trust ($CI^2$) which tracks a set of futures contracts on an index composed of a diversified group of commodities futures. Also in this case, we use changes in the commodity price indices as currency factor predictors, namely

$$\Delta CI^1_t = \log\left(\frac{CI^1_t}{CI^1_{t-1}}\right)$$

$$\Delta CI^2_t = \log\left(\frac{CI^2_t}{CI^2_{t-1}}\right)$$

# 3. Data

For the purpose of illustrating our method, collect and analyze three main types of time series data: a) currency spot and future returns; b) international reference equity market indices; c) a set of currency factor predictors. Firstly, we consider the following countries and their respective foreign exchange rates: Australia (AUD), Canada (CAD), Switzerland (CHF), Europe (EUR), Great Britain (GBP), Hong

DRAFT

Kong (HKD), Japan (JPY), Norway (NOC), New Zealand (NZD), Poland (PLN), Sweden (SEK), and South Africa (ZAR). For each currency, we collect spot and future closing prices. Secondly, we retrieve the equity prices of the reference stock market of each country[1]. Thirdly, we retreive daily time series observations related to the exogenous predictors discussed in Subsection 2.4. The analyzed time period ranges from 27 October 2008 to 30 December 2022.

## 4. Empirical results

In this Section we present the results of our methodology. The model is back-tested using a dynamic approach. For the in-sample estimation, we choose a 5-year rolling window, each one shifted by 22 working days, i.e. a trading month. For each of them, we produce out-of-sample predictions for the following month. Predictions are obtained by the selecting, for each rolling window, the best performing model across DA, RT and SVM in terms of cross-validation error. To ease computational burden, we use Bayesian optimization for training the set of different models and tuning their hyperparameters. Bayesian optimization finds an optimal set of hyperparameters for a given model by minimizing the objective function of the model, strategically selecting new hyperparameters for each iteration and tipically outperforming parameter grid search.

Figure 1 (upper panel) shows the cumulative payoff to investing in global equity portfolios, and specifically the cumulative payoff to a USD 1 invested in equally weighted global equity portfolios under different currency hedging frameworks from the perspective of a US investor. The payoffs to the linear DCF hedged portfolio, our proposed DCF hedged portfolio and the unhedged portfolio are highlighted. The profit and loss metric is complemented with two performance measures as highlighted in the central and lower panels of Figure 1, which illustrate the Sharpe and the Maximum Drawdown, respectively. Both measures are obtained with a rolling window of two years. Finally, we also report the empirical distribution of returns for the unhedged and hedged portfolio, along with the Kolmogorov-Smirnov test statistics and associated p-value.

We notice that hedged portfolios, on average, outperform the unhedged alternative. This suggests that hedging for global currency risk factors is determinant to the profitability of international equity porfolios. This is further confirmed by the sistematically superior values of the Sharpe ratios and Maximum Drawdown. Moreover, our proposed ML-hedging produces superior performances with respect to the linear hedged portfolio. The Kolmogorov-Smirnov test provides evidence against the null hypothesis that the returns of the two portfolios come from the same continuous probability distribution at all conventional significance levels ($p = 0.00073$).

DRAFT

## 5. Conclusion

In this paper we have provided a dynamic method to hedge exposure of international equity portfolios against foreign exchange risk via predictive ML models for global factor and currency returns. The method is based on the currency returns predictions that have been produced by training several ML models constructed using the principal factor components of the currency return time series. The comparison of various model performance metrics leads us to draw the conclusion that a significant currency risk hedging of global stock portfolios can be achieved by making accurate predictions of global factor returns, such as those derived with non linear machine learning models. Automated portfolio management via ML does not only offer a better hedge against currency risk, but also enhanced portfolio returns with respect to linear forecasting models.

[1]For the EURO area, we select the DAX as reference stock market index.

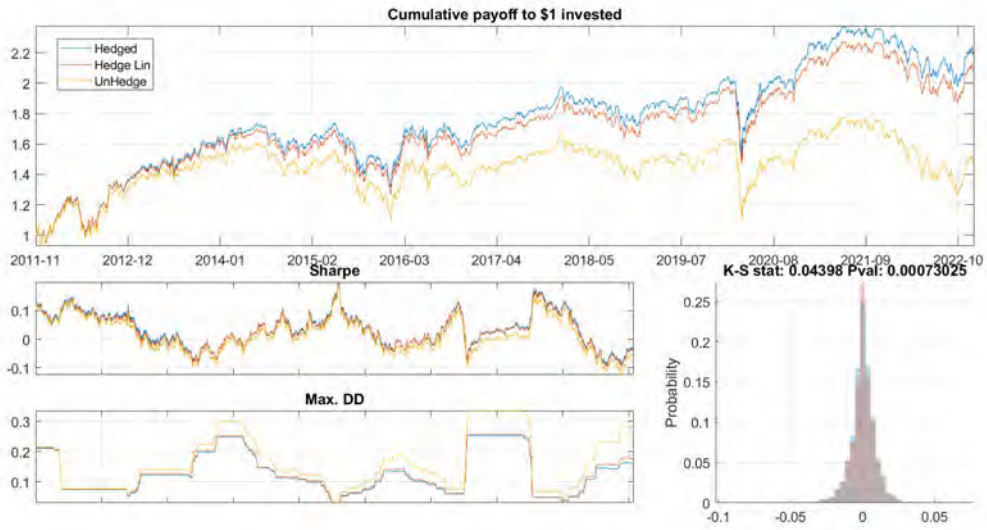



Figure 1: **Performance of ML-hedged, linear hedge and unhedged portfolios**. The upper panel shows the cumulative payoff to a USD 1 invested in equally weighted global equity portfolios under different currency hedging frameworks (ML-hedged, linear hedge and unhedged) from the perspective of a US investor. The central and lower panels of show the Sharpe and the Maximum Drawdown, both measures are obtained with a two-year rolling window. The right lower panel reports the empirical distribution of returns for the unhedged and hedged portfolio, along with the Kolmogorov-Smirnov test statistics and associated p-value.

DRAFT

## References

[1] Gardner, G.W., Stone, D., 1995. Estimating currency hedge ratios for international portfolios. Financial Analysts Journal 51, 58–64.

[2] Larsen, Jr, G.A., Resnick, B.G., 2000. The optimal construction of internationally diversified equity portfolios hedged against exchange rate uncertainty. European Financial Management 6, 479–514.

[3] Meese, R.A., Rogoff, K., 1983. Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of international economics 14, 3–24.

[4] Verdelhan, A., 2018. The share of systematic variation in bilateral exchange rates. The Journal of Finance 73, 375–418.

# InstanceSHAP: An instance-based estimation approach for Shapley values

Golnoosh Babaei[a] and Paolo Giudici[b]

[a]University of Pavia; golnoosh.babaei01@universitadipavia.it
[b]University of Pavia; Department of Economics and Management; paolo.giudici@unipv.it

**Abstract**

Explanations are necessary objects of a decision-making problem. However, complex Machine learning (ML) models that are usually used to provide decisions lack explanations. Model-agnostic explanation methods are the solutions for this problem and can find the contribution of each variable to the prediction of any ML model. Among these methods, SHapley Additive exPlanations (SHAP) is the most commonly used explanation approach which is based on game theory and requires a background dataset when interpreting an ML model. In this study we evaluate the effect of the background dataset on the explanations. In particular, we propose a variant of SHAP, InstanceSHAP, that use instance-based learning to produce a background dataset for the Shapley value framework. More precisely, we focus on Peer-to-Peer (P2P) lending credit risk assessment and design an instance-based explanation model, which uses a more similar background distribution. Experimental results reveal that the proposed model can effectively improve the ordinary shapley values and provide more robust explanations.

***Keywords:*** Feature attribution; Shapley values; Machine Learning; Explainability.

DRAFT

## 1. Introduction

The field of credit risk management has grown rapidly in the last decades. This growth has led to a significant improvement in the performance, specifically accuracy, of prediction models, usually ML models. The trade-off between model complexity and interpretability makes it difficult to understand why sophisticated ML models perform so well. This problem prevents final users of decision-making models from trusting the predictions (10).

A key solution to this issue is the recent introduction of model-agnostic explanation approaches to understand the contribution of each predictor towards the overall prediction (2). (7) proposed a model-agnostic attribution method, SHapley Additive exPlanations (SHAP), based on the Shapley value from game theory. The Shapley value, quantifies the average contribution of a feature when it is added to all possible coalitions without the feature of interest. SHAP has been a popular method in the feature attribution literature (5) and many efforts have been devoted to improve it. For example, (1) address the problem of dependent variables in the SHAP algorithm. (3) mentions the heavy computational costs of calculating the Shapley values. Also in another study by (6), a rigorous analysis is performed to find where Shapley values are mathematically suboptimal. As computation of shapley values is based on the coalitions of the features, when dimension of the data increases, SHAP becomes computationally expensive. To reduce this complexity, SHAP (7) includes various explainers (e.g. TreeExplainer) for different ML models. A prerequisite of these explainers is a background dataset that is usually randomly

sampled from the training dataset. More precisely, when a feature is removed from a coalition, it is not clear how to have it inside the model. However, Background data is the solution for this issue by estimating the missing feature as an expectation over a background distribution (9). In this paper, we try to improve SHAP explanations by focusing on the background data. In particular, despite other studies that search for the correlation between variables and use conditional sampling for dependent features, we study the similarities among observations to improve estimation of the shapley values. For this purpose, we use an instance-based approach to provide a conditional selection for the background data.

The rest of this paper is organized as follows. Section 2 represents the methods we use in this paper. Experimental study, including description of the selected data to validate the proposed model and the results are presented in Section 3. Finally, Section 4 concludes the study and contains some remarks for future study.

# 2. Methodology

## 2.1 SHAP

SHAP values are based on the definition of Shapley values (7), which finds the marginal contribution of each variable towards the predictions. From the game theory point of view, features of an instance in the data set behave as players in a coalition, and Shapley values represent the distribution of the prediction among the features according to their contribution. In this paper we use the approximation of the Shapley values proposed in (7) that simulates the absence of a feature using the marginal expectation over a background distribution $D$. The Shapley value of variable $i$ for $i = 1, \ldots, m$, is calculated as follows:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus i} \frac{|S|!(n-1-|S|)!}{n!}[v(S \cup i) - v(S)], \tag{1}$$

Here, $v(S)$ is equal to $= E_{x' \sim D[f(x_S, x^i)]}$. $E_{x' \sim D}$ denotes the expected value under the distribution $D$ and $f(x_S, x'_F \setminus S)$ represents prediction of the model with feature values included in $S$ and values $x^{\hat{a}^2}$ for feature values not in $S$. Also, $n$ is the total number of variables, $S$ is a subset of predictors.

When there are $n$ variables, we will have these calculations for $2^n$ coalitions so number of coalitions increases exponentially. This is the main disadvantage of shapley values that has led to approximations such as Kernel SHAP (7). In our application, as we employ random forest for the credit scoring model, we utilize a non-conditional version of TreeExplainer (8).

## 2.2 Instance-based Learning

The instance-based method has the ability to find some train observations that are the most similar instances to the test data that we pass through the explainers to calculate shapley values. Therefore, instead of random sampling from train data using the instance-based method we can find the similar samples and provide a more similar distribution for estimation of shapley values.

The weights for an instance-based model can be calculated using the distance between train and test instances. In this paper, we have a loan assessment problem so can define the default likelihood distance between loans i (train data) and j (test data), as follows

$$d_{ij} = |PD_i - PD_j| \tag{2}$$

Here $PD_i$ and $PD_j$ are the probabilities of default for loans i and j, respectively. It is obvious that we expect a higher amount of weight when observations are more near to each other. However, this is not an optimal weighing approach so as suggested by (4) we use kernel regression and follow their method to find the weights between train and test data. As the authors mention, kernel weights for instance-based modeling Kernel regression is a statistical technique to find non-linear relation between a pair of random variables. Similar to their scenario, we evaluate each observation (loan) based on risk and return but

DRAFT

Table 1: Summary statistics for the LendingClub data[a].

| Type | Variable | Mean | STD | Min | Max |
|---|---|---|---|---|---|
| Numerical | Loan Amount | 14588.273 | 8970.4714 | 500 | 40000 |
| | Annual Income | 77369.565 | 117821.8 | 0 | 110000000 |
| | Debt To Income (DTI) | 18.567739 | 13.08763 | -1 | 999 |
| | Open Accounts | 11.60594 | 5.5755073 | 0 | 90 |
| | Inquiries In The Last 6months | 0.6125763 | 0.9018182 | 0 | 8 |
| | Delinquencies In 2Years | 0.312875 | 0.8754565 | 0 | 42 |
| | Public Records | 0.2083065 | 0.5901983 | 0 | 86 |
| | Interest Rate | 13.170133 | 4.8283 | 5.31 | 30.99 |
| Categorical | LC Grade | 7 Grades, From A(Safest) to G(Riskiest) | | | |
| | Loan Purpose | 8 Categories, medical, car and etc. | | | |
| | Home Ownership | 4 categories, Rent, Own, Rent and Any | | | |
| Binary | Loan Status | 1=Failed Loans, 0=Fully Paid Loans | | | |

[a]See https://www.kaggle.com/datasets/jonchan2003/lending-club-data-dictionary

since the focus of our paper is proposing an explainable model and improve the explainability we utilize interest rate, already available in data, as the return indicator and we do not calculate another measure to evaluate the return of loans. Kernel weights as supposed by (4) between the $i^{th}$ train observation ($i = (1, \ldots, n)$) and the $j^{th}$ test observation ($j = (1, \ldots, m)$) are calculated using the following equation:

$$W_{ij} = \frac{K(\frac{d_{ij}}{h})}{\sum_{j=1}^{n} K(\frac{d_{ij}}{h})} \tag{3}$$

Here, $h$ ($h$¿0) is the bandwidth which follows the definition given by (4). After finding the optimal $h$, the weights between train and test observations are calculated. These weights are then used to find the similar background distribution for estimation of Shapley values.

DRAFT

# 3. Application

## 3.1 Data Description

We use the most common dataset in the P2P Lending Credit Scoring literature. This dataset includes information of individuals applied for a loan on the Lending Club P2P lending platform. The time horizon we consider for our paper is from 2007 to 2019. In general all observations are grouped into two categories: Class=zero, including borrowers who paid their loans and Class=one, representing borrowers who failed to pay the loan.

LC dataset is a big dataset which includes more than 100 variables. However, we select only mostly common used variables in the literature. After preprocessing the data, including dealing with missing values and encoding, explanatory analysis of the selected variables are as presented in Table 1. Loan status (a binary variable) is used as the target variable in our model.

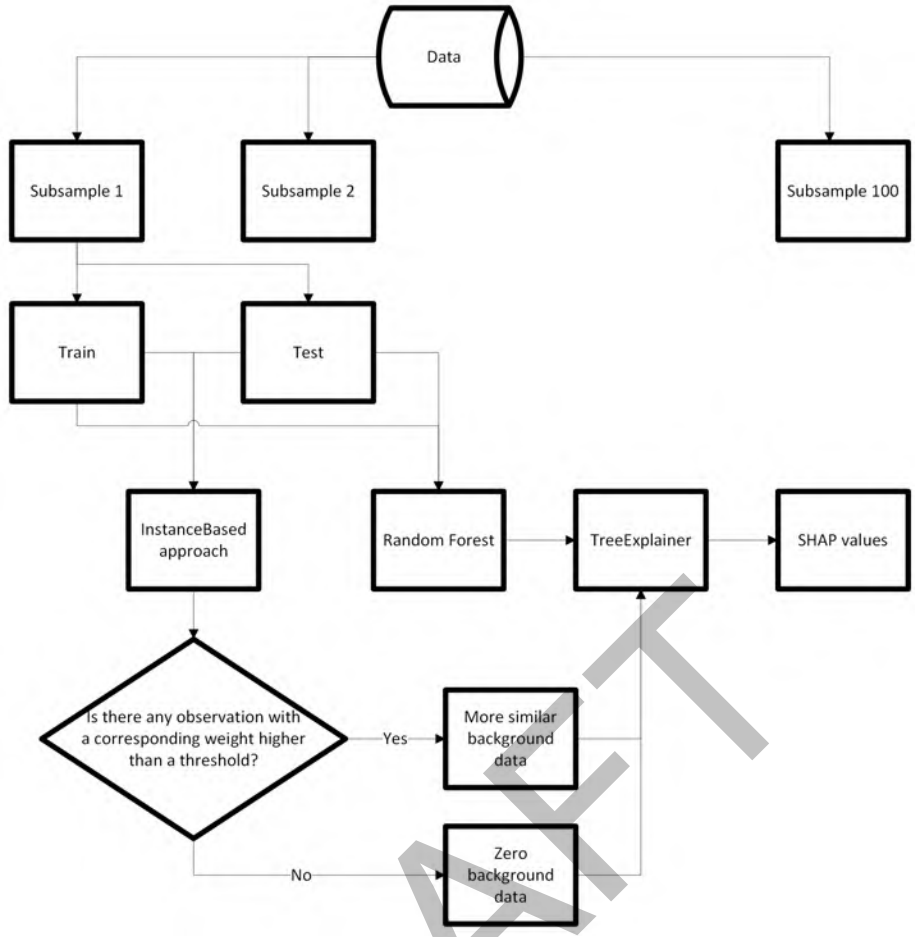



Figure 1: InstanceSHAP Flowchart

## 3.2 Empirical Approach Setting

Our proposed algorithm, InstanceSHAP, is illustrated in Figure 1.

As shown in Figure 1, we sample the processed data to a set of subsamples while the probability of default is being kept. Then to start the ML model that in our case is a random forest, we split subsamples to train and test data. We repeat the following operations for each set of train and test data generated by each sub sample and finally we take the mean of the results among the subsamples. The first step to find the weights between train and test observations is finding the distance among them which is calculated using Equation 2. After that, kernel weights using the calculated distance values and the optimal bandwidth are found based on Equation 3 and the description in (4). It should be mentioned that for optimizing the bandwidth we consider [0.25,1.5) with a step equal to 0.1.

Having the similarities between train and test data, we select the observations from train data for which the corresponding weight values are higher than a threshold that is set to be equal to the average of the calculated weights. In this case if no observations meet this condition, zero values are considered as the background data to estimate the missing features.

Finally, the instance-based background distribution is used in a Tree explainer. We note that we used the interventional (a.k.a., non-conditional) version of SHAP (default setting). Imagine we have 10 samples, after finding the shapley values on the explanation set for each of ten samples, we obtain 100 SHAP values for each variable in a single observation. To present the superiority of our proposed InstanceSHAP to an ordinary SHAP method in which whole train data is considered as the background distribution, we also find shapley values for each variable in each observation for each subsample and finaly to compare the InstanceSHAP with the ordinary SHAP, we consider the fluctuation of SHAP values using the concentration measure, Gini index.

DRAFT

Table 2: Comparison of the explanation models

| Method | InstanceSHAP | OrdinarySHAP |
|---|---|---|
| Gini Index | 0.54 | 0.53 |

Empirically, the original data is a huge dataset and contains over 1 million rows. To decrease computational time, we sample 1000 observations from the original data. Then, we set 70% of the dataset for training, and use the remaining, 30%, for testing. According to our proposed algorithm, we find the most similar train observations to the test data based on two variables. Later, in the explanation step of the algorithm, we use the retrieved observations which have a more similar distribution to the explanation data (test data) as the background data in the Tree explainer. In addition, for the purpose of comparison and validation of InstanceSHAP, we find the ordinary Shapley values. Here, ordinary shapley values mean the contribution values that are found using the whole train data as the background data in the explainer. After finding shapley values in the considered sample we go to the first step and repeat all steps for another sample. we repeat the process 100 times. In each round of the calculation, we find the Gini index of shapley values given by both explanation methods we consider in our application, InstanceSHAP and OrdinarySHAP. After finding all 100 Gini indices, we find the mean of Gini for each method. The results are presented in Table 2.

It can be seen that the value of Gini in InstanceSHAP is slightly higher than that of for OrdinarySHAP that shows, even for a relatively simple Random Forest model as the one we employ.

We should mention that since the main focus of our paper is to improve explanations provided by Shapley values, we use a simple Random Forest model with the default setting on Python. Analyzing the effect of hyperparameter optimization and machine learning models on explanation methods could be a direction for future study, which will likely lead to an even higher superiority of our proposed model.

DRAFT

## 4. Conclusion

Through an empirical study, we have shown that SHAP explanations fluctuate when background data changes. In particular, we showed that providing a background data with more similar distribution to the test data leads to better explanations. Using Gini index and comparing the indices of our proposed InstanceSHAP and OrdinarySHAP (using train data as the background data), we observe higher values for InstanceSHAP, even with a simple model. In general, our results suggest that the robustness of explanations caused by the choice of the background dataset should not be ignored. Future work should include: (i) extending to background data selection; and (ii) robustness of explanations.

## References

[1] Aas, K., Jullum, M., & LÃ¸land, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502.

[2] Burkart, N., Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317.

[3] Covert, I., Lee, S. I. (2021, March). Improving KernelSHAP: Practical Shapley value estimation using linear regression. In International Conference on Artificial Intelligence and Statistics (pp. 3457-3465). PMLR.

[4] Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 249(2), 417-426.

[5] Janzing, D., Minorics, L., Blobaum, P. (2020, June). Feature relevance quantification in explainable AI: A causal problem. In International Conference on artificial intelligence and statistics (pp. 2907-2916). PMLR

[6] Kwon, Y., Zou, J. (2022). WeightedSHAP: analyzing and improving Shapley based feature attributions. arXiv preprint arXiv:2209.13429.

[7] Lundberg, S.M., & Lee, S. (2017). A unified approach to interpreting model predictions. In: Proceedings of NIPS2017, 4768-4777.
[8] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610.
[9] Merrick, L., Taly, A. (2020, August). The explanation game: Explaining machine learning models using shapley values. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 17-38). Springer, Cham.
[10] Ribeiro, M. T., Singh, S., Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

DRAFT

# Networks & Nature Based Solutions: an application for Milan hydric resources

Alessia Forciniti[a] and Emma Zavarrone[a]

[a]IULM University; alessia.forciniti@iulm.it, emma.zavarrone@iulm.it

**Abstract**

This paper proposes an innovative model to study the process of co-creation of stakeholders in the definition of *Nature-Based Solutions (NBS)*. Thematic groups of NBS with a focus on water well-being in Milan are identified. A random sample of oral interviews has been converted into textual corpus, combined text analytics and network analysis to detect latent thematic groups. The corpus has been divided into two gender-based sub-corpora to identify demo-diversity in building of sustainable innovation.The results revealed gender-based differences. Male stakeholders see NBS as a linear process based on circular economy and intermediation of the territorial authorities. Female stakeholders identify them as a progressive transformative process based on science, culture and citizen co-operation.

***Keywords:*** nature based solutions, natural language processing, network analysis

DRAFT

## 1. Introduction

In 2019, the European Commission acquired through an innovative research project the concept of *Nature-Based Solutions-NBS* such as replicable models of strategies, actions, and adaptive interventions in nature aimed at increasing the sustainability of urban systems and performing the recovery of degraded ecosystems.

The perspective is oriented to foster socio-economic well-being of cities, by increasing inclusivity, re-generation of the disadvantages urban areas characterised by dilapidated infrastructures, pollution, high unemployment rates and poverty (e.g., Frantzeskaki *et al.*, 2017), by reducing violence and social tensions. This transformative potential is based on research and innovation and it is addressed to both citizens and other stakeholders in a perspective at several levels: ecological balance, creation of cultural, social, and economic advantages. However, the framework concerning the development of NBS highlights operational and technical criticalities that see solutions and approaches incompatible with local communities (e.g.; European Commission, 2020). The literature suggests to solve this limitation through cooperative governance models (Zingraff-Hamed *et al.*, 2020) based on the citizen engagement in the planning of nature-based adaptations (e.g.; Wamsler, 2020). Some authors as Burton & Mustelin (2013) and Mees *et al.*, (2015) propose a model of 'inclusive urban regeneration' which takes shape through the transdisciplinarity of different stakeholders, from residents to local authorities, from companies to academics, for a better fairness, relevance, acceptance, and co-creation of solutions oriented to management of public assets. One of the current management of public resources concerns the ongoing climate crisis that impacts the hydrogeological cycle and reflects itself in the supply of drinking water, health. In Italy, in compliance with L.R. 15 March 2016, n. 4 and the R.R. 7/2017, have been introduced the principles of hydraulic and hydrological invariance for reducing the hydrological impact of the transformation activities of the territory, and different organisations have developed to support municipalities in sustainable urban drainage systems.

The objective of this study is to investigate the process of co-creation through the dialogue with local stakeholders, by detecting the key categories characterising the NBS on water which can improve the quality of life in the metropolitan city of Milan. We explored two research questions (RQs):
$RQ_1$: What are the thematic groups of NBS on water that propose the stakeholders in Milan?
$RQ_2$: Are there gender differences in the definition of thematic groups?

To answer these RQs, in this paper we present an innovative protocol which moves itself from a sample of oral interviews of stakeholders, to the identification of the latent topics on emerging needs and intervention areas to improve the hydrological systems in Milan. After we operated a conversion of video-audio interviews in a textual corpus, we combined text analytics to network analysis as our best performing approach for detecting subgroups of words interpreted as thematic groups of NBS. In addition, we split the corpus in two sub-corpora based on stakeholders gender for detecting demo-diversities in building of sustainable innovation.

In the following, section 2 introduces the data; section 3 describe the methods; section 4 presents the results; section 5 shows the conclusion and future developments.

## 2. Data collection

Data collection is a population of 101 oral interviews (57.42 % male and 42.57 % female) conducted on corporate managers engaged in ecological regeneration in Milan. The selection was based on four macro-categories of topics proposed by CAP Group, the main company - founded in 1928 - that manages the integrated water service of Milan and some municipalities in Monza and Brianza. The fundamental pillars proposed are: *a) innovation; b) ecological transition; c) people; d) environment*.

To answer $RQ_1$, we used the population of interviews, while to satisfy $RQ_2$ we filtered the interviews by gender, obtaining two sub-groups. The textual transcription of the all oral interviews presents 1,957 types, 3,866 tokens, and a lexical richness by the *type-token ratio (TTR)* equal to 50.62 %, suggesting a language little diversified in describing the sustainable well-being of 'water'. The TTR within each pillars (Figure 1) ranges between 81% and 83%, denoting a little greater variability about the environment.

Additional details on corpus and sub-corpora are available in Table 1.

DRAFT

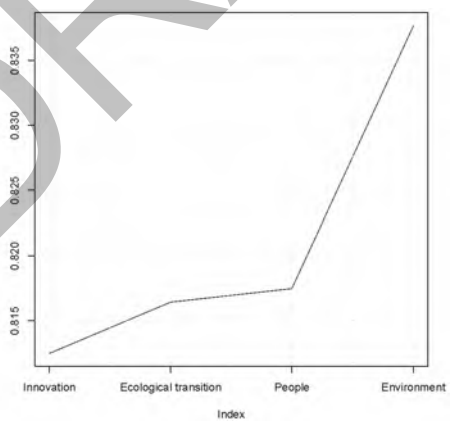



Figure 1: Lexical diversity plot

## 3. Methods

Our methodology is based on a 3-step model (Figure 2) which combines textual analysis to Network Analysis for detecting thematic groups. The model is based on a less conventional strategy compared to the most common methods (e.g., Hofmann, 1999; Blei *et al.*, 2003), but very effective in several studies (e.g.; Paolillo & Forciniti, 2021).

Table 1: Statistics on corpus and sub-corpora

| Textual collection | Types | Tokens | TTR | DTM[a] | Sparsity |
|---|---|---|---|---|---|
| Total corpus | 1,957 | 3,866 | 50.6% | $\mathbf{DTM}_{T101x915}$ | 72.85% |
| Male sub-corpora | 722 | 1,393 | 51.8% | $\mathbf{DTM}_{M58x329}$ | 66.30% |
| Female sub-corpora | 1,235 | 2,473 | 49.9% | $\mathbf{DTM}_{F43x586}$ | 57.29% |

[a]**DTM**: *Document Term Matrix or Lexical Table. For details concerning the interpretation of data in Table 1 refers to paragraph 3.*

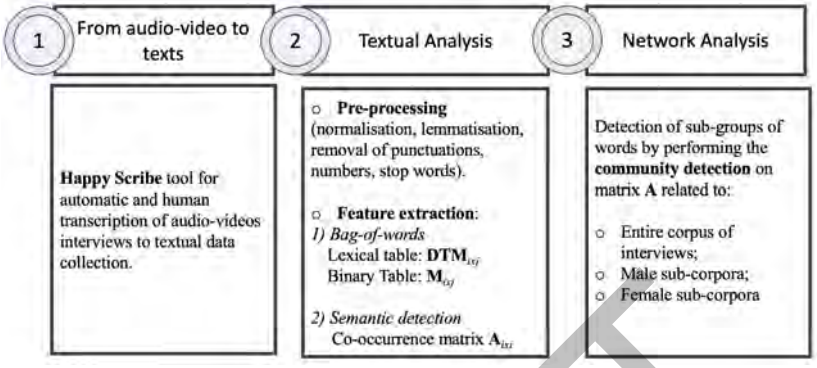



Figure 2: Flowchart of model

DRAFT

1. **STEP 1:** *Transformation of video-audios in textual data collection*
   We converted audio-video interviews to texts with 85-99% of accuracy using *Happy Scribe*, an automatic and human transcription tool for more than 120 languages.
2. **STEP 2:** *Textual Analysis*
   The corpus of the interviews was pre-treated through: a) normalisation; b) the removal of punctuation, numbers, stop word list for the Italian language based on standardised nomenclature of languages ISO 639 and c) lemmatisation. For feature extraction, the pre-processed texts were analysed at first by using a *bag-of-words* approach, and then by studying the relationship among the words to detect the semantic. We built the lexical table **DTM** (Document Term Matrix) based on a weighting scheme of frequency $n_{ij}$ which measures the occurrence of the term $i^{th}$ *(i = 1, ..., p)* in the document $j^{th}$ *(j = 1, ..., q)*. The **DTM** was converted into a binary matrix **M** where $m_{ij}$ is equal to 1 if the $i^{th}$ terms occurs in the $j^{th}$ document, and 0 otherwise.
   To keep tracks of each couple of terms side by side in a certain order, the matrix **M** was transformed into a *terms x terms* co-occurrence table **A**, multiplying $\mathbf{MM}^T$.
3. **STEP 3:** *Network Analysis*
   The matrix **A** may be interpreted such as an adjacency matrix (Wasserman & Faust, 1994) formally represented by a graph G (V, E) where V is a finite set of nodes of words and E a finite set of linkages (co-occurrences). To identify sub-groups of words densely interconnected to one another and poorly connected to other parts of the network (Newman & Girvan, 2004), we adopted the community detection. We used *fast-greedy* algorithm (Clauset *al.*, 2004), an agglomerative hierarchical clustering method for very large network, whose main advantage is the stopping criteria for choosing the number of groups based on optimisation of modularity *(Q)* (Newman & Girvan, 2004). The empirical range of its maximisation is between [0.3;0.7].

In our model, we performed the fast-greedy algorithm on matrices **A** related to both entire corpus of interviews and to gender sub-corpora in order to detect the thematic groups of NBS on water .

## 4. Results

To answer $RQ_1$, the findings on the entire corpus of interviews will be described to determine the thematic groups of NBS on water proposed by stakeholders in Milan. Later, to reply to $RQ_2$, the networks concerning the sub-corpora of male and female stakeholders will be showed in order to highlight gender differences in the definition of thematic groups.

Table 2: Summary of community detection

| Textual collection | Co-occurrence Matrix **A** | Modularity $Q$ |
|---|---|---|
| Total corpus | $\mathbf{A}_{T915x915}$ | 0.71 |
| Male sub-corpora | $\mathbf{A}_{M329x329}$ | 0.40 |
| Female sub-corpora | $\mathbf{A}_{F586x586}$ | 0.34 |

### 4.1 $RQ_1$: thematic groups of NBS detected on the entire corpus

For a better graphic intelligibility, we present the first four communities of NBS obtained on the most frequent 75 terms (Figure 3).

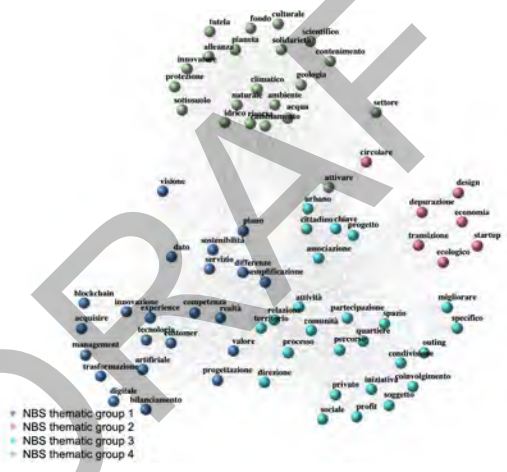



Figure 3: Thematic groups of entire corpus ($RQ_1$)

The first thematic group (blue cluster) is based on simplification of the processes, where innovation and digitalisation have key role. The second group (pink cluster) sees the overcoming of the linear economies through the ecological transition of start-ups, sewage treatment plants.The third group (turquoise cluster) is oriented to sustainable design and cooperation. The last group (green group) is the most specific of network, in which water, safety and the climate change emerge. Reading from bottom to top, we see the shift from the transformation of systems to the value of sharing and geological issues.

### 4.2 $RQ_2$: gender differences in the definition of thematic groups

The Figure 4 shows the network of male sub-corpora through two communities detected on the most frequent 50 terms.

The first group (turquoise cluster) describes the hydric resource with references to pollution of plastic, to depuration systems and circular economy. The second group (green cluster) is based on mediation of the local authorities (municipalities or other stakeholders).

DRAFT

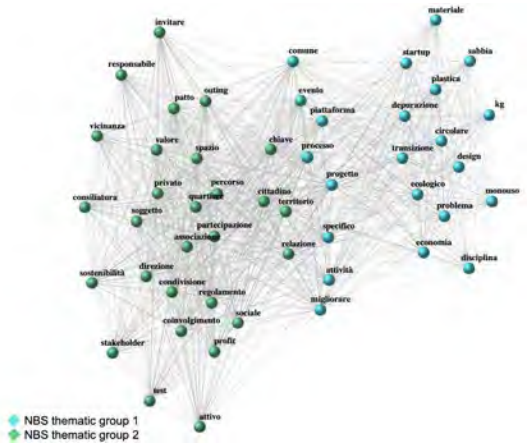



Figure 4: Thematic groups of male sub-corpora ($RQ_2$)

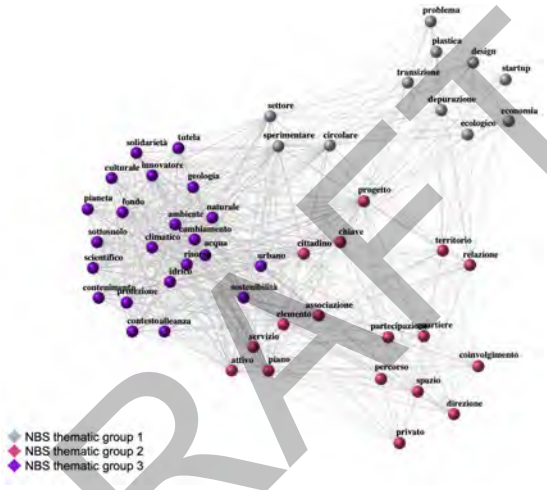



Figure 5: Thematic groups of female sub-corpora ($RQ_2$)

DRAFT

The Figure 5 shows the network of female sub-corpora and three communities detected on the most frequent 50 terms.

The first group (grey cluster) is a generalist view of the problem showing pollution, plastics and sewage systems. The second group (pink cluster) interprets the NBS as path of innovation of services, in contrast to the male perspective based on a transferring of the problem to the local authorities. The last group (purple cluster) is based on climate change, urban protection, science, culture and most of all highlights the gender difference.

## 5. Conclusion and future developments

In this paper, we combined text analytics and network analysis to detect thematic patters related to NBS on the water, by analysing the interviews of stakeholders which operates in the Milan city. At first, we studied the total perspective of interviewed, and at later we focused on gender-based sub-groups of interviews in order to detect potential gender differences in planning of strategies for urban systems development. The first investigation showed thematic groups based on: a) the simplification and digitalisation of systems; b) recycling-based economies; c) sewage treatment plants; (RQ1). However, by looking at the based-gender perceptions, it emerges that male stakeholders see NBS as a linear process

that moves from circular economy to the intermediation the territorial authorities as bridge. While, female stakeholders see NBS as a path of progressive transformation where the citizen take a key role and where are important welfare, solidarity, science and culture (RQ2).

Future developments might investigate additional demo-diversity features by means of techniques oriented to emotion and tone of voice recognition for better determining a diversification in defining of NBS on sustainability.

## References


[1] Blei, M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research, **31**, pp. 993-1022 (2003)

[2] Burton, P.; Mustelin, J.: Planning for Climate Change: Is Greater Public Participation the Key to Success? Urban Policy Res. **31**, pp. 399-415 (2013)

[3] Clauset, A., Newman,M. E. J., Moore, C. : Finding community structure in very large networks. hys. Rev. E **70** (2004)

[4] European Commission: Nature-based solutions (2019). Available via DIALOG. https://web.archive.org/web/20190923161801/http://ec.europa.eu/research/environment/index.cfm?pg=nbs

[5] European Commission, Directorate-General for Research and Innovation: Nature-based solutions towards sustainable communities: analysis of EU-funded projects. Publications Office of the European Union (2020). https://data.europa.eu/doi/10.2777/877034

[6] Frantzeskaki, N., Borgstrom S., Gorissen, L., Egermann, M., Ehnert, F.: Nature-Based Solutions Accelerating Urban Sustainability Transitions in Cities: Lessons from Dresden, Genk and Stockholm Cities. In: Kabisch, N., Korn, H., Stadler, J., Bonn, A. (eds.) Nature-Based Solutions to Climate Change Adaptation in Urban Areas. Theory and Practice of Urban Sustainability Transitions. Springer, Cham (2017)

[7] Hofmann, T.: Probabilistic Latent Semantic Analysis. Proceedings of the XV Conference on Uncertainty in Artificial Intelligence (UAI1999) (1999)

[8] Mees, H.L.P., Driessen, P.P.J., Runhaar, H.A.C.: Cool governance of a 'hot' climate issue: Public and private responsibilities for the protection of vulnerable citizens against extreme heat. Reg. Environ. Chang.**15**, pp. 1065 - 1079 (2015)

[9] Newman, M. E. & Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E, **69**, (2004)

[10] Paolillo, M. & Forciniti, A.: L'impatto del Covid-19 sull'opinione pubblica: una strategia di analisi per lo studio della comunicazione su Twitter. In: Favretto, A., Maturo, A., Tomelleri, S.(eds.) L'impatto sociale del Covid-19, pp.310-318 (2021). Milano: Franco Angeli

[11] Wamsler, C., Alkan-Olsson, J., Bjorn, H., Falck, H. *et al.*: Beyond participation: When citizen engagement leads to undesirable outcomes for nature-based solutions and climate change adaptation. Clim. Chang **158**, pp. 235-254 (2020)

[12] Wasserman, S. & Faust, K.: Social Network Analysis. Cambridge University Press (1994).


DRAFT

# The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization

Maria Gabriella Grassia[a], Marina Marino[a], Rocco Mazza[b], Agostino Stavolo[a]

[a] Department of Social Sciences, University of Naples "Federico II", Naples, Italy;
mariagabriella.grassia@unina.it; marina.marino@unina.it;
agostino.stavolo@unina.it
[b] Department of Political Science, University of Bari "Aldo Moro", Bari, Italy;
rocco.mazza@uniba.it

**Abstract**

In recent years, social media has become the main field for sourcing textual data. In particular, microblogging platforms such as Twitter make it possible to study users' online discussions by understanding and analysing the opinions, comments, and experiences that users share on different issues. One of the most debated issues in recent years is the voluntary termination of pregnancy. In particular, the United States Supreme Court's Roe v. Wade ruling has brought the abortion debate back to the forefront. Indeed, after the annulment of the ruling that restricted the right to abortion in the U.S., the response of online users has been crucial. Indeed, the aim of the paper is to understand what the major topics of discussion have been in the wake of the ruling. To do this, semantic clusters of terms were created through a symmetrical matrix reduction technique, Symmetric Non-Negative Matrix factorization.

**Key words:** Twitter, Symmetric non-negative matrix factorization, Lexical matrix decomposition

DRAFT

## 1. Introduction

On 24th June 2022, the U.S. Supreme Court overturned the sentence Roe v. Wade, which established the constitutional right to abortion in the United States in 1973. The Republican-appointed justices voted to strike down the federal right, while the other Democratic justices voted against it.

With the annulment of the ruling, individual U.S. states now have the power to establish their own laws regarding the right, or not, of a woman to have a termination of pregnancy, which Roe v. Wade allowed by the 24th week. In the absence of a law from Congress regulating abortion at the federal level, it will mean that each state can decide whether to allow abortions, whether to ban them always or under certain circumstances.

This event produced legal-normative consequences regarding whether or not voluntary termination could be accessed, but also social consequences in that it opened up a wide debate about the freedom to choose and to protect women's bodies. In addition, people seeking abortions will have to move from the city where they live and travel to more distant clinics or hospitals that allow the practice.

The outcome of the ruling has created a wide online debate among citizens who have expressed positions for or against the incident. This is because social networks and microblogging platforms are used as tools of political exchange: the debate on microblogging platforms has been studied before (Graells-Garrido et al. 2020; Sharma et al., 2017; Zang et al. 2016). Data from social media platforms such as Twitter offer new possibilities to study these dynamics because it provides a public arena for information gathering and opinion formation (Grassia et al. 2022).

So, the following work contributes to the study of the online debate on the issue of abortion through the analysis of tweets posted by users following the ruling. To understand the issues that users focused on in the online discussion, we used Symmetric Non-Negative matrix factorization (symNMF), which is a symmetric matrix reduction method used in clustering operations to create semantic clusters of terms.

## 2. Symmetric Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) is an unsupervised matrix decomposition method that decomposes a matrix $\mathbf{X} \in R^{n \times m}$ by a product of two factor matrices $\mathbf{X} \approx \mathbf{WH}$, where $\mathbf{W} \in R^{n \times k}$ and $\mathbf{H} \in R^{k \times m}$, both with nonnegative elements. The product WH is an approximate factorization of rank at most $k$ that it is assumed to satisfy the condition $k \ll min\ \{m,n\}$ (Gaujoux and Seoighe, 2010). The value of the parameter $k$ shows the numbers of factors to be used to explain data (Casalino et al. 2016).

The NMF has the advantage that provides easily interpretable results because the extracted latent features are parts of the original data. Recently, NMF is widely used in text mining and document clustering but also for topic modeling in short texts. However, NMF is not a general clustering method that can be applied in every condition, as it is dependent on the linear or non-linear structure of clusters. (Kuang et al. 2015).

To solve the problem, it is used the symmetric version of NMF, the Symmetric Non-Negative Matrix Factorization (symNMF). This method factorizes a symmetric matrix input $\mathbf{A}$ by the product of two matrices $\mathbf{A} \approx \mathbf{HH^T}$, where $\mathbf{H}$ is the cluster assignment and its transpose $\mathbf{H^T}$ (Jia et al. 2021). Specifically, the matrix $\mathbf{H}$ is a nonnegative matrix of size $n \times k$, and $k$ is the number of clusters requested.

Suppose the data points of the same group have high similarity values and the data points of the different groups have weak similarity values. So, a better approximation of $\mathbf{A}$ defines the cluster structure because the largest entry in the i-th row of $\mathbf{H}$ indicates the clustering assignment of the i-th data point, according to the nonnegativity of $\mathbf{H}$. The usual approach for approximating the input matrix $\mathbf{A}$ is to minimize the Frobenius norm $min_{H \geq 0} ||A \approx HH^T||^2_F$, and it can be related to a generalized form of many clustering objectives (Kuang et al. 2012).

SymNMF has been shown to be more effective for nonlinearly separable data than NMF (Kuang et al. 2015). It is important to know that symNMF is equal to spectral clustering, but Vangara et al. (2021) showed that it assumes better performance than k-means and spectral clustering. Kabir et al. (2020) stated that symNMF is more functional for clustering because it converts spectral clustering into an optimization problem with stationary point solutions. It has also been used as a graph clustering method (Luo et al.2021) and for topic extraction from a lexical matrix (Grassia et al. 2022; Yan et al., 2013)

## 3. Methodology

According to the open access academy API, we extracted English-language tweets using #RoeVsWade in the week following the outcome of the ruling (June 24-30, 2022). The volume of data extracted was very large: 943,696 tweets were extracted. In this regard, we eliminated retweets and identified 214,469 documents.

As part of an information mining process, it is necessary to process texts and obtain a set of structured data that can be elaborated using statistical techniques. So, several text pre-processing operations were carried out to transform the textual data (i.e., tweets) into structured data.

Documents are parsed and tokenized, resulting in a set of distinct strings (*tokens*) separated by blanks, punctuation marks, or other types of special characters (e.g., hashtags). These tokens correspond to the terms used in the vocabulary. The particular scheme achieved by tokenization is commonly known as bag-of-words (BoW), as it treats each document as a multiset of its tokens, without regard to grammatical and syntactic roles.

Once documents have been atomized into their basic components, pre-processing is necessary to reduce linguistic variability (Uysal and Gunal, 2014). First, all characters of the terms were changed to lowercase. To account for language variety, the following were carried out other *normalization* operations, such as correcting misspelled terms or removing numbers (Misuraca and Spano, 2018).

To reduce morphological variability, *lemmatization* was carried out, where each term was returned to its canonical form (verbs are returned to the present infinitive, nouns, and adjectives to the masculine singular). When texts have been pre-processed, it is possible to construct the so-called *vocabulary* by stacking identical terms and counting the number of occurrences of each vocabulary (type) in the document collection. To avoid uninformative terms, the vocabulary can be trimmed by removing so-called *stop-words*, i.e., the common terms used in the specific language and domain analysed (prepositions, conjunctions, etc.). For the same reason, rare terms with a low number of occurrences are usually removed from the vocabulary.

DRAFT

These phases returned a database composed of 2.819,642 tokens, 63,150 types and 214,469 documents. In the final stage of the pre-processing process, we applied the matrix vector space model of documents and words. Each document can be seen as a vector in *p*-dimensional vector space spanned by the terms belonging to the vocabulary. We created a term-document matrix, where *term frequency* is used to express the relative importance of each term in each document. Raw frequency weights are calculated as the number of occurrences of a term in a document and correspond to the absolute frequency. This system of weights results in the creation of a sparse matrix, which, in the case of NMF, leads to numerous problems with the interpretability of the result.

According to Yan et al. (2013), to reduce the sparsity of the term-document matrix, we transform it into a *co-occurrence* matrix, that represents the number of times two terms $w_i$ and $w_j$ co-occur together. For each pair of vectors, we define the *cosine* similarity measure that expresses the association of terms. Cosine similarity measures the similarity between two vectors of an inner product space. It is defined by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. Thus, we created a similarity matrix, and we applied the symNMF to the similarity matrix for defining semantic clusters of terms.

## 4. Preliminary results

Specifically, we define four semantic clusters.

Table 1: Semantic clusters by SymNMF

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|
| Odd | Trample | Hypocrisy | Healthy |
| Prevent | Witness | Demand | Daughter |
| Process | Terrible | Duty | Disappoint |
| Progress | Upheld | Moral | Embarrassment |
| Miscarriage | Trash | Legalize | Disregard |
| Oppose | Violation | Fail | Harm |
| Overturn | Stripping | Democracy | Freedom |
| Practice | Norm | Jesus | Mother |
| Respect | Rights | Life | Children |
| Vote | Amendment | Church | Violence |

DRAFT

From the analysis of the clusters extracted from the matrix, we noticed how there was a strong aversion to the outcome of the judgment. Table 1 shows the top ten terms associated with the clusters created.

We can see that the first cluster focuses on the narrative of the ruling and the reversal of the right to abortion. The Supreme Court overturned the ruling legalizing abortion, reaffirming that authority now reverts to the people and state representatives. All of this led to a strong outrage from Twitter users against the ruling, highlighting how the result will lead to negative consequences for the right to abortion. Through the overturning of the ruling numerous women's rights are going to be violated, primarily the freedom to choose one's own body and choose whether to continue the pregnancy. This is also interpreted in the second cluster, which highlights a number of terms such as 'trample', 'trash' and 'terrible' which indicate a strong aversion of online users to the decision taken.

Added to this are the users' criticism of the Catholic world and the church, pointing out how the church is hypocritical in not recognizing individuals who decide to have an abortion full legality. Abortion presents a profound moral issue, and the church and bishops have called the incident a "historic day" as defenders of the so-called "pro-life" movements. The rationale behind pro-life movements is to equate the gynaecological practice with the killing of an embryo, which acquires the status of a person at the moment of conception, and to consider abortion a highly damaging intervention for health.

The latest cluster created, however, shows parents' concern for their children's future. Users wonder what will happen after the ruling and hope that their children will be able to grow up in a state that

guarantees more rights than they had. And it is especially the issue of violence that creates discontent, as people are frightened and perplexed about the future of girls should they experience violence and have difficulty accessing abortion practices.

## 5. Conclusions

Nowadays, short texts have become an interesting form of text information, such as social media posts, question titles, and comments on posts. Short texts from the Internet are often extremely short, noisy, and ambiguous, imposing great challenges to clustering. The biggest difficulty is that each short text only holds very few word tokens.

This preliminary work aims to show how symNMF is an optimal technique for creating term clusters. About what has been analysed, it has been shown how Twitter users had a strong reaction to the overturning of the U.S. Roe v. Wade ruling by expressing strong criticism of the Supreme Court's choice.

For future developments, we are improving the level of efficiency in the construction of clusters by automatically defining the number of $k$ clusters to be detected and adopting of specific metrics that effectively measure the quality of clusters.

## References


[1] Casalino, G., Del Buono, N., Mencar, C.: Non-negative matrix factorizations for intelligent data analysis. Non-negative Matrix Factorization Techniques (pp. 49-74), Springer, Berlin, (2016).

[2] Gaujoux, R., Seoighe, C.: A flexible R package for nonnegative matrix factorization. BMC bioinformatics, 11(1), 1-9, (2010).

[3] Graells-Garrido, E., Baeza-Yates, R., Lalmas, M.: Representativeness of abortion legislation debate on twitter: A case study in Argentina and Chile. In Companion Proceedings of the Web Conference 2020 pp. 765-774, (2020).

[4] Grassia M.G., Marino M., Mazza R., Stavolo A.: Analysis of the public debate on DDL Zan on Twitter: an application of the Structural Topic Model in Proceedings of the 16th International Conference on Statistical Analysis of Textual Data, Vol. 1, pp. 67-73, (2022).

[5] Grassia M.G., Marino M., Mazza R., Misuraca M., Stavolo A.,: Topic modeling for analyzing the Russian propaganda in the conflict with Ukraine in Book of Short Papers of the ASA Conference 2022 - Data-Drive Decision Making, *In press*, (2022)

[6] Jia, Y., Liu, H., Hou, J., Kwong, S., Zhang, Q.: Self-supervised symmetric nonnegative matrix factorization. IEEE Transactions on Circuits and Systems for Video Technology (2021)

[7] Kuang, D., Ding, C., Park, H.: Symmetric nonnegative matrix factorization for graph clustering. In Proceedings of the 2012 SIAM international conference on data mining (pp. 106-117). Society for Industrial and Applied Mathematics, (2012).

[8] Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering, in Partitional Clustering Algorithms pp. 215-243. Springer, Cham (2015).

[9] Luo, X., Liu, Z., Jin, L., Zhou, Y., Zhou, M.: Symmetric nonnegative matrix factorization-based community detection models and their convergence analysis, IEEE Transactions on Neural Networks and Learning Systems (2021)

[10] Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In International Conference on the Statistical Analysis of Textual Data, pp. 17-28. Springer, (2018).

[11] Sharma, E., Saha, K., Ernala, S. K., Ghoshal, S., De Choudhury, M.: Analyzing ideological discourse on social media: A case study of the abortion debate. In Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas, (2017)

[12] Uysal A, Gunal S: The impact of preprocessing on text classification. Information Processing and Management 50(1):104–112, (2014)

[13] Vangara, R., Rasmussen, K. Ø., Chennupati, G., Alexandrov, B.: Determination of the number of clusters by symmetric non-negative matrix factorization. In Big Data III: Learning, Analytics, and Applications Vol. 11730, pp. 104-113. SPIE, (2021)

[14] Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y. Learning topics in short texts by non-negative matrix factorization on term correlation matrix, in Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 749-757, Society for Industrial and Applied Mathematics, (2013).

[15] Zhang, A. X., Counts, S.: Gender and ideology in the spread of anti-abortion policy. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems pp. 3378-3389, (2016).


DRAFT

# A comparison of different techniques for handling missing covariate values in propensity score methods

Anna Zanovello[a], Alessandra R. Brazzale[b], and Omar Paccagnella[b]

[a]Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua;
anna.zanovello.1@studenti.unipd.it
[b]Department of Statistical Sciences, University of Padua;
alessandra.brazzale@unipd.it, omar.paccagnella@unipd.it

**Abstract**

In observational studies, confounders may bias the estimates of the causal effects. Propensity score adjustment is commonly used to correct these estimates for such bias. The presence of missing data is another problem that often characterises the statistical analysis. The aim of our contribution is to discuss the performance of the four main methods for dealing with confounding based on the propensity score when combined with different techniques for imputing missing values according to the Multiple Imputation approach. The discussion will be based on the insight achieved from an extensive simulation study, which embraces multiple scenarios, on the estimation of the ATT.

***Keywords:*** ATT, confounding, imputation, missing data, propensity score

DRAFT

## 1. Motivation and rationale

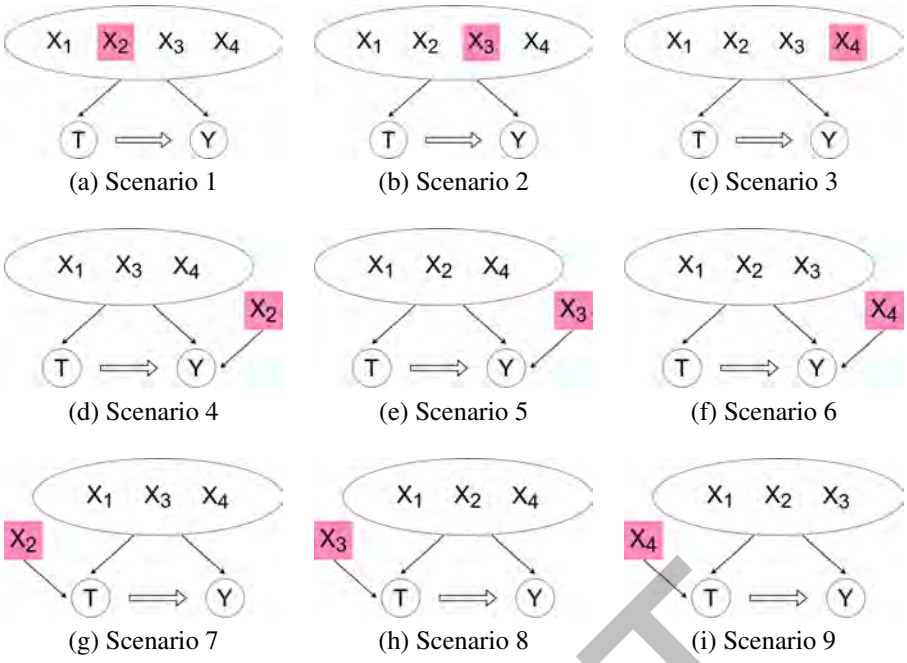
When observational or non randomised data are investigated, propensity score methods are powerful solutions to estimate causal treatment effects, as they minimise the influence of observed confounding variables (6). While this approach has been widely investigated in the literature, limited attention has been paid to the role of missing values in the covariates used to compute the propensity scores (3; 5).

By means of an extended and diversified simulation study, we aim at comparing the performance of the four main methods based on the propensity score to treat confounding, while taking different techniques of imputation of missing data into account. The motivation which inspired our work was suggested by an observational study on smokers who were discharged alive after being hospitalized with a diagnosis of acute myocardial infarction: the outcome variable is mortality within three years of hospital discharge, while the treatment was receipt of smoking cessation counseling prior to hospital discharge (1).

## 2. The simulation design

Nine different scenarios were investigated, summarised in Fig. 1. The following assumptions were made on the four covariates $X = (X_1; X_2; X_3; X_4)$:

- $X_1$ and $X_2$ are continuous, simulated from two independent normal distributions;
- $X_3$ is dichotomous, created from a Bernoulli distribution;
- $X_4$ is ordinal, assuming values $(1, 2, 3, 4)$ with probabilities $(0.3, 0.2, 0.4, 0.1)$, respectively.




(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

(d) Scenario 4 (e) Scenario 5 (f) Scenario 6

(g) Scenario 7 (h) Scenario 8 (i) Scenario 9

Figure 1: Simulation scheme: $Y$ binary response, $T$ binary treatment, $X_1$ and $X_2$ continuous covariate, $X_3$ binary covariate, $X_4$ discrete covariate. In pink: covariate with missing values.

DRAFT

### Treatment assignment

Let $T$ be the dichotomous variable which assumes value 1 for those who received the treatment with probability $\pi_T$. Values of $T$ were generated according to the following four different working assumptions:

**Assumption A:** Linear
$\text{logit}(\pi_T) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4$
**Assumption B:** Quadratic with no interaction term
$\text{logit}(\pi_T) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_2^2 + \alpha_6 X_4^2$
**Assumption C:** Linear with interaction terms
$\text{logit}(\pi_T) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_7 X_1 X_3 + \alpha_8 X_2 X_4$
**Assumption D:** Quadratic with interaction terms
$\text{logit}(\pi_T) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_2^2 + \alpha_6 X_4^2 + \alpha_7 X_1 X_3 + \alpha_8 X_2 X_4$

with $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8) = (0.8, -0.25, 0.6, 0.4, 0.05, -0.25, 0.4, -0.15)$ in scenarios 1–3 and 7–9, while in scenarios 4–6 some parameters had to be assumed equal to 0:

- Scenario 4: $\alpha_2 = \alpha_5 = \alpha_8 = 0$
- Scenario 5: $\alpha_3 = \alpha_7 = 0$
- Scenario 6: $\alpha_4 = \alpha_6 = \alpha_8 = 0$.

### Response variable

Let $Y$ be a dichotomous outcome variable which assumes the value 1 for those who experience the event under investigation with probability $\pi_Y$. We simulated the outcomes according to

$$\text{logit}(\pi_Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \gamma T,$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (-1, 0.3, -0.25, 0.35, 0.5)$ and $\gamma = -0.7$ in scenarios 1–6. Similar to the simulation of treatment, the following assumptions had to be made to some parameters in the remaining scenarios:

- Scenario 7: $\beta 2 = 0$;
- Scenario 8: $\beta 3 = 0$;
- Scenario 9: $\beta 4 = 0$.

### Missing values

We assume that only one explanatory variable at a time includes missing values, as illustrated by Fig. 1. For each condition, 50 fake datasets of sample size equal to 1,000 were generated.

## 3. Handling missing data

For each of the nine scenarios considered by our simulation study, the proportion of missing values was fixed at 40% of the sample size. Assuming a Missing-At-Random (MAR) mechanism, missing data were generated according to the amputation procedure proposed by (8). The amputation was repeated 10 times for each simulated dataset to take into account the variability inherent the procedure itself.

Following Rubin's (1987) methodology (7), Multiple Imputation (MI) was adopted. Each imputation was repeated 5 times for each simulated dataset, for a total of 500 fake datasets. Missing data were imputed following a MICE (Multiple Imputation by Chained Equations) approach by three different solutions: i) Predictive Mean Matching (PMM); ii) Weighted Predictive Mean Matching (WPMM); iii) regression imputation (REGRESS). According to the type of the variable with missing values, that is continuous, binary or discrete, a Bayesian linear model, a Bayesian logistic model or a proportional odds logistic model was adopted for the regression imputation solution, respectively. See (9) for further details on these procedures and their implementation in numerical computing environment R. All fully observed covariates were involved in each imputation model, while the outcome and treatment variables were excluded.

The single imputation approach, based on the unconditional mean imputation (MEAN) solution, was included in our analysis as a further comparison.

DRAFT

## 4. Propensity score methods

The individual propensity scores were estimated through a logistic regression model which includes all covariates. The Average Treatment Effect on Treated (ATT) was estimated using the four main different propensity score approaches (2): i) matching (to the nearest neighbour); ii) stratification (based on five strata obtained according to the quintiles of the estimated propensity scores); iii) covariate adjustment; iv) Inverse Probability of Treatment Weighting (IPTW) (according to Imbens' estimator (4)).

The propensity score methods were also benchmarked on the complete case analysis (CCA), that is by analysing the dataset where missing values are just deleted and not replaced.

## 5. Main results

### Performance metrics

The performance of the different estimators were compared by means of their relative bias and mean squared error. For every combination of scenario, assumption, strategy used to handle missing data and propensity score adjustment method, let $\text{ATT}_r$ be the true values of the ATT and $\widehat{\text{ATT}}_{r,i}$ be the estimates of the ATT for the $r = 1, \ldots, 50$ fake datasets and their $i = 1, \ldots, 10$ amputations. The average relative bias is defined as

$$\overline{\text{Bias}} = \frac{1}{50} \sum_{r=1}^{50} \left( \frac{\frac{1}{10} \sum_{i=1}^{10} \widehat{\text{ATT}}_{r,i} - \text{ATT}_r}{\text{ATT}_r} \right),$$

Table 1: Best and worst combinations of PS adjustment methods and missing data imputation techniques for Assumption A (linear relationship), by simulation scenario.

| Incomplete variable | | continuous | | dichotomous | | ordinal | |
|---|---|---|---|---|---|---|---|
| | | best | worst | best | worst | best | worst |
| confounding | $\overline{\text{Bias}}$ | IPTW<br>CCA | Stratification<br>MEAN | Cov. adjustment<br>MI | Matching<br>MEAN | IPTW<br>CCA | Stratification<br>MI |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>MEAN | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>MEAN | Matching<br>CCA |
| predictor of $Y$ | $\overline{\text{Bias}}$ | Cov. adjustment<br>MI | Stratification<br>CCA | IPTW<br>MI | Stratification<br>MI | Cov. adjustment<br>MI | Stratification<br>CCA |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA |
| predictor of $T$ | $\overline{\text{Bias}}$ | Matching<br>WPMM | Cov. Adjustment<br>CCA | Cov. Adjustment<br>CCA | Stratification<br>CCA | Matching<br>WPMM | Stratification<br>CCA |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA |

MI: Multiple Imputation (all methods). Imputation: all the multiple and single imputation methods

while the average mean squared error is

$$\overline{\text{EQM}} = \frac{1}{50}\sum_{r=1}^{50}\sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(\widehat{\text{ATT}}_{r,i} - \text{ATT}_r\right)^2}.$$

## Take-home messages

Generally, no specific combination worked markedly better or worse for all the nine investigated scenarios of Fig. 1. The largest differences were observed among the four propensity score adjustments rather than for the varying techniques to handle missing data. In terms of relative bias, it seems that stratification tends to produce the worst results, while covariate adjustment often outperforms the other three techniques. However, this cannot be generalised to all circumstances.

If we focus on the single simulation scenarios, lower performances in terms of bias are generally observed in the first three cases, that is when the missing data are present in a confounding variable. The effect of missing data in variables which are only predictors of treatment or of the response is null or minimal. Also the type of variable with missing values seems to have limited effects on the performance of the estimators. In particular, as far as the last six scenarios goes, that is, when the incomplete variable is either a predictor of treatment or of the response, there is almost no difference whether it is continuous, dichotomous or discrete.

Tables 1 and 2 highlight, for the two opposite Assumptions A ("additive and linear") and D ("quadratic with interaction terms"), the most efficient and the least performing combinations of propensity score adjustment and missing data imputation. In most cases, the best choice to treat confounding is covariate adjustment, while it is more difficult to provide general guidelines with respect to the treatment of missing data. According to relative bias, the best combination seems to be with complete case analysis when missing data are in confounding variables, and with multiple imputation in all other cases. However, if we focus on the mean squared error, multiple imputation shows better performances.

DRAFT

Table 2: Best and worst combinations of PS adjustment methods and missing data imputation techniques for Assumption D (quadratic with interaction terms' relationship), by simulation scenario.

| Incomplete variable | | continuous | | dichotomous | | ordinal | |
|---|---|---|---|---|---|---|---|
| | | best | worst | best | worst | best | worst |
| confounding | $\overline{\text{Bias}}$ | Cov. adjustment<br>CCA | Stratification<br>all methods | Cov. adjustment<br>CCA | Stratification<br>CCA | Cov. adjustment<br>CCA | IPTW<br>MI |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>MEAN | Matching<br>CCA |
| predictor of $Y$ | $\overline{\text{Bias}}$ | Cov. adjustment<br>MEAN | Stratification<br>CCA | Matching<br>WPMM | Stratification<br>CCA | IPTW<br>REGRESS | Stratification<br>CCA |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA |
| predictor of $T$ | $\overline{\text{Bias}}$ | Cov. adjustment<br>CCA | Stratification<br>MEAN | Stratification<br>MEAN | IPTW<br>MI | Cov. adjustment<br>MI | Stratification<br>CCA |
| | $\overline{\text{EQM}}$ | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA | Cov. adjustment<br>Imputation | Matching<br>CCA |

MI: Multiple Imputation (all methods). Imputation: all the multiple and single imputation methods

# References


[1] Austin, P.C.: A tutorial and case study in Propensity Score Analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behav. Res. **46**, 119–151 (2011)

[2] Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav. Res. **46**, 399–424 (2011)

[3] Choi, J., Dekkers, O.M. and le Cessie, S.: A comparison of different methods to handle missing data in the context of propensity score analysis. Eur. J. Epidemiol. **34**, 23–36 (2019)

[4] Imbens, G.W.: Nonparametric estimation of average treatment effects under exogeneity: A review. Rev. Econ. Stat. **86**, 4–29 (2004)

[5] Malla, L., Perera-Salazar, R., McFadden, E., Ogero, M., Stepniewska, K. and English, M.: Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review. J. Comp. Eff. Res. **7**, 271–279 (2018)

[6] Rosenbaum, P.R. and Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)

[7] Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York (1987)

[8] Schouten, R.M., Lugtig, P. and Vink, G.: Generating missing values for simulation purposes: a multivariate amputation procedure. J. Stat. Comput. Simul. **88**, 2909–2930 (2018)

[9] van Buuren, S. and Groothuis-Oudshoorn, K.: *mice*: Multivariate Imputation by Chained Equations in R. J. Stat. Soft. **45**, 1–67 (2011)


DRAFT

# A New Penalized Estimator for Sparse Inference in Gaussian Graphical Models: An Adaptive Non-Convex Approach

Daniele Cuntrera[a], Vito M.R. Muggeo[a], and Luigi Augugliaro[a]

[a]Universitá degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche;
daniele.cuntrera@unipa.it, vito.muggeo@unipa.it,

luigi.augugliaro@unipa.it

DRAFT

**Abstract**

A new penalized estimator for sparse inference in Gaussian Graphical Models is proposed in this paper. It is based on the adaptive non-convex penalty function first presented in (4). In comparison to other estimators based on non-convex penalty functions, such as SCAD and MCP, the proposed estimator has a number of advantages because it allows controlling the degree of the non-convexity of the objective function through a second tuning parameter, which eliminates the inferential issues associated with the existence of multiple local minima. A simulation study is used to assess the proposed estimator's performance.

***Keywords:*** Gaussian Graphical Models, high-dimensional data, non-convex penalty function, penalized inference, sparse inference

## 1. Introduction

Let $X$ be a $p$-dimensional random variable with joint distribution $p(x)$ and let $\mathscr{V} = \{1,\ldots,p\}$ be the associated vertex-set. A *probabilistic graphical model* can be defined as the triplet $(X, p(x), \mathscr{G})$, where $\mathscr{G} = \{\mathscr{V}, \mathscr{E}\}$ is a graph whose edge-set $\mathscr{E}$ encodes the conditional dependence and independence structure among the $p$ random variables, that is $X_h$ and $X_k$ are stochastically independent given the remaining random variables iff $(h,k) \notin \mathscr{E}$. For a comprehensive treatment of the probabilistic graphical models, the interested reader can refer, for example, to (9) and (11).

The Gaussian Graphical Models (GGM) are probabilistic graphical models based on the assumption that $X$ follows a $p$-dimensional Gaussian distribution:

$$f(x;\mu,\Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-1/2(x-\mu)^\top\Sigma^{-1}(x-\mu)\}, \tag{1}$$

where $\mu$ is the vector of expected values of $X$, whereas $\Sigma$ is the covariance matrix. The inverse of the covariance matrix, denoted by $\Theta = \Sigma^{-1}$, is called *precision matrix* and its off-diagonal elements, denoted as $\theta_{hk}$, are the parametric tools by which density (1) factorizes according to the undirected graph $\mathscr{G}$, formally:

$$(h,k) \notin \mathscr{E} \Leftrightarrow X_h \perp\!\!\!\perp X_k \mid X_{\mathscr{V}\setminus\{h,k\}} \Leftrightarrow \theta_{hk} = 0.$$

The interested reader is referred to (9) for a proof. The problem of finding a parsimonious representation of the model (1) was originally studied in (5), and it is known in the literature as *covariance selection problem*.

## 2. Sparse inference via adaptive non-convex penalty function

Suppose that a set of $n$ independent and identically distributed observations is drawn from the distribution (1) and denote the corresponding GGM by $\{X, f(x;\mu,\Theta),\mathscr{G}\}$, where $\mathscr{G}$ is the undirected graph whose edge set encodes the conditional independence structure among the $p$ random variables. In principle, inference on the factorization of the density (1), and consequently on $\mathscr{G}$, can be carried out by maximizing the following profile log-likelihood function:

$$\ell(\Theta) = \log\det\Theta - \mathrm{tr}(S\Theta),$$

where $S$ denotes the empirical covariance matrix. Then, the edge-set $\mathscr{E}$ the can be estimated by $\widehat{\mathscr{E}} = \{(h,k); \hat{\theta}_{hk} \text{ is significantly different from zero}\}$.

Although the procedure described above is theoretically well-founded, the application to real datasets is limited for two main reasons. Firstly, the number of measured variables is often larger than the sample size, implying the non-existence of the maximum likelihood estimator of the precision matrix. Secondly, the maximum likelihood estimator will exhibit very high variance even when the sample size is large enough. In terms of GGMs, this evidence translates into the assumption that $\Theta$ has a sparse structure; consequently, a number of authors have proposed a penalized approach to estimate $\Theta$ and $\mathscr{G}$ at the same time. We refer the interested reader to (1) for an extensive review of the more recent penalized methods proposed in the literature for GGMs.

Following the approach presented in (12), in this paper, we propose to estimate $\Theta$ and $\mathscr{G}$ using a penalized approach based on the usage of the non-convex penalty function introduced in (4). Formally, the proposed estimator is defined as follows:

$$\widehat{\Theta} = \arg\min_{\Theta\succ 0} -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P(|\theta_{hk}|/\nu_{hk}), \tag{2}$$

where $\ell(\cdot)$ is the log-likelihood and $P(|\theta_{hk}|/\nu_{hk}) = \nu_{hk}\int_0^{|\theta_{hk}|/\nu_{hk}} \exp\{-x^2/2\}\mathrm{d}x$. In the proposed estimator, $\rho > 0$ is a tuning parameter aimed to control the amount of sparsity in $\widehat{\Theta}$ and, consequently, in the corresponding estimated graph $\widehat{\mathscr{G}} = \{\mathscr{V}, \widehat{\mathscr{E}}\}$, where $\widehat{\mathscr{E}} = \{(h,k) : \hat{\theta}_{hk} \neq 0\}$. When $\rho$ is large enough, some $\hat{\theta}_{hk}$ are shrunken to zero resulting in the removal of the corresponding link in $\widehat{\mathscr{G}}$; on the other hand, when $\rho$ is equal to zero, and the sample size is large enough, $\widehat{\Theta}$ coincides with the maximum likelihood estimator of the concentration matrix, which implies a fully connected estimated graph.

The additional parameter $\nu$ allows the penalty function to be flexible enough to mimic the widely popular penalty functions proposed in the literature, both convex (e.g. LASSO (10)) and non-convex ones (e.g. SCAD (6), MCP (13)). For example, noting that

$$\rho \lim_{\nu\to\infty} P'(|\theta_{hk}|/\nu_{hk}) = \rho,$$

we can conclude immediately that, for large enough $\nu$-values, the results given by the proposed model are approximately equivalent to the one given by the graphical lasso (gLASSO) model (12).

Figure 1 shows the derivatives of the most well-known penalties and the proposed penalty in (4), for different values of $\nu$. Graphically it is possible to notice how it is a trade-off between the L1 and L0 penalty, taking (in the middle) the form of a non-convex penalty.

As can be guessed from Figure 1, $\nu$ determines the degree of nonconvexity, thus affecting the presence of multiple local solutions, especially in contexts where the number of covariates is huge ((3; 7)). For sufficiently small values of $\nu$ and $\rho$, the uniqueness of the solution to the problem (2) is not guaranteed. As discussed in (4), it is possible (and easy) to find, for every value of $\rho$, the lowest value of $\nu_{min,\rho}$ such that the solution is unique having the highest degree of non-convexity. It is important to emphasize that the value $\nu_{min}$ is not to be understood as the best value to use (in terms of, e.g., minimizing the prediction error committed) but as the lower-bound of the search space for the optimal value of $\nu$.

DRAFT

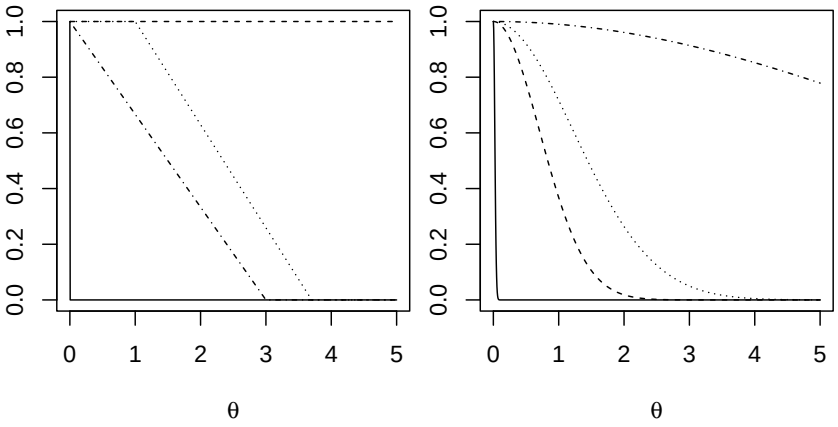



Figure 1: Plot of $P'(\cdot)$ functions over $\theta > 0$. Left panel: L0 (solid line), LASSO (dashed line), SCAD (dotted line, $\gamma = 3.7$) and MCP (dot-dashed line, $\gamma = 3$). Right panel ANP fixing $\nu$ at 0.001 (solid line), 1 (dashed line), 3 (dotted line) and 100 (dot-dashed line)

## 3. Computational Aspects: An efficient ADMM algorithm

In this section, we propose an efficient Alternating Direction Method of Multipliers (ADMM) methodologically grounded on the results given in (2). We begin redefining the solution of our minimization problem (2) as the solution of the following equality-constrained minimization problem, with matrix variables $\Theta$ and $Z$:

$$\begin{aligned} \min_{\Theta, Z \succ 0} \quad & -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right) \\ \text{s.t.} \quad & \Theta - Z = 0. \end{aligned}$$

According to the standard ADMM theory, the augmented scaled Lagrangian function takes the form:

$$\mathscr{L}(\Theta, Z, U) = -\ell(\Theta) + \rho \sum_{h,k=1}^{p} P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right) + \frac{\tau}{2}\|\Omega - Z + U\|_F^2 - \frac{\tau}{2}\|U\|_F^2,$$

where $\tau > 0$ is a penalty parameter, $U \succ 0$ is the scaled dual matrix and $\|\cdot\|_F$ denotes the Frobenius norm, respectively. Using $\mathscr{L}(\Theta, Z, U)$, the solution of the initial minimization problem (2) can be computed through the following procedure:

1: **repeat**
2: $\quad \Theta^{k+1} = \arg\min_{\Theta \succ 0} -\ell(\Theta) + \frac{\tau}{2}\|\Theta - Z^k + U^k\|_F^2$,
3: $\quad Z^{k+1} = \arg\min_{Z \succ 0} \frac{\tau}{2}\|\Theta^{k+1} - Z + U^k\|_F^2 + \rho \sum_{h,k}^{p} P(|Z_{hk}|/\nu_{hk})$,
4: $\quad U^{k+1} = U^k + \Theta^{k+1} - Z^{k+1}$
5: **until** convergence criterion is met

As we shall show, the main advantage of the ADMM algorithm consists of the ability to split the initial complex problem into a series of simpler problems, each of which is easy to solve. In the remaining part of this section, we shall study in greater detail the minimization problems defined in Step 2 and 3.

*Updating* $\Theta$. The problem in Step 2 has been studied in (2), where the authors show that the update of the precision matrix estimator admits the solution in closed form. To get more insight into the updating formula, consider the first-order optimality condition of the problem in Step 2, which can be rewritten in the following more convenient form:

$$\tau\Theta - \Theta^{-1} = \tau(Z^k - U^k) - S. \tag{3}$$

Let $Q\Lambda Q^\top$ be the spectral decomposition of $\tau(Z^k - U^k) - S$, then from the equation (3) we can immediately conclude that $\Theta^{k+1}$ can be written as $Q\widetilde{\Lambda}Q^\top$, where $\widetilde{\Lambda}$ is a diagonal matrix whose elements are the

DRAFT

solutions of the equation $\tau\widetilde{\Lambda} - \widetilde{\Lambda}^{-1} - \Lambda = 0$, that is:

$$\tilde{\lambda}_{ii} = \frac{\lambda_{ii} + \sqrt{\lambda_{ii}^2 + 4\tau}}{2\tau},$$

which are always positive since $\tau > 0$.

*Updating Z.* Before delving into the technical details related to the update of the matrix $Z$, we note that the objective function in Step 3 takes the following additive structure:

$$\frac{\tau}{2} \sum_{h,k=1}^{p} \left\{ (\theta_{hk}^{k+1} + U_{hk}^{k} - Z_{hk})^2 + \rho P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right) \right\},$$

which implies that the minimization problem in Step 3 can be split into $p(p+1)/2$ univariate optimization problems that can be solved in parallel. Therefore, in the remaining part of this section, we shall focus on how to solve the subproblem:

$$Z_{hk}^{k+1} = \arg\min_{Z_{hk}} \frac{\tau}{2}(\theta_{hk}^{k+1} + U_{hk}^{k} - Z_{hk})^2 + \rho P\left(\frac{|Z_{hk}|}{\nu_{hk}}\right).$$

Following the approach presented in (4), we solve the problem above using the local linear approximation (LLA) method (14), that is, $Z_{hk}^{k+1}$ is computed as solution of a sequence of new minimization problems involving a new objective function obtained replacing the penalty function with a suitable local approximation. Formally, $Z_{hk}^{k+1}$ is obtained by the following iterative procedure:

1: Let $\widetilde{Z}_{hk}^{k}$ be a starting value
2: **repeat**
3: Let $w_{hk} = \exp\{-(\widetilde{Z}_{hk}^{k}/\nu_{hk})^2/2\}$
4: $\widetilde{Z}_{hk}^{k+1} = \arg\min_{\widetilde{Z}_{hk}} \frac{1}{2}(\theta_{hk}^{k+1} + U_{hk}^{k} - \widetilde{Z}_{hk})^2 + \frac{\rho}{\tau} w_{hk}|\widetilde{Z}_{hk}|$
5: **until** convergence criterion is met
6: Return $Z_{hk}^{k+1} = \widetilde{Z}_{hk}^{k+1}$

In Step 4, we immediately recognize a weighted lasso problem; therefore, using the results given in (8), we have that the updating step of $\widetilde{Z}_{hk}^{k+1}$ admits the following solution in closed form:

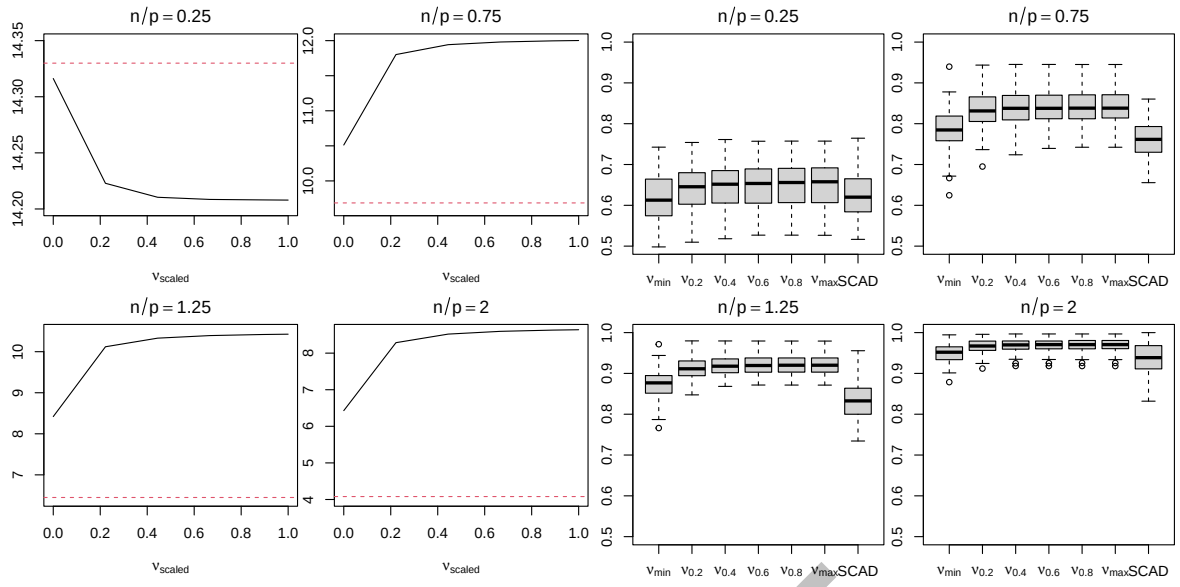
$$\widetilde{Z}_{hk}^{k+1} = S(\theta_{hk}^{k+1} + U_{hk}^{k}; \frac{\rho}{\tau} w_{hk}),$$

where $S(x;\lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator.

DRAFT

## 4. Simulation Study

We perform some simulation experiments to compare the behaviour of the estimator with that obtained using the SCAD penalty function. As discussed above, the gLASSO model is a limiting case of our proposal, i.e. as the tuning parameters $\nu$ go to infinity, the estimates given by the estimator (2) are asymptotically equivalent to those given by the gLASSO model. Therefore, in this study, we aimed to evaluate the effects of $\nu$ on the entire path of $\widehat{\Theta}$. Specifically, we used an evenly spaced sequence of six $\nu$ values, from $\nu_{min}$ to $\nu_{max} = 3$, where the largest $\nu$ value was chosen after a preliminary study to ensure that our estimates were approximately equal to those of GLASSO. To assess the effect of $\nu$ for different values of the ratio $n/p$, we set $p = 50$ and considered four different sample sizes, i.e. $n = (13, 38, 63, 100)$. To simulate a sparse precision matrix, we defined the edge set as a collection of $J$ disjoint edge sets encoding star structures. Formally, $\mathscr{E} = \cup_{j=1}^{J} \mathscr{E}_j$, where $\mathscr{E}_j = \{(j,k) : k = (j+1), \ldots (j+k_j)\}$. In our setting, $k_j = 24$ and, consequently, $J = 2$. The corresponding off-diagonal entries of $\Theta$ are simulated using a uniform distribution on the interval $(0.65, 1)$. In contrast, diagonal entries are computed to make




(a) Frobenius distances (b) AUC

Figure 2: Frobenius distances (left panel) and AUC (right panel) for SCAD (red dotted line in the left panel) and our proposal varying $\nu$.

the resulting precision matrix positively definite. For each possible combination of $\nu$ and $n$, we run 100 simulations.

The performance of the evaluated models is measured in terms of the accuracy of the graph structure recovery and the Frobenius distance between the estimated and the actual precision matrix. Regarding the first aspect, we summarise the coefficient paths using the AUC based on the ROC curves in each simulation run. Regarding the Frobenius distances, we take the minimum value of the total path of each estimator. We average the minima across replicates: in this way, we look at the best performance that all estimators can achieve.

Figure 2 (a) shows the curves of the Frobenius distances varying $\nu$ and the value obtained using the SCAD penalty. When the number of parameters is much larger than the number of observations, SCAD has the maximum average Frobenius distance. At the same time, it is observed that with our proposal, the distance decreases as $\nu$ increases (although the decrease is not large in absolute terms). As the ratio $n/p$ increases, it is observed that the lowest value of the distance to the true matrix $\Theta$ is obtained with the smallest value of $\nu$, i.e. $\nu_{min}$. The advantage with respect to the opposite limit case (i.e. gLASSO) increases as the ratio $n/p$ increases. Finally, by considering the capacity of the correct selection of non-zero coefficients (and thus the box plots of the AUCs), it is observed that our proposal always gives better results than SCAD. For different values of $\nu$, the AUC tends to improve as the value of $\nu$ increases, but the differences decrease as the ratio $n/p$ increases.

## 5. Conclusions

This paper presents a new penalized estimator for sparse inference in Gaussian Graphical Models. The estimator is based on an adaptive non-convex penalty function and can control the degree of non-convexity through a second tuning parameter $\nu$, avoiding inferential issues associated with multiple local minima. The simulation study results show the proposed estimator's improved performance compared to non-convex penalty functions like SCAD. From the evaluations based on the simulations study, it can be seen that it may be crucial to propose a method for selecting the $\nu$ parameter, which certainly depends on the $n/p$ ratio

DRAFT

**Acknowledgements.** Luigi Augugliaro and Vito M.R. Muggeo gratefully acknowledge financial support from the University of Palermo (FFR2021-22).

## References

[1] AUGUGLIARO, L., MINEO, A. M., AND WIT, E. C. *$\ell_1$-Penalized Methods in High-Dimensional Gaussian Markov Random Fields*. John Wiley & Sons, Ltd, 2016, ch. 8, pp. 201–265.

[2] BOYD, S., PARIKH, N., CHU, E., PELEATO, B., AND ECKSTEIN, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*, 1 (2011), 1–122.

[3] BREHENY, P., AND HUANG, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics 5*, 1 (2011), 232.

[4] CUNTRERA, D., MUGGEO, V. M. R., AND AUGUGLIARO, L. Variable selection with unbiased estimation: the CDF penalty. In *51th Scientific Meeting of the Italian Statistical Society: Book of Short Papers* (2022), pp. 1835–1840.

[5] DEMPSTER, A. Covariance selection. *Biometrics 28*, 1 (1972), 157–175.

[6] FAN, J., AND LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*, 456 (2001), 1348–1360.

[7] FAN, J., AND LV, J. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*, 8 (2011), 5467–5484.

[8] FRIEDMAN, J., HASTIE, T., HOEFLING, H., AND TIBSHIRANI, R. Pathwise coordinate optimization. *The Annals of Applied Statistics 2*, 1 (2007), 302–332.

[9] LAURITZEN, S. L. *Graphical Models*. Oxford University Press, Oxford, 1996.

[10] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[11] WHITTAKER, J. *Graphical Models in Applied Multivariate Statistics*. Wiley & Sons, Chichester, 1990.

[12] YUAN, M., AND LIN, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, 1 (2007), 19–35.

[13] ZHANG, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*, 2 (2010), 894–942.

[14] ZOU, H., AND LI, R. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*, 4 (2008), 1509–1533.

DRAFT

# A tool for assessing weak identifiability of statistical models

Antonio Di Noia[a,b], Francesco Denti[c], and Antonietta Mira[b,d]

[a]ETH Zurich, Zurich, Switzerland
[b]Università della Svizzera italiana, Lugano, Switzerland
[c]Università Cattolica del Sacro Cuore, Milan, Italy
[d]Università dell'Insubria, Varese, Italy

**Abstract**

The notion of identifiability has a long history in the statistical literature, with econometrics providing the first theoretical contributions. On the one hand, within the frequentist paradigm, identifiability represents a critical issue to tackle, closely tied to the feasibility of the model estimation. On the other hand, identifiability issues in the Bayesian framework could be overcome by complementing the non-identifiable likelihood with additional prior beliefs summarized via an informative prior distribution. Unfortunately, since estimation is still feasible, unidentifiabily may remain unnoticed and silently hinder posterior consistency. This contribution provides a tool to inspect whether the model specification is weakly identified. Our procedure is based on estimating the intrinsic dimension of posterior samples. The methodology is illustrated with a simulated example.

***Keywords:*** Identifiability, Intrinsic Dimension, Bayesian models, Monte Carlo Markov Chain.

DRAFT

## 1. Introduction

Identifiability plays a relevant role in the specification of statistical models. As the modeling complexity increases, identifiability issues are more likely to arise. Generally, a statistical model is identifiable in the standard parametric setting if the parameter values uniquely determine the data distribution and vice versa. In other words, given a model specification, it is possible to perform consistent inference on the parameters with the available data. Moreover, a model is *weakly identified* if the available data do not provide sufficient information to perform correct inferences, leading, for example, to very similar likelihood values in the maximum likelihood region, even if the one-to-one correspondence between likelihood and parameter values holds in that region.

Econometrics is an example of a research field in which identifiability plays a crucial role, originating a long research track on both its theoretical and applied sides. For example, the seminal paper [12] established a strong link between the information matrix and the identifiability of a statistical model under some regularity conditions on the data distribution. Moreover, the study introduced the notion of *local identifiability*, a form of identifiability restricted to open neighborhoods within the parameter space. Successively, [2] generalized this result to any distribution along with a very general criterion to study the identifiability of a specific parametric model based on the Kullback-Leibler divergence. Finally, the extension to non-parametric model specifications has been addressed in [11].

This branch of research has provided theoretically solid conditions for model identifiability. Unfortunately, checking these conditions on increasingly complex models is not immediate. Even worse, these

conditions are far from satisfied in modeling settings such as mixture, overparameterized, or spatial models. The problem of identifiability in mixture models is addressed in [5] and [10]. Other notable weakly identifiable models are latent variable models; see, for example, [9], and [1]. All these approaches are explicitly developed into a frequentist likelihood-based framework. The Bayesian approach could bypass the issue by the specification of informative priors. These priors can play the role of identifiability constraints, which are necessary to derive a well-defined posterior. However, the problem is not automatically solved but rather ignored. Because identifiability issues are not immediately apparent, there is the risk that, under the Bayesian framework, they are just overlooked.

Moreover, specifying informative priors to aid model estimation has consequences. Indeed, suppose we focus on frequentist properties of the Bayesian approach. In that case, if the prior is very informative, the posterior distribution of unidentified models could converge to a point mass that is not the true value. For this reason, highly informative priors would eventually lead to strongly biased inference.

In the Bayesian framework, weak identifiability could be present whenever the posterior distribution is "too" close to the prior according to some distributional. In particular, the marginal distributions of the non-identifiable parameters would not scale with the sample size and, consequently, will strongly impact the intrinsic dimension (id) of the posterior distribution support. We speculate that the id of the posterior distribution support could provide interesting insights into whether or not a specific set of parameters is identified, thus eventually representing an indication of some identifiability issues in the adopted model specification.

Although many definitions of id appeared in the literature, a fairly intuitive one is that the id is the dimension of the latent manifold on which the data-generating probability distribution has support. [3] is a comprehensive review of different approaches to id estimation. In this contribution, we focus on nearest-neighbor methods since they can effortlessly be embedded in a theoretically coherent modeling framework, allowing for uncertainty quantification of the estimates. As an example, [7] proposes a general and robust approach to id estimation based on ratios of distances between a unit and its nearest neighbors of arbitrary order. The intuition that motivates our proposal is based on the idea that weakly identified Bayesian models lack posterior consistency in the dimensions of the unidentified parameters.

The remainder of this contribution is organized as follows. In Section 2, we discuss the id estimation. Successively, in Section 3, we describe our methodology in more detail, providing some fundamental theoretical justifications and an example of application on a simulated scenario. We conclude the contribution with some final remarks in Section 4.

DRAFT

## 2. Intrinsic dimension estimation

Consider a dataset $\{\boldsymbol{x}_i\}_{i=1}^n$ with $n$ observations and $D$ observed features. Moreover, assume it is a realization from a Poisson process with a locally constant (or alternatively, locally homogeneous) intensity function. Define $r_{i,l}$ as the distance between unit $i$ and its $l$-th nearest neighbor. Then, let $n_1, n_2$ be two integers with $n_2 > n_1$ and define the ratio $\mu_{i,n_1,n_2} = r_{i,n_2}/r_{i,n_1}$. In [7], the authors introduced the Gride model, proving that $\mu_{i,n_1,n_2}$ has density

$$f\mu_{i,n_1,n_2}(\mu) = \frac{d(\mu^d - 1)^{n_2-n_1-1}}{\mu^{d(n_2-1)+1}B(n_2-n_1, n_1)}, \quad \mu > 1 \tag{1}$$

where $B(\cdot,\cdot)$ is the Beta function and $d$ is the id. This result elegantly generalizes the Two-NN estimator proposed in [8], since when $n_1 = 1$ and $n_2 = 2$, the density in (1) reduces to a $Pareto(1, d)$. These novel distributional results allow the derivation of efficient estimators for $d$ in both the frequentist and Bayesian paradigms. In this contribution, we consider the unbiased maximum likelihood estimator of $d$, which has closed form when $n_1 = 1$ and $n_2 = 2$. Therefore, the Two-NN estimator is given by

$$\widehat{d} = \frac{n-1}{\sum_{i=1}^n \log(\mu_i)}, \tag{2}$$

with $1-\alpha$ confidence interval

$$\tau_{1-\alpha} = \left[ \frac{\widehat{d}}{q_{IG_{n,n-1}}^{1-\alpha/2}}, \frac{\widehat{d}}{q_{IG_{n,n-1}}^{\alpha/2}} \right]$$

where $q_{IG}^{\alpha/2}$ denotes the $\alpha/2$-order quantile of the Inverse Gamma distribution. Alternatively, in the Bayesian framework, it suffices to set a prior $d \sim Gamma(a, b)$ and derive the posterior distribution

$$d|\mu_1, \ldots, \mu_n \sim Gamma\Big(a + n, b + \sum_{i=1}^{n} \log(\mu_i)\Big), \tag{3}$$

whose mean is asymptotically equivalent to (2). It must be pointed out that, if compared to the Gride estimator based on values of $n_2 > 2$, the Two-NN is more likely to satisfy the local homogeneity assumption of the Poisson process because it only uses information up to the scale of the second nearest neighbor of each statistical unit. On the other hand, it is potentially affected by the presence of noise in the data.

## 3. Identifiability and the intrinsic dimension of the posterior support

Let us introduce the following non-identifiable setting. To fix notation, let $\mathcal{Y}$ be a sample space and let $Y_1, \ldots, Y_N$, $Y_j \in \mathcal{Y}$ be a random sample whose generating process is described by some statistical model $\mathcal{M}_\theta$. Here, $\theta \in \Theta$ is a parameter vector which indexes the family $(\mathcal{M}_\theta)_{\theta \in \Theta}$. Suppose that the aim is to model $Y_1, \ldots, Y_N$ adopting a family of models $(\mathcal{G}_\lambda)_{\lambda \in \Lambda}$ with $\Theta \subset \Lambda$. Here, we assume $\dim(\Lambda) > \dim(\Theta)$. Since the function $f : \Lambda \times \mathcal{Y}^N \to \mathbb{R}^+$ is not injective once we condition on the sample $Y_1, \ldots, Y_N$, we can readily conclude that the model $\mathcal{G}_\lambda$ is not identified, since the likelihood $\ell(\lambda) := f(\lambda; Y_1, \ldots, Y_N)$ is flat on a set of $\Lambda$ characterized by positive Lebesgue measure.

In the frequentist setting, this characteristic compromises the optimization procedure. On the Bayesian side, instead, one can always set a prior distribution over the parameter vector $\lambda$ such that the posterior distribution is well-defined. This caveat allows posterior inference, being it exact or approximated using suitable sampling algorithms - such as Markov Chain Monte Carlo (MCMC) simulation. Since, under the Bayesian framework, unidentifiability does not have the same disruptive effects in the estimation as under the frequentist one, sometimes the model feasibility is taken for granted, leading to ignoring identifiability issues that may exist in the model specification.

To understand how a posterior is affected by the non-identifiability setting described above, let us split the parameter space into its identifiable and unidentifiable components: $\Lambda = \Lambda' \times \Lambda''$. Here, $\Lambda''$ is the non-identified component. If we let $\lambda|Y_1, \ldots, Y_N$ be the posterior random variable, it yields a marginal component $\lambda'|Y_1, \ldots, Y_N$ that degenerates to a constant as $N \to \infty$. This also implies that the id $d$ of the support of $\lambda|Y_1, \ldots, Y_N$ converges to $\dim(\Lambda'') < \dim(\Lambda)$. In real settings, the posterior distribution is often not available in closed form, but it is possible to sample from it employing MCMC algorithms. In this case, the id $d$ of the $n \times D$ matrix, where each row is a sample from the posterior, can be estimated using the methodologies previously presented in Section 2, bringing valuable insights on the identifiability of the model. Notice that here $D = \dim(\Lambda)$ and the id $d = \dim(\Lambda'')$.

Before proceeding to a simulated example, a cautionary note must be given. In well-identified models, when $N$ is finite, we expect the id to be equal to $\dim(\Theta) = \dim(\Lambda)$. When $N \to \infty$, instead, the entire posterior converges to a point mass, resulting in a degenerate id equal to 0. This argument suggests that the id of the posterior distribution support is determined by the magnitude of variability each marginal component has w.r.t. each other. In an unidentifiable setting, we expect the variability to be much higher for the specific unidentified parameters set. Conversely, the well-identified model leads to a posterior distribution whose marginal components have variances of comparable scales for any $N$ resulting in an id of $\dim(\Theta)$, although as $N \to \infty$ the variances vanish to 0 jointly with the id.

DRAFT

## 3.1 A simulated example

To provide a more concrete example, we consider the following simulated setting. Suppose we want to estimate a parameter vector $(\beta_0, \beta_1, \ldots, \beta_4)^\top \in \Lambda$ representing the coefficients of a regression line. According to the notation introduced above, $\mathcal{G}_\lambda$ is the linear model, while $\lambda = (\beta_0, \beta_1 \ldots, \beta_4)^\top$. We consider the following two data-generating processes:

$$Y_j = \beta_0 + \beta_1 + \beta_2 X_{2j} + \beta_3 X_{2j} + \beta_4 X_{4j} + \varepsilon_j, \quad j = 1, \ldots, N, \tag{4}$$

and

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + \varepsilon_j, \quad j = 1, \ldots, N. \tag{5}$$

We assume $X_{kj} \overset{ind.}{\sim} Unif\,[u_k, v_k]$ for $k = 1, \ldots, 4$, with $\boldsymbol{u} = (-2, -3, -4, 0)$ and $\boldsymbol{v} = (2, 3, 0, 6)$. Moreover, $\varepsilon_j \sim \mathcal{N}(0, 1)$ and, crucially, the priors are chosen to be highly non-informative $\beta_l \sim Cauchy(0, 10),\ l = 0, 1, \ldots, 4$.

Notice that model (4) is parametrized by a parameter vector whose dimension is smaller than $\dim(\Lambda)$ leading to the unidentifiability of some parameters. Thus, data generated from (4) do not allow the identification of $(\beta_0, \beta_1 \ldots, \beta_4)^\top$ and this makes the id of the support of the 5-dimensional posterior shrink to 2, which is the number of the unidentified parameters. On the other hand, we expect the id of the posterior distribution support obtained on data generated by (5) to be around 5 for finite sample sizes and 0 asymptotically in $N$. We must point out that the choice of non-informative priors allows a fast convergence of id to 2. Assuming informative priors could lead to incorrect conclusions about the id of the posterior and, consequently, on the identifiability of the model when the sample size is small.

To validate our hypothesis, posterior sampling is performed using efficient HMC algorithms available in the STAN software [4]. At the same time, the id estimation can be carried out via the routines in the R package `intRinsic` [6]. As a first step, we fit the model obtaining the Markov chains of the parameters. In detail, if we take an MCMC sample of $n = 1000$ and $N = 150$ statistical units, we obtain a Two-NN id estimate of 2.12 with 95% confidence interval $[1.99, 2.26]$ on model (4). On model (5), we obtain an estimated id of 4.87 with 95% confidence interval $[4.58, 5.18]$. If we instead use the Bayesian id estimator, setting a non-informative prior $d \sim Gamma(10^{-3}, 10^{-3})$ we get as posteriors a $Gamma(10^3 + 10^{-3}, 471.03 + +10^{-3})$ with mean 2.12 for model (4) and a $Gamma(10^3 + 10^{-3}, 205.22 + 10^{-3})$ with mean 4.87 for model (5). The posterior densities are reported in Figure 1. As expected, for finite sample sizes, the id of the support of the identifiable posterior is equal to the ambient dimension (i.e., 5) because all the marginal components are shrinking at the same rate to a point mass. The non-identifiable posterior scales with the sample size only in its identifiable marginals, while the non-identifiable ones do not scale with $N$ and keep the variability induced by the prior distributions. This source of variability dominates the variability of the identified marginals. This last fact holds asymptotically and for finite sample sizes. Thus, the non-identifiable marginals determine an id of 2.

DRAFT

## 4. Conclusions

In this paper, we proposed an easy-to-use tool to detect identifiability issues in Bayesian models. The methodology is based on estimating the id of samples from the posterior distribution. The estimation of the id is carried out employing the Two-NN, a method that recently appeared in the literature. The simulated example shows that the id of the posterior sample is equal to the number of non-identified parameters. The presented results could be easily extended to other modeling settings, such as generalized linear models or mixture models.

## References

[1] Allman, E. S., Matias, C., and Rhodes, J. A. (1974). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37 (6A):3099–3132.

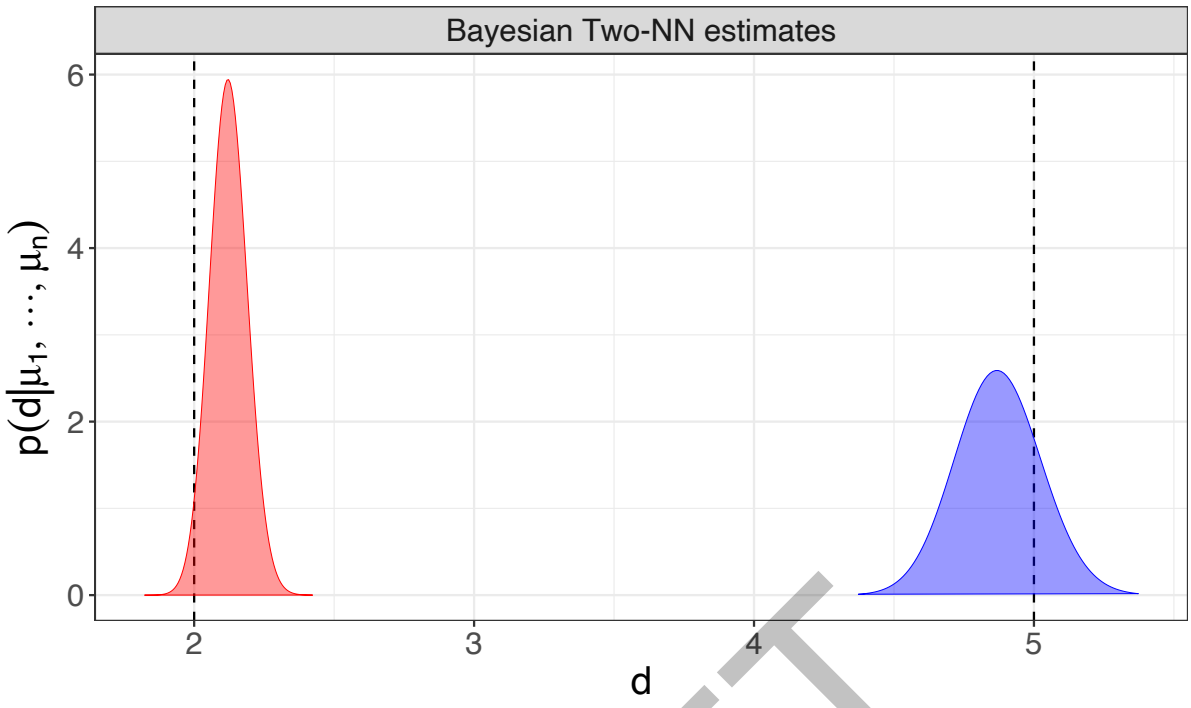



Figure 1: The figure displays the id posterior densities of the MCMC samples obtained by the two models. The id obtained from model (4) is displayed in red, while the id estimated from model (5) is in blue.

DRAFT

[2] Bowden, R. (1973). The theory of parametric identification. *Econometrica*, 41 (6):1069–1074.

[3] Campadelli, P., Casiraghi, E., Ceruti, C., and Rozza, A. (2015). Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework. *Mathematical Problems in Engineering*, 2015.

[4] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

[5] Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89 (425):259–267.

[6] Denti, F. (2021). intRinsic: an R package for model-based estimation of the intrinsic dimension of a dataset.

[7] Denti, F., Doimo, D., Laio, A., and Mira, A. (2022). The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1).

[8] Facco, E., D'Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1).

[9] Goodam, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61 (2):215–231.

[10] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and em algorithm. *SIAM Review*, 26 (2):195–239.

[11] Roehrig, C. S. (1988). Conditions for identification in nonparametric and parametric models. *Econometrica*, 56 (2):433–447.

[12] Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39 (3):577–591.

# Computing Highest Density Regions with Copulae

Nina Deliu[a] and Brunero Liseo[b]

[a]MEMOTEF, Sapienza Università di Roma, nina.deliu@uniroma1.it

[b]MEMOTEF, Sapienza Università di Roma, brunero.liseo@uniroma1.it

**Abstract**

We investigate the problem of deriving highest density regions (HDRs) from multivariate data samples. We are interested in estimating minimum volume sets that contain a given probability. In the case of unknown distribution probabilities $f$, the problem involves their estimation, which may be challenging in multidimensional settings. Motivated by the ubiquitous role of copula modelling in modern statistics, we explore their use in the context of HDR estimation. Rather than directly estimating the multivariate $f$, we propose to estimate the marginals and their dependence structure, i.e., the copula structure, separately. We evaluate this new method, using both a parametric and a nonparametric approach, in a number of synthetic experiments and considering a real dataset.

***Keywords:*** Highest density regions, Copula modelling, Kernel density estimation

DRAFT

## 1. Introduction

A ubiquitous problem in statistics is to derive statistical intervals or *regions* (in the multivariate setting) for population parameters or other unknown quantities. Given a random sample of data, they provide a way to quantify the uncertainty about a quantity of interest, or simply a way to summarize the information contained in a distribution. In this work, we are interested in statistical *regions* for summarizing probability distributions and we focus on one approach to addressing this problem: *highest density regions* (HDRs, 1). As the name suggests, an HDR specifies the set of points of highest density: the density for every point inside the region is greater than that for every point outside it. More specifically, as we will better discuss in Section 2, the concrete problem is to estimate minimum volume sets of the form $R(f_\alpha) = \{\mathbf{x}\colon f_{\mathbf{X}}(\mathbf{x}) \geq f_\alpha\}$, such that $P(\mathbf{X} \in R(f_\alpha)) \geq 1-\alpha$, where $f_{\mathbf{X}}$ is the probability distribution of the variable of interest $\mathbf{X} \in \mathbb{R}^d$ and $1-\alpha$, with $\alpha \in (0,1)$, a prespecified coverage probability. In principle, HDRs can be derived for any probability distribution and their scope can be widely different. The following are possible applications of HDRs.

**Forecasting** The goal is to obtain a "prediction region" for a set of observable variables in order to inform any required action (for illustrative examples, see e.g., 1; 2);

**Anomaly detection** The goal is to detect abnormal observations from a sample: if a data point does not belong to a region of normal or concentrated data, then it is regarded as anomalous (see e.g., 3);

**Unsupervised or semi-supervised classification** Identify areas or clusters with a relatively high concentration of a given phenomenon, e.g., areas with remarkably high coronavirus incidence (4).

Such regions are of interest in Bayesian analysis as well, in the formulation of *highest posterior density regions* and *credibility regions* (5; 6). In that context, they are based on a posterior distribution.

Because of their flexibility "to convey both multimodularity and asymmetry in the forecast density", HDRs are argued to be a "more effective summary of the forecast distribution than other common forecast regions" (1). However, to build an accurate HDR, one needs to know (or accurately estimate) the

underlying probability distribution. Methods for estimating $f$ such as the kernel density estimator (KDE, 7) or the local likelihood approach (8) work very well for unidimensional problems, but they may be inefficient for multidimensional problems (9). For example, the bandwidth selection in KDE, recognized as the most crucial and difficult step (see e.g., Chapter 2 in 10), has no definite and unique solution. Further, high-dimensional data pose challenges also from the algorithmic/computational perspective. Altogether, these aspects hamper the ability to derive an appropriate HDR.

The scope of this work is to propose an alternative approach to build HDRs using *copulae* so as to overcome the direct estimation of the multivariate $f$. Specifically, copulae allow to relax the estimation of multivariate random vectors, by separately estimating the marginals and their dependence structure, i.e., the copula model (11). Motivated by the ubiquitous role of copula modelling in modern multivariate statistics, specifically multivariate density estimation, we explore its use the context of HDR estimation.

## 2. Problem Setting: Highest Density Regions

Assume we have access to a sample $\mathbf{s}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of independent and identically distributed (iid) observations, drawn from a probability measure $\mathbb{P}$. Each data point can be multidimensional, that is $\mathbf{x}_i \in \mathbb{R}^d$, with $d \geq 1$. We denote by $x_i^{(j)}$ the $j$-th coordinate of $\mathbf{x}_i$, for $j = 1, \dots, d$ and $i = 1, \dots, n$. For simplicity, we restrict our analysis to bivariate data points with $d = 2$, and we focus on continuous random variables $\mathbf{X} = (X^{(1)}, X^{(2)}) \in \mathbb{R}^2$. Let $f_{\mathbf{X}}$ denote the probability density function (PDF) of $\mathbf{X}$ and $F_{\mathbf{X}}$ its cumulative density function (CDF). Then, given a coverage probability $1-\alpha$, with $\alpha \in (0, 1)$, the $100(1-\alpha)\%$ HDR is defined as the subset $R(f_\alpha)$ of the sample space of $\mathbf{X}$ such that:

$$R(f_\alpha) \doteq \{\mathbf{x} \colon f_{\mathbf{X}}(\mathbf{x}) \geq f_\alpha\}, \tag{1}$$

where $f_\alpha$ is the largest constant such that $P(\mathbf{X} \in R(f_\alpha)) \geq 1 - \alpha$.

It follows from the definition that the boundary of an HDR consists of those values of the sample space with equal density. Hence a plot of a bivariate HDR is a form of contour plot. One of the most distinctive properties of HDRs is that, of all regions of probability coverage $100(1-\alpha)\%$, the HDR has the smallest region possible in the sample space. Clearly, an HDR always contains the global mode, and in the case of multimodal distributions, it often consists of several disjoint subregions, each containing a local mode. This provides useful information which is "masked" by other types of statistical regions.

To estimate an HDR for $\mathbf{X}$ according to Eq. (1), one needs to know the density function $f_{\mathbf{X}}$. If this is known, the typical way to compute HDRs is the density quantile approach (1), which is based on the following rationale. Let $\mathbf{Y} = f_{\mathbf{X}}(\mathbf{X})$ be the random variable obtained by transforming $\mathbf{X}$ by $f_{\mathbf{X}}$ (bounded and continuous in $\mathbf{x}$). Consider a set of independent observations from the distribution of $\mathbf{X}$, say $\mathbf{s}_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$. It follows that independent observations from the distribution of $\mathbf{Y}$ can be obtained as $\{f_{\mathbf{X}}(\mathbf{x}_1), \dots, f_{\mathbf{X}}(\mathbf{x}_m)\}$. Consider now the ordered sample $\{f_{(1)}, \dots, f_{(m)}\}$ with $f_{(j)}$ the $j$-th largest of $f_{\mathbf{X}}(\mathbf{x}_i)$, $i = 1, \dots, m$, so that $f_{(j)}$ is the $(j/m)$ sample quantile of $\mathbf{Y}$. Then, given a constant $\alpha \in [0, 1]$, and denoted with $\lfloor \cdot \rfloor$ the floor operator, we have that:

$$\hat{f}_\alpha \doteq f_{\lfloor \alpha m \rfloor} \to f_\alpha, \quad \text{and} \quad R_m(\hat{f}_\alpha) \doteq \{\mathbf{x} \colon f_{\mathbf{X}}(\mathbf{x}) > \hat{f}_\alpha\} \to R(f_\alpha), \quad \text{as } m \to \infty.$$

Basically, the HDR can be derived based on the sample quantile of $\mathbf{Y} = f_{\mathbf{X}}(\mathbf{X})$.

However, in most real-world scenarios, the density function is unknown. If we have access to a sample of iid observations $\mathbf{s}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, then we can estimate it, and subsequently obtain an estimate of the $100(1-\alpha)\%$ HDR by using the again the density quantile approach with

$$\hat{R}_n(\hat{f}_\alpha) \doteq \{\mathbf{x} \colon \hat{f}(\mathbf{x}) > f_{\lfloor \alpha n \rfloor}\}, \tag{2}$$

where $\hat{f}$ is a possibly consistent estimator of $f$.

Note that for small $n$ it may not be possible to get a reasonable density estimate. Besides, with few observations and no prior knowledge of the underlying density function, there seems little point in attempting to summarize the sample space. In higher dimensions, the difficulty of selecting an appropriate region is even greater due to the density estimation challenges (see e.g., 9; 10).

DRAFT

## 3. Using Copulae for Deriving HDRs

A general $d$-dimensional copula $C\colon [0,1]^d \to [0,1]$ is a joint cumulative density function whose $d$ marginals are uniform over $[0,1]$. Consider, without major loss of generality, the bivariate case $d=2$, with $F_{\mathbf{X}}$ the joint CDF of the random vector $\mathbf{X} = (X^{(1)}, X^{(2)})$, and $F_{X^{(1)}} = F_1$ and $F_{X^{(2)}} = F_2$ its marginals. Then, it follows from the *probability-integral transform* (12) that the joint distribution of $(F_1, F_2)$ is a copula, say $C_{\mathbf{X}}$, and its expression can be derived by noting that

$$C_{\mathbf{X}}(u^{(1)}, u^{(2)}) = \mathbb{P}(F_1(X^{(1)}) \leq u^{(1)}, F_2(X^{(2)}) \leq u^{(2)}) = F_{\mathbf{X}}(F_1^{-1}(u^{(1)}), F_2^{-1}(u^{(2)})).$$

Letting $u^{(j)} \doteq F_j(x^{(j)})$, $j=1,2$, this yields the following result due to Sklar (13):

$$F_{\mathbf{X}}(x^{(1)}, x^{(2)}) = C_{\mathbf{X}}(F_1(x^{(1)}), F_2(x^{(2)})), \quad \forall \mathbf{x} = (x^{(1)}, x^{(2)}) \in \mathbb{R}^2.$$

In summary, we can decompose the bivariate CDF $F_{\mathbf{X}}$ into a composition of the two marginal distribution functions and a two-dimensional copula $C_{\mathbf{X}}$. $C_{\mathbf{X}}$ is the copula of $F_{\mathbf{X}}$ and describes the dependence structure of $F_1$ and $F_2$. We refer to (11) for book-length treatment of the foregoing ideas.

In case the bivariate distribution has density $f$, and if this is available, it holds further that

$$f_{\mathbf{X}}(x^{(1)}, x^{(2)}) = c_{\mathbf{X}}(F_1(x^{(1)}), F_2(x^{(2)}))f_1(x^{(1)})f_2(x^{(2)}),$$

with $c$ being the copula density and $f_1$ and $f_2$ the marginal densities. The main advantage of this representation over the one involving the joint PDF is that an estimate of $f_{\mathbf{X}}$ can be obtained by estimating the marginals and the copula density separately, evading potential high-dimensional data challenges (see e.g., 14). Furthermore, copulae offer a flexible framework that can capture complex dependency structures.

If $\hat{c}$ is an estimate of the copula density, we propose to estimate the $100(1-\alpha)\%$ HDR as

$$\hat{R}_n(\hat{f}_\alpha) = \{\mathbf{x}\colon \hat{c}_{\mathbf{X}}(\hat{F}_1(x^{(1)}), \hat{F}_2(x^{(2)}))\hat{f}_1(x^{(1)})\hat{f}_2(x^{(2)}) > f_{\lfloor \alpha n \rfloor}\},$$

with $\hat{f}_j$ and $\hat{F}_j$ consistent estimators of the marginals $f_j$ and $F_j$, $j=1,2$. While here, for the sake of space, we focus on the bivariate case, we emphasize that the approach can be easily extended to higher dimensions, as copulae naturally apply to multidimensional contexts (see e.g., *vine copula* methods; 14).

DRAFT

## 4. Empirical Evaluation

**Simulation studies** We start with simulation studies, considering the following four data-generation scenarios, with constant parameters fixed at $\mu_1 = 0, \mu_2 = 1, \sigma_1 = \sigma_2 = 2$, and $w_1 = 1 - w_2 = 0.7$ over an increasing number of sample sizes (from 50 to 10,000). For copula specifications, we refer to (11).

**SC1: Gaussian marginals $\mathcal{N}$ – Gaussian copula $C^{\text{Gauss}}$**

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = \mathcal{N}(\mu_2, \sigma_2), \quad C = C^{\text{Gauss}}_{\rho=0.7}$$

**SC2: Gaussian $\mathcal{N}$ & Student $t$ marginals – Clayton copula $C^{\text{Clay}}$**

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = t_{\nu=10}, \quad C = C^{\text{Clay}}_{\alpha=2}$$

**SC3: Gaussian $\mathcal{N}$ & Gaussian mixture marginals – Student t copula $C^{\text{t}}$**

$$f_1 = \mathcal{N}(\mu_1, \sigma_1), \quad f_2 = w_1\mathcal{N}(\mu_2, \sigma_2) + w_2\mathcal{N}(\mu_2 + 10, \sigma_2), \quad C = C^{\text{t}}_{\rho=0.4,\nu=6}$$

**SC4: Gaussian $\mathcal{N}$ mixture marginals – Gaussian copula $C^{\text{Gauss}}$**

$$f_1 = w_1\mathcal{N}(\mu_1, \sigma_1) + w_2\mathcal{N}(\mu_1 + 10, \sigma_1), \quad f_2 = w_1\mathcal{N}(\mu_2, \sigma_2) + w_3\mathcal{N}(\mu_2 + 10, \sigma_2), \quad C = C^{\text{Gauss}}_{\rho=0.7}.$$

For each scenario, we evaluate the following three methods.

**Method1: Direct estimation of the bivariate density** We use the nonparametric KDE (7), and consider the asympotically optimal solution proposed in (15) for the bandwidths selection.

**Method2: Indirect fully-parametric copula-based estimation of the bivariate density** For the estimation of the marginals, we consider the true data-generation processes models (with no misspecification) and maximum likelihood fitting. For the copula model, we perform both model selection (with the AIC criterion) and parameter estimation (with maximum likelihood estimation).

**Method3: Indirect fully-nonparametric copula-based estimation of the bivariate density** For the marginal densities, we use the standard KDE. For the copula model, we use a KDE approach with the *transformation local likelihood estimator* of (16). We use the R KDECOPULA package, adopting the method with quadratic polynomials and nearest-neighbor bandwidths (17).

To quantify the performance of the methods in the simulation study, we call *positive* those points which should be outside the region and *negative* the others. Let FP, TP, FN and TN be, respectively, the number of false positive, true positive, false negative, and true negative points. Well-established measures of inefficiency are the False Negative (Positive) Rates (FNR and FPR), and the Total Error Rate (ER):

$$\text{FNR} = \frac{FN}{FN+TP}; \quad \text{FPR} = \frac{FP}{FP+TN}; \quad \text{ER} = \frac{FN+FP}{FN+FP+TN+TP}.$$

All methods are evaluated based on $\alpha = 0.05$, that is a coverage probability of 95%.

**Results** As depicted in Figure 1, the three methods lead to slightly different results, with the two copula-based approaches outperforming the direct KDE (Method1). Compared to the nonparametric Method3, the parametric copula-based approach (Method2) shows the lowest ER across all different scenarios and sample sizes, with an exception for the largest sample sizes, where the difference is negligible.

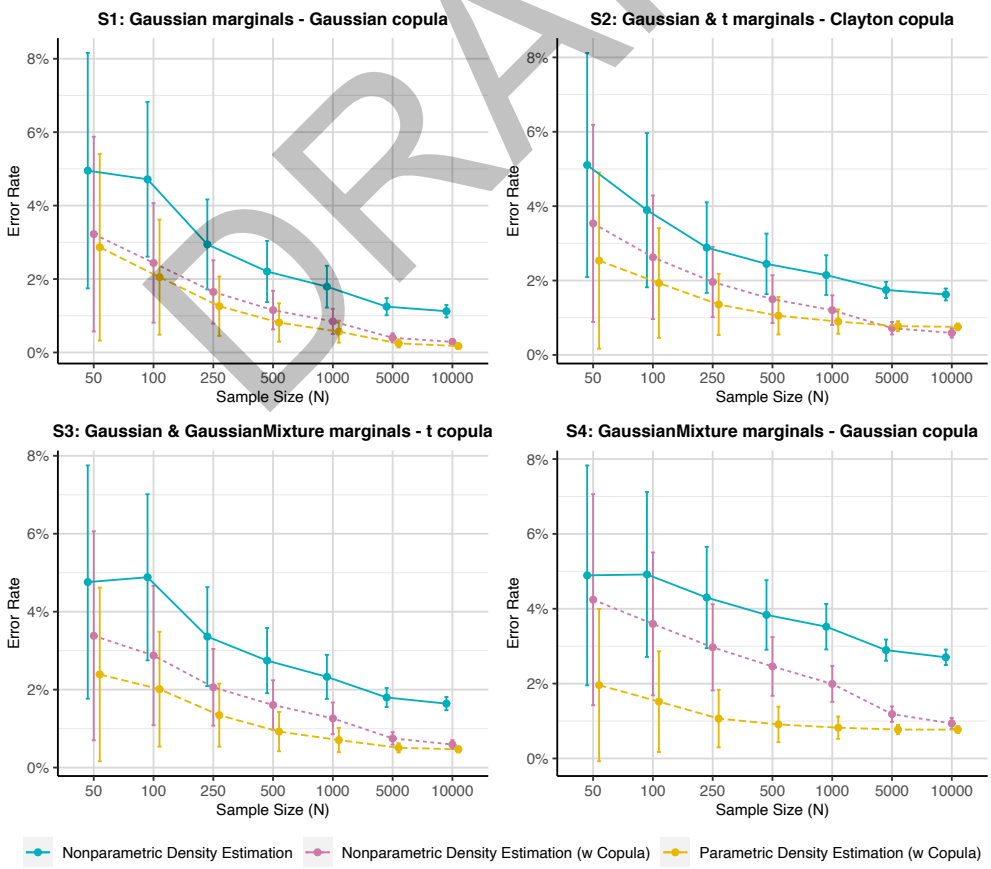



Figure 1: Total error rate (mean and error bars (mean $\pm$ SD), averaged across $10^4$ MC samples) of the three compared methods across the different scenarios and for different sample sizes.

DRAFT

Looking at the FPR and FNR, results are very similar across different scenarios and we only discuss scenario S2. As displayed in Figure 2, Method1 has the best FPR performance, with a value close to 0. Rather than being a result of optimal performance, this is due to the fact almost or all data points were classified as highest density points, with no *positives* detected. This is also reflected in its FNR, highlighting a low ability to correctly place *positives* outside the HDR. The two copula-based approaches result in overall better performances, with a slight superiority of the parametric Method2. This was expected as simulation schemes consisted of only parametric copula families, and there is no misspecification.

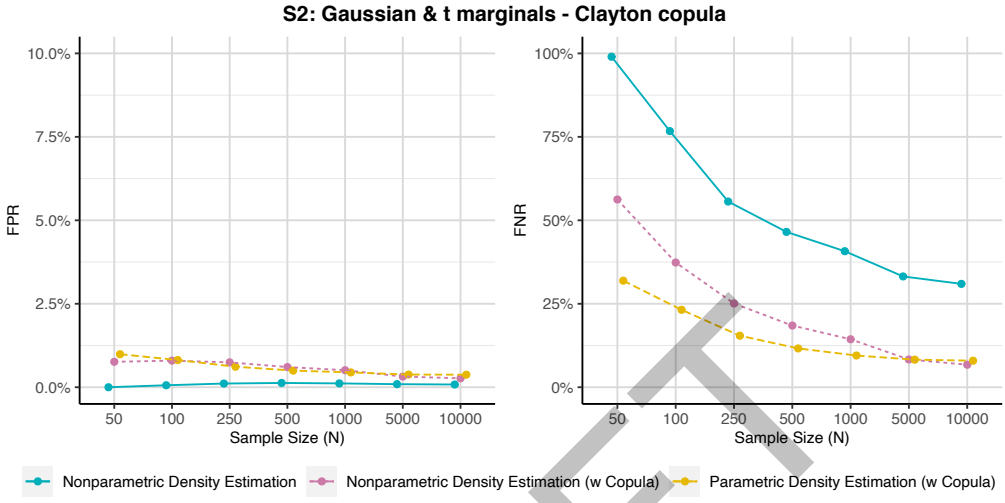



Figure 2: FPR and FNR (averaged across $10^4$ MC samples) of the three compared methods for different sample sizes, relatively to scenario S2.

**Application to MAGIC Data** We apply the proposed methods for constructing a HDR for the joint distribution of two variables from the MAGIC dataset (https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope). The data simulate the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope. We focus on gamma observations (overall $n = 12,332$), and consider the two variables "fConc1" and "fM3Long". In such a case (as deduced from Figure 3), the parametric approach is inappropriate for both the estimation of the marginal distribution and, more importantly, the copula model. Thus, we illustrate the derived HDR using the nonparametric Method1 and Method3 only. While in absence of the underlying truth it is not possible to reliably evaluate the two methods, it seems that the copula-based approach (Method3; right plot), more sensibly excludes the tail data points (which may be expected to have a lower density) from the HDR.

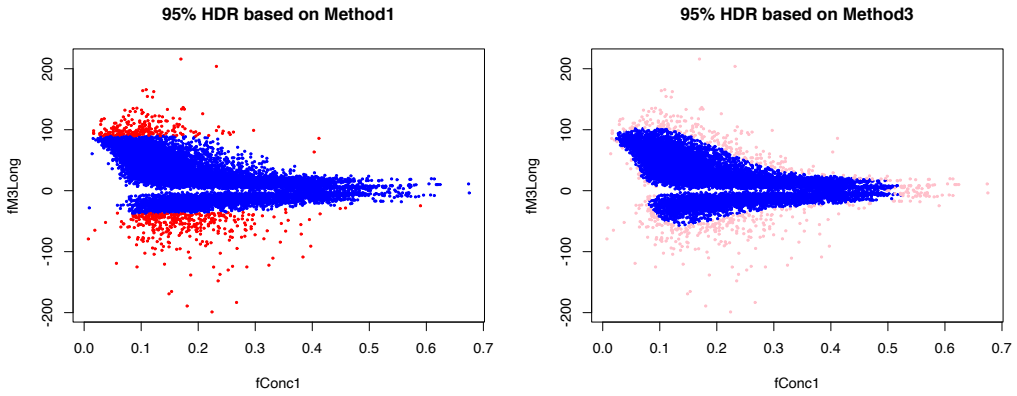



Figure 3: 95% HDR for two variables from the MAGIC dataset with Method1 and Method2.

DRAFT

## 5. Concluding Remarks

In this work, we proposed an alternative strategy for deriving HDRs in multivariate contexts using *copulae*, and evaluated both a parametric and a nonparametric approach. Compared to traditional kernel density estimation, the copula-based HDR resulted in lower missclasification errors in a number of simulation scenarios and possibly in real data. Although in this work we focused on the bivariate case ($d = 2$), we expect to see remarkable advantages over an increased number of variables $d > 2$. In fact, the extension of the common KDE to high dimensions has proven challenging in terms of both computational efficiency and statistical inference. We aim to pursue such a direction in future work, exploring, e.g., the use of vine copulae to construct flexible dependence models for an arbitrary number of variables using only bivariate building blocks.

## References


[1] Rob J Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

[2] Jae H Kim, Iain Fraser, and Rob J Hyndman. Improved interval estimation of long run response from a dynamic linear model: A highest density region approach. *Computational Statistics & Data Analysis*, 55(8):2477–2489, 2011.

[3] Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[4] Paula Saavedra-Nieves. Nonparametric estimation of highest density regions for COVID-19. *Journal of Nonparametric Statistics*, 34(3):663–682, 2022.

[5] George EP Box and George C Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, 2011.

[6] Noyan Turkkan and T Pham-Gia. Computation of the highest posterior density interval in Bayesian analysis. *Journal of Statistical Computation and Simulation*, 44(3-4):243–250, 1993.

[7] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[8] Nils Lid Hjort and M Chris Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24(4):1619–1647, 1996.

[9] Han Liu, John Lafferty, and Larry Wasserman. Sparse nonparametric density estimation in high dimensions using the rodeo. In *Artificial Intelligence and Statistics*, pages 283–290. PMLR, 2007.

[10] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

[11] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[12] George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2021.

[13] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[14] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[15] José E. Chacón, Tarn Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21(2):807–840, 2011.

[16] Gery Geenens, Arthur Charpentier, and Davy Paindaveine. Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873, 2017.

[17] Thomas Nagler. kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities. *Journal of Statistical Software*, 84(7):1–22, 2018.


DRAFT

# Parameter estimation via Indirect Inference for multivariate Wrapped Normal distributions

Francesca Labanca and Anna Gottard

Department of Statistics, Computer Science, Applications (DiSIA), Florence University, Florence, Italy;
francesca.labanca@unifi.it, anna.gottard@unifi.it

**Abstract**

Multivariate circular observations, i.e. points on a $p$-dimensional torus, can be seen as random angular quantities. A way to model this kind of observation is using the multivariate wrapped normal distribution. In this paper, we tackle the problem of estimating the multivariate wrapped normal distribution parameters using a given set of samples. Here we focus on the parameters in the variance-covariance matrix that characterize the dependence among the variables of interest. For this purpose, we propose an Indirect Inference approach that relies on a Normal auxiliary model defined on the real space with the same dimension as the target model's parameter space. This approach provides an efficient and accurate estimation method for the parameters of the multivariate wrapped normal distribution.

***Keywords:*** Indirect Inference; Multivariate directional data; Multivariate Wrapped Normal; Statistical computation and simulation.

DRAFT

## 1. Introduction

In a wide variety of fields, including natural and physical sciences, measurements involve directions and can be characterized as random angular quantities. Such data are commonly referred to as *directional data*, and their analysis and understanding rely on specific statistical models and procedures for non-Euclidean space; see e.g. Mardia and Jupp (2000) (12); Jona Lasinio and Gelfand (2012) (10). Several approaches have been proposed in the literature for defining distributions for directional data, including the embedding, intrinsic, and wrapping approaches, which are widely used and relevant for analyzing directional data. In the embedding approach, the sample space is obtained by radial projection of a distribution on the real space, onto the unit circle. In the intrinsic approach, the directions are represented as points on the circle, and probability distributions are defined on the circle directly. Instead, the wrapping approach constructs circular random variables as the modulo $2\pi$ version of real random variables. The circular distribution is obtained by wrapping the corresponding real distribution on the $p$-torus $\mathbb{T}^p$, the cartesian product of $p$ circles. For each approach, several distributions have been proposed. In a multivariate setting, the most popular proposals are respectively the *Inverse Stereographic Gaussian* distribution (IS), see e.g. Selvitella (2019) (16), the *Von Mises* (VM) and the *Wrapped Normal* (WN).

In directional statistics, the VM and the WN distributions often play a similar role as the Normal distribution on the real space. The most used distribution, in the univariate framework, is the VM due to the analytical tractability of the maximum likelihood estimators. In the case of a high concentration

of the univariate distributions around the mean, both the VM and the WN can be approximated as Gaussian distributions, and they are a good approximation of each other. In the multivariate framework, both distributions present major problems with maximum likelihood estimation. The VM multivariate distribution has an unknown normalizing constant, and, as suggested by Mardia (2016) (13), inference can be achieved via the score-matching estimator, see e.g. Yu, Drton, and Shojaie (2022) (17). Conversely, likelihood-based inference for the WN distribution presents major difficulties as it involves infinite sums. In literature, among the main approximate inference methods for the parameters inference of a WN distribution, we can mention the iterative reweighted maximum likelihood estimating equation algorithm for univariate WN, also available in the R package `circular`, proposed by Agostinelli (2007) (1); the two steps Expectation-Maximization algorithm proposed by Fisher and Lee (1994) (8), recently generalized and improved by Nodehi et al., (2021) (14); the combination of maximum likelihood and moment matching proposed by Kurz and Hanebeck (2015) (11); the data augmentation approach in a Bayesian framework proposed by Coles (1998) (6) and Ravindran and Ghosh (2011) (15). This leads to the main goal of this paper: proposing an alternative parameter estimation approach for both univariate and multivariate WN distributions using Indirect Inference. Indirect Inference (II), see e.g. Gourieroux, Monfort, Renault (1993) (9), is a powerful simulation-based method for estimating parameters in complex statistical models whose likelihood is, for some reason, intractable. The method entails simulating data from an auxiliary model, whose likelihood is tractable, and then estimating parameters in the target model using an auxiliary criterion on the simulated data. Interesting use of the Indirect inference method for GARCH-type models can be found in Calzolari, Halbleib, and Parrini (2014) (4); Calzolari and Halbleib (2018) (5). Another application of this method can be found in Bee's (2022) (2) work on the wrapped stable distribution.

The remainder of this paper is organized as follows. Section 2. describes the multivariate WN model. Section 3. describes the II procedure in a general framework and introduces a possible application for inference on WN models. Section 4. gives final comments and remarks.

DRAFT

## 2. Wrapped Normal distribution

Given a discrete or continuous random variable $\boldsymbol{X}$ defined on $\mathbb{R}^p$ and its cumulative distribution function (CDF) $G$, the wrapped counterpart is defined as $\boldsymbol{Y} = \boldsymbol{X}\ (mod\ 2\pi)$, where the modulo is applied component-wise, and the corresponding wrapped CDF is

$$F(\boldsymbol{y}) = \sum_{\boldsymbol{j}\in\mathbb{Z}^p} \left[G(\boldsymbol{y} + 2\pi\boldsymbol{j}) - G(2\pi\boldsymbol{j})\right], \qquad \boldsymbol{y} \in [0, 2\pi)^p.$$

Similarly, for any r.v. $X$ admitting a density function $g$, the wrapped version of the density function can be defined as

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{j}\in\mathbb{Z}^p} g(\boldsymbol{y} + 2\pi\boldsymbol{j}), \qquad \boldsymbol{y} \in [0, 2\pi)^p.$$

The support of the resulting wrapped random variable is the $p$-torus $\mathbb{T}^p$ and its wrapped density function $f$ is obtained from the density function $g$ wrapping the $\mathbb{R}^p$ space on $\mathbb{T}^p$ and summing the densities of all identified points component-wise.
In this framework, a multivariate Wrapped Normal distribution is obtained by choosing $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Consequently, the random variable $\boldsymbol{Y} = \boldsymbol{X}\ (mod\ 2\pi)$ has a multivariate Wrapped Normal distribution $WN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the exact same parameters as $\boldsymbol{X}$. Due to the model definition, the mean vector parameter is not identifiable: the parameter $\boldsymbol{\mu}$ is expressed as $\boldsymbol{\mu} = \boldsymbol{\mu}^* + 2\pi\boldsymbol{j}$, where $\boldsymbol{\mu}^*$ is the mean direction and a parameter on $[0, 2\pi)^p$ and, since the winding number j is not observed, it is impossible to identify $\boldsymbol{\mu}$. However, $\boldsymbol{\mu}^*$ is identifiable. In this paper, we focus on estimating the variance-covariance matrix, limiting to the case of a moderately large total variance. For very large variances, the WN distribution converges to a circular uniform distribution on the $p$-torus, and its parameters are not defined. The WN density distribution is defined as an infinite series and has no closed form, making it impossible to obtain

parameter estimates via direct maximization of the WN likelihood. Thus, different techniques must be used.

## 3. Indirect Inference for Wrapped Normal models

In this section, we first introduce the indirect inference estimating procedure for a general model and then concentrate on the case of the wrapped normal distribution.

Consider a multivariate random variable $\boldsymbol{Y}$ with values in $\mathcal{Y}$ and let $\mathcal{M}(\boldsymbol{\theta})$ be an adequate model for it, admitting a density function $f$, that depends on a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Suppose that $\mathcal{M}(\boldsymbol{\theta})$ is associated with a complex or intractable likelihood, but such that drawing random samples from $\mathcal{M}(\boldsymbol{\theta})$ is feasible once given a set of values $\boldsymbol{\theta}$. Let $\widetilde{\mathcal{M}}_a(\boldsymbol{\beta})$ be an auxiliary model with density $f_a$ and parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ that are easy to estimate.

To provide estimates for $\boldsymbol{\theta}$, the Indirect Inference method utilizes the auxiliary model $\widetilde{\mathcal{M}}_a(\boldsymbol{\beta})$, which is miss-specified and whose approximated maximum likelihood estimator can be inconsistent for $\boldsymbol{\theta}$. Using data simulated from the original model $\mathcal{M}(\boldsymbol{\theta})$, it is possible to correct the asymptotic bias under the hypothesis of the existence of an unknown bijective function between the two parameter spaces. Such a function is called *binding function* $b : \boldsymbol{\theta} \to \boldsymbol{\beta}(\boldsymbol{\theta})$. The choice of the auxiliary model impacts the functional form of the binding function and, thus, the properties of the estimators. The estimation of the auxiliary model parameters is conducted using both observed data and simulated samples generated from the original model $\mathcal{M}(\boldsymbol{\theta})$. The parameter estimates for $\mathcal{M}(\boldsymbol{\theta})$ are obtained via a calibration procedure that involves a comparison between the estimates obtained from the two aforementioned sources of auxiliary model information.

Now, consider the case of the wrapped normal distribution $\mathcal{M}(\boldsymbol{\theta}) = WN_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$. We propose to adopt precisely the Normal distribution as a suitable choice for the auxiliary model $\widetilde{\mathcal{M}}_a(\boldsymbol{\beta}) = N_p(\boldsymbol{\mu}^*, \boldsymbol{\beta})$, due to the intrinsic connection between the two models and the existence of closed-form, easy-to-compute, estimators for this auxiliary model. This is a case of so-called *exact identification*, with a one–to–one correspondence between the parameters $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ versus $(\boldsymbol{\mu}^*, \boldsymbol{\beta})$. The mean vector $\boldsymbol{\mu}^*$ can be estimated using the circular mean, as suggested by Mardia and Jupp (2000) (12). Thus, for sake of simplicity, we assume $\boldsymbol{\mu}^*$ known and we define $\boldsymbol{\theta}$ as the parameters vector of $\boldsymbol{\Sigma}$.

The II estimation procedure can be summed up in the following three steps.

(1) **PML**: Compute $\widehat{\boldsymbol{\beta}}$, the *Pseudo-Maximum-Likelihood* estimate of $\boldsymbol{\Sigma}$, using the variance-covariance matrix maximum likelihood estimator for the auxiliary model $\widetilde{\mathcal{M}}_a(\boldsymbol{\beta}) = N_p(\boldsymbol{\mu}^*, \boldsymbol{\beta})$ and using independent and identically distributed random sample $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ from the model of interest $\mathcal{M}(\boldsymbol{\theta}) = WN_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$:

$$\widehat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}^*)^T (\mathbf{y}_i - \boldsymbol{\mu}^*)}{n},$$

The simple estimation of the auxiliary model parameters leads to biased (inconsistent) estimates, called *naive estimates*. Nevertheless, since all the random vectors $\mathbf{y}_i$ are sampled from $WN_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$, they hold large information content about $\boldsymbol{\Sigma}$, thus the equation above gives an implicit relation between $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$, pseudo-true value, and the unknown binding function $\boldsymbol{\beta}(\boldsymbol{\Sigma})$ is necessarily related to the equation above.

(2) **Simulation**: Simulate $R$ samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(R)}$ of size n from the *true* model $WN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and compute $\widetilde{\boldsymbol{\beta}}^{(1)}, \ldots, \widetilde{\boldsymbol{\beta}}^{(R)}$, $R$ naive estimates. Regarding the size of the simulated samples, one may adopt a number $n^*$ equal to the observed sample size $n$. Larger $n^*$, such as, for instance, $n^* = H \cdot n$, with a certain integer $H \geq 1$, can be adopted to improve estimates precision. In most available applications of II, the value of $H$ has usually been chosen between 10 and 100, and the main reason is that for larger $H$, the estimator variance of the indirect estimator is smaller. Whether a much larger value of $H$ is used (between 500 and 1000), it is possible to obtain a smoother function $\widetilde{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$, numerically closer to a continuous function.

DRAFT

(3) **Estimation**: Estimate of $\widehat{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}(\boldsymbol{\Sigma}))$ is obtained as solution of the optimization problem

$$\min_{\boldsymbol{\Sigma}} ||(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^{(r)}(\boldsymbol{\Sigma}))||.$$

It is worth noting that the arbitrary choice of the Euclidean norm is possible because the parameter vectors of the true and auxiliary models have the same dimension. Whenever the number of parameters of the auxiliary model is greater than the one of the true model, the estimation is dependent on the choice of the norm induced by a positive definite matrix. The tentative values for the true model parameters $\boldsymbol{\Sigma}$ are chosen iteratively until the equality $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^{(r)}(\boldsymbol{\Sigma}))$ is fulfilled. This iterative procedure for choosing tentative values of $\boldsymbol{\Sigma}$ is called *calibration*.

Generally, an analytic solution for the optimization problem in the Estimation step does not exist and the II estimator cannot be written in closed form. However, as proposed by Calzolari et al., 1999 (3), the problem can be solved numerically, by adopting as updating equation at the $i$-th step

$$\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Sigma}}_{i-1} - \delta * \left(Jac(\boldsymbol{\beta}(\boldsymbol{\Sigma}))|_{\widehat{\boldsymbol{\Sigma}}_{i-1}}\right)^{-1} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}_{i-1})).$$

where

- $\widehat{\boldsymbol{\Sigma}}_i$ is the value of the calibrated parameters after $i$ iterations;
- $\delta$ is a scale constant in $[0, 1]$, which reduces the step size in the given direction;
- $\left(Jac(\boldsymbol{\beta}(\boldsymbol{\Sigma}))|_{\widehat{\boldsymbol{\Sigma}}_{i-1}}\right)^{-1}$ is the inverse of the Jacobian matrix, evaluated in the $(i-1)$-th parameter estimate and it determine the direction of the $i$-th step.
  The derivatives are computed numerically by finite difference method.

For a better understanding of the iterative implementation, the numerical procedure is summarized in Figure 1.

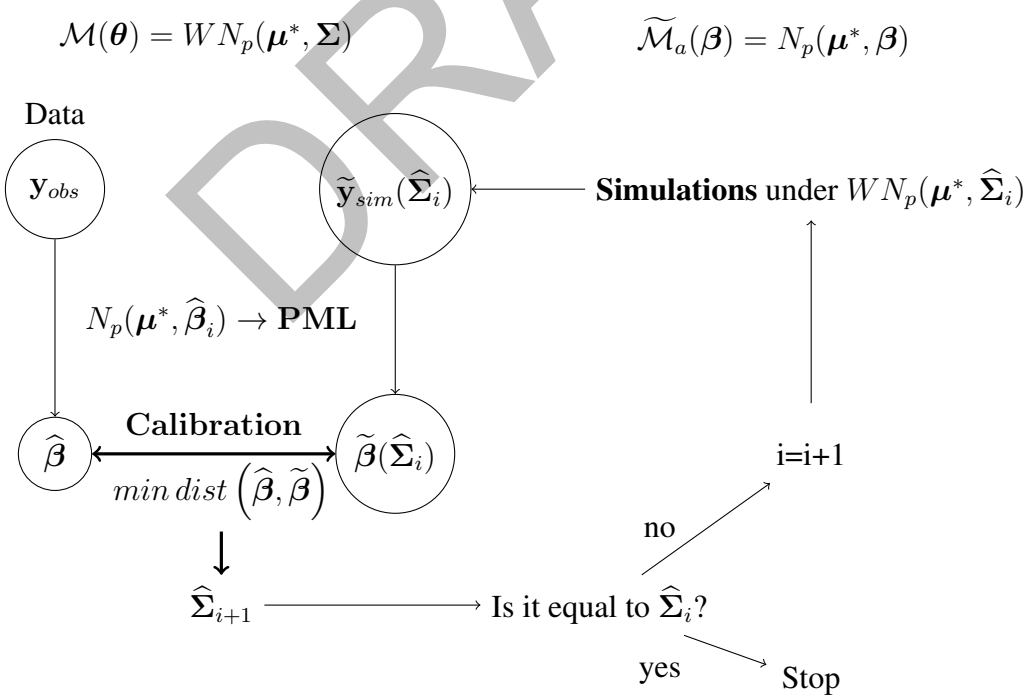



Figure 1: Diagram of the Indirect Inference estimation procedure for the variance-covariance matrix of a $p$-variate Wrapped Normal distribution, when using the standard $p$-variate Normal distribution as auxiliary model

DRAFT

## 4. Final remarks

We propose a novel approach to estimate the parameters of univariate and multivariate Wrapped Normal models via Indirect Inference with a Normal distribution as the auxiliary model. The II procedure explained in Section 3.is well-defined for all sample space dimensions, and has a low computational cost. Consequently, it can be used in a general context when the model of interest is a generic multivariate Wrapped Normal. The II estimators are similar to the auxiliary model estimators for very small variances and small $p$. Thus the convergence is sure. However, a direct approximation of the WN distribution based on truncated infinite series is sufficient. Nevertheless, computation becomes fastly infeasible when the dimension of the truncated sum increases. In such situations, the II estimators are a good and fast alternative to the estimator based on the approximate likelihood. For large variances, it is not guaranteed that the II calibration procedure is coming to convergence. This issue is also related to a structural problem implied by the construction of the WN distribution: the distribution is not distinguishable from a uniform distribution on the hypertorus, and its parameters tend to be undefined.

## References


[1] Agostinelli, C.: Robust estimation for circular data. Ann. Mat. Pura. Appl. **51**, 5867–5875 (2007)

[2] Bee, M.: Estimating the wrapped stable distribution via indirect inference. Commun. Stat. - Simul. Comput., **51** (11), 6371–6387 (2022)

[3] Calzolari, G., Di Iorio, F., Fiorentini, G.: Indirect estimation of just-identified models with control variates. Quad. DiSIA. **46**, (1999)

[4] Calzolari, G., Halbleib, R., Parrini, A.: Estimating GARCH-type models with symmetric stable innovations: Indirect inference versus maximum likelihood. CSDA **76**, 158–171 (2014)

[5] Calzolari, G., Halbleib, R.: Estimating stable latent factor models by indirect inference. J. Econom., **205** (1), 280–301 (2018)

[6] Coles, S.: Inference for circular distributions and processes. Stat. Comput. **8**, 105–113 (1998)

[7] Ferrari, C.:The wrapping approach for circular data Bayesian modeling. VDM (2009) ISBN: 9783639279993

[8] Fisher, N.I., Lee, A.J.: Time Series Analysis of Circular Data. R. Stat. Soc. Series B Stat. Methodol. **56**, 327–339 (1994)

[9] Gourieroux, C., Monfort, A., Renault, E.: Indirect Inference. J. Appl. Econ. **8**, 85–118 (1993)

[10] Jona-Lasinio, G., Gelfand, A., Jona-Lasinio, M.: Spatial analysis of wave direction data using wrapped Gaussian processes. Ann. Appl. Stat. **6** (4) 1478 – 1498 (2012)

[11] Kurz, G., Hanebeck, U.D.: Parameter estimation for the bivariate wrapped normal distribution, 2015 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, 2015, pp. 1192-1198, doi: 10.1109/CDC.2015.7402373.

[12] Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley and Sons. (2000) ISBN: 0471953334

[13] Mardia, K.V., Kent, J.T., Laha, A.K.: Score matching estimators for directional distributions. (2016) doi: 10.48550/ARXIV.1604.08470
https://arxiv.org/abs/1604.08470

[14] Nodehi, A., Golalizadeh, M., Maadooliat, M. et al.: Estimation of parameters in multivariate wrapped models for data on a p-torus. Comput. Stat. (2021) doi: 10.1007/s00180-020-01006-x

[15] Ravindran, P., Ghosh, S.: Bayesian analysis of circular data using wrapped distributions. Stat. Theory Pract. **5**, 547–561 (2011)

[16] Selvitella, A.: On geometric probability distributions on the torus with applications to molecular biology. EJS **13** (2), 2717–2763 (2019)

[17] Yu, S., Drton, M., Shojaie, A.: Generalized score matching for general domains. IMA Inf. Inference **11** (2), 739–780 (2022)


DRAFT

# A Comparison of Distribution-Free Control Charts

Michele Scagliarini[a]

[a] Department of Statistical Sciences, University of Bologna; michele.scagliarini@unibo.it

**Abstract**

In this study the statistical properties of several non-parametric control charts are investigated and compared. It is considered the problem of monitoring data streams when the distributional form of the monitored data is the generalized inverse Gaussian distribution. Two simulation studies are performed to assess the performance of the monitoring algorithms in various scenarios. The aim is to identify the most suitable monitoring algorithm considering jointly the ability in detecting shifts in location and/or scale and the percentage of missed alarms.

**Keywords:** change detection, control charts, non-parametric tests; simulation experiments

DRAFT

## 1. Introduction

Distribution-free control charts have received increasing attention in non-manufacturing fields because they can be used without any assumption on the distribution of the data to be monitored. This feature makes them particularly suitable for monitoring environmental phenomena often characterized by highly skewed distribution. In this work, we compare using Monte Carlo simulations, the performance of several non-parametric change point control charts for monitoring data distributed according the Generalised Inverse Gaussian (GIG) distribution. The choice of the GIG distribution is motivated by the fact that on the one hand it is often used to describe environmental radioactivity data, but on the other hand it has never been considered in connection with non-parametric control charts. For our purposes, aware of being non-exhaustive, we consider: the non-parametric change-point control chart proposed by [2] which is based on the Mann-Whitney (MW) statistic; the distribution-free control chart, based on Recursive Segmentation and Permutation (RS/P) proposed by [1]; the two non-parametric control charts based on the Kolmogorov-Smirnov (KS) and Cramer-von-Mises (CvM) statistics proposed by [4].

## 2. Methodology

In the Change Point Model (CPM) framework, it is assumed that the time ordered observations $x_1$, $x_2$,... are generated by the random variables $X_1$, $X_2$,... with unknown distribution functions $F_1$, $F_2$,... , respectively. A change point occurs at instant $\tau$ when $F_\tau \neq F_{\tau+1}$ and it is usually assumed that the observations are independent and identically distributed between every pair of change points. Therefore, the distribution of the sequence can be described by the following model: $X_i \sim F_0$ if $0 < i \leq \tau_1$, $X_i \sim F_1$ if $\tau_1 < i \leq \tau_2, \ldots,$ $X_i \sim F_k$ if $\tau_k < i \leq m$, where $0 < i < \tau_1 < \tau_2 < \cdots < \tau_k < m$ denote $k$ unknown change points. Within this framework, it is of interest to test $H_0$: $k$=0, versus $H_1$: $k \neq 0$ , to estimate the number of change points, $k$, and to estimate their locations $\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_k$ .

For testing the above described hypothesis system [4] proposed two change point control charts designed to detect arbitrary changes in the process distribution. The first is based on the Cramer-von-

Mises test and the second uses the Kolmogorov-Smirnov statistic. Both the control charts test for a change point immediately following any observation $x_k$ by partitioning the observations into two samples, $S_1=(x_1,\ldots,x_k)$ and $S_2=(x_{k+1},\ldots,x_t)$, and comparing the corresponding empirical distribution functions $\hat{F}_{S_1}(x)$ and $\hat{F}_{S_2}(x)$. The CvM test uses a statistic based on the square of the average distance between the empirical distributions $W_{k,t}=\int_{-\infty}^{\infty}\left|\hat{F}_{S_1}(x_i)-\hat{F}_{S_2}(x_i)\right|^2 dF_t(x)$, while the KS test uses a statistic defined as the maximum difference between the empirical distributions $D_{k,t}=\sup_x\left|\hat{F}_{S_1}(x_i)-\hat{F}_{S_2}(x_i)\right|$. Further details can be found in [4].

To detect location (mean) shifts [2] proposed a control chart based on the Mann-Whitney $U$-statistic: $U_{k,n}=\sum_{i=1}^{k}\sum_{j=k+1}^{n}D_{ij}$ for $1\le k\le n-1$, where $D_{ij}=\operatorname{sgn}(X_i-X_j)=\{1 \text{ if } X_i>X_j;0 \text{ if } X_i=X_j;-1 \text{ if } X_i<X_j\}$. The variance of $U_{k,n}$ depends on the split point $k$, so for this reason it is suitably standardized, thus obtaining the statistic $T_{k,n}$ [2]. Therefore, the test for the presence of a change point and the estimate of its time of occurrence are given by maximizing $T_{k,n}$ over $k$: $T_{\max,n}=\max_{1\le k\le n-1}\left|T_{k,n}\right|$. For further details see [2].

The RS/P control chart [1] allows to detect location and scale (variability) shifts. Initially, a set of elementary test statistics, $T_k$, designed to detect the possible shifts are calculated. For detecting changes in the process location, the statistics $T_k$ and the possible change points are computed using a forward recursive-segmentation approach. The algorithm starts with $k$=0 and proceeds in $K$ successive stages. At the beginning of stage $k$, the interval [1, $m$] is partitioned into $k$ subintervals, each having a length greater or equal to $l_{MIN}$. The quantities $K$ and $l_{MIN}$ can be appropriately chosen by the researcher [1]. At stage $k$, one of these subintervals is split, adding a new potential change point. The new change point is selected maximizing the function $f_k=\sum_{i=1}^{k+1}(\hat{\tau}_i-\hat{\tau}_{i-1})\left(\bar{x}(\hat{\tau}_{i-1},\hat{\tau}_i)-\bar{\bar{x}}\right)^2$, conditionally on the results of the previous stages, and the control statistic $T_k$ is equal to the attained maximum value of function $f_k$. Here $0=\hat{\tau}_0<\hat{\tau}_1<\ldots<\hat{\tau}_k<\tau_{k+1}=m$ are the boundaries of the new partition and $\bar{x}(a,b)=\sum_{i=a+1}^{b}\bar{x}_i\Big/(b-a)$. Then, these statistics are standardized and aggregated to obtain an overall statistic and the $p$-value is computed using a permutation approach. For detecting changes in scale, the steps are the same as for the process location, but the function $f_k$, used for the recursive segmentation, is different. Complete details can be found in [1].

DRAFT

## 3. Simulation Studies and Discussion

Environmental radioactivity data are usually positive and their distribution is often right skewed [5]. To analytically model these data good results have been obtained with the generalized inverse Gaussian distribution. The generalized inverse Gaussian distribution [3] has density function given by $f(x)=\frac{(\psi/\chi)^{\frac{\lambda}{2}}}{2K_\lambda\left(\sqrt{\psi\chi}\right)}x^{\lambda-1}e^{-\frac{1}{2}\left(\chi x^{-1}+\psi x\right)}$ for $x>0$, where $K_\lambda()$ is the modified Bessel function of the third kind with order $\lambda$. It is denoted with $N^{-1}(\lambda,\chi,\psi)$ and the domain of variation of the parameters is given by $\lambda\in R$, $(\chi,\psi)\in\Theta_\lambda$ where: $\Theta_\lambda=\{(\chi,\psi):\chi\ge 0,\psi>0\}$ if $\lambda>0$; $\Theta_\lambda=\{(\chi,\psi):\chi>0,\psi>0\}$ if $\lambda=0$; $\Theta_\lambda=\{(\chi,\psi):\chi>0,\psi\ge 0\}$ if $\lambda<0$.

As far as the control charts is concerned, there are several measures for assessing their statistical properties. The most commonly used measure is the Average Run Length (ARL). The Run Length (RL) of a control chart is a discrete random variable that is defined as the number of plotted statistics before an out-of-control point is observed on the chart. The ARL is the expected value of this random variable. The control charts introduced in the previous Section can be designed to achieve the same in-control

average run length, $ARL_0$ [1,2, 4]. Thus, to allow a fair comparison, we set $ARL_0$=200 for all the monitoring algorithms and focus our attention on their out-of-control performance.

Without loss of any generality, we consider as an initial benchmark the situation where the data come from a GIG distribution with parameters equal to: $\chi_0 = 0.596400$, $\psi_0 = 0.048177$ and $\lambda_0 = -0.738610$. In the first simulation study, for each run of the $k$=5·10$^4$ replications, we generate $n$ observations: the first $m$ observations are generated from $X_0 \sim N^{-1}(\lambda_0, \chi_0, \psi_0)$ while the observations from $m$+1 to $n$ are generated as $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_0) + \delta$ (location shift) and $X_1 \sim \varepsilon \cdot N^{-1}(\lambda_0, \chi_0, \psi_0)$ (scale shift), with $\delta$=1, 1.5, 2.5 and $\varepsilon$=1.5, 2, 2.5, respectively. In our case, without loss of any generality we set $m$=50 and $n$=400. The first observation >$m$ for which the control chart signals an alarm is the out-of-control run length ($RL_1$). The average of the $k$ run lengths is our estimate of the out-of-control average run length ($ARL_1$). In each run, if at the end of the $n$ observations the control chart did not detect any alarm, we defined this event as a missed alarm. Therefore, we compute for each control chart the percentage of missed alarms obtained in the $k$ replications. The results are summarised in Tables 1 and 2, where for each control chart and each simulated scenario the following quantities are reported: %missAL, the percentage of missed alarms; $ARL_1$, the estimated average length; SDRL, the standard deviation of the $RL_1$.

In the second simulation study the observations from $m$+1 to $n$ are generated with specific changes in the parameters. More specifically, as far as the parameter $\chi$ is concerned, we considered cases where the "in-control" parameter $\chi_0$ had been increased by 50%, 100% and 150%, while the other parameters remain unchanged. Therefore, the data from observations $m$+1 to $n$ are generated from $X_1 \sim N^{-1}(\lambda_0, \chi_1, \psi_0)$ with $\chi_1$ equal to $1.5\chi_0$, $2\chi_0$ and $2.5\chi_0$. For the parameter $\psi$ we considered cases where the "in-control" parameter $\psi_0$ had been decreased by 25%, 50%, 75%, while the other parameters remain unchanged. Therefore, the data from observations $m$+1 to $n$ are generated from $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_1)$ with $\psi_1$ equal to $0.75\psi_0$, $0.5\psi_0$ and $0.25\psi_0$. For the parameter $\lambda$ we considered cases where the "in-control" parameter $\lambda_0$ had been increased by 25%, 50% and 75%, while the other parameters remain unchanged. Therefore, the data from observations $m$+1 to $n$ are generated from $X_1 \sim N^{-1}(\lambda_1, \chi_0, \psi_0)$ with $\lambda_1$ equal to $0.75\lambda_0$, $0.5\lambda_0$ and $0.25\lambda_0$. The results are summarised in Tables 3-5.

Table 1: Simulation results for $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_0) + \delta$

| | RS/P (location) | RS/P (scale) | MW | CvM | KS |
|---|---|---|---|---|---|
| | | | $\delta$=1.0 | | |
| **%missAL** | 8.3 | 83.5 | 0.0 | 0.0 | 0.0 |
| **$ARL_1$** | 2.2 | 83.9 | 4.2 | 3.9 | 3.0 |
| **SD** | 19.6 | 94.6 | 1.1 | 0.9 | 0.7 |
| | | | $\delta$=1.5 | | |
| **%missAL** | 0.8 | 89.0 | 0.0 | 0.0 | 0.0 |
| **$ARL_1$** | 1.1 | 117.6 | 3.9 | 3.6 | 2.9 |
| **SD** | 5.6 | 97.1 | 0.9 | 0.8 | 0.5 |
| | | | $\delta$=2.0 | | |
| **%missAL** | 0.1 | 90.3 | 0.0 | 0.0 | 0.0 |
| **$ARL_1$** | 1.0 | 122.6 | 3.7 | 3.5 | 2.8 |
| **SD** | 2.3 | 97.4 | 0.8 | 0.7 | 0.5 |

Before commenting on the results, it is important to mention that we are aware that the scenarios examined in the simulations, although quite extensive, do not include all possible cases.

With these reservations in mind, the results of the first simulation study reveal that the RS/P control chart has good performance in terms of $ARL_1$ for detecting location shifts (see the values with a shaded background in Table 1). However, for $\delta$=1 it has a relatively high percentage of missed alarms (8.3%) if compared with the MW, CvM and KS control charts. For scale changes (Table 2) the results show that for all the shift magnitudes considered ($\varepsilon$=1.5, $\varepsilon$=2 and 2.5) the KS monitoring algorithm has the best performance both in terms of $ARL_1$ and missed alarms (values with a shaded background in Table

DRAFT

2). Furthermore, the MW control chart gives slightly better performance when compared with the CvM counterpart. It can also be noted that for small shifts ($\varepsilon$=1.5), both MW and CvM have relatively high percentages of missed alarms (9.1% and 9.8%) if compared with the KS control chart (2.4%). Surprisingly, the RS/P monitoring algorithm, even if it has been specifically designed to detect scale changes, gives the worst performance.

Table 2: Simulation results for $X_1 \sim \varepsilon \cdot N^{-1}(\lambda_0, \chi_0, \psi_0)$

| | RS/P (location) | RS/P (scale) | MW | CvM | KS |
|---|---|---|---|---|---|
| | | | $\varepsilon$=1.5 | | |
| **%missAL** | 95.7 | 94.4 | 9.1 | 9.8 | 2.4 |
| **ARL$_1$** | 189.7 | 116.4 | 86.4 | 89.7 | 64.6 |
| **SD** | 127.5 | 98.2 | 83.5 | 85.2 | 69.2 |
| | | | $\varepsilon$=2.0 | | |
| **%missAL** | 95.0 | 93.5 | 1.8 | 2.1 | 0.4 |
| **ARL$_1$** | 183.6 | 109.1 | 43.2 | 47.0 | 33.8 |
| **SD** | 132.2 | 98.9 | 52.9 | 56.8 | 42.7 |
| | | | $\varepsilon$=2.5 | | |
| **%missAL** | 94.2 | 92.4 | 0.02 | 0.04 | 0.01 |
| **ARL$_1$** | 158.0 | 89.1 | 15.8 | 16.8 | 13.8 |
| **SD** | 136.1 | 90.3 | 16.3 | 17.6 | 14.5 |

Table 3: Simulation results for changes in the parameter $\chi$

| | RS/P (location) | RS/P (scale) | MW | CvM | KS |
|---|---|---|---|---|---|
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_1, \psi_0)$ with $\chi_1 = 1.5\chi_0$ | | |
| **%missAL** | 96.0 | 94.7 | 10.4 | 10.8 | 2.8 |
| **ARL$_1$** | 199.9 | 121.0 | 93.3 | 95.9 | 69.3 |
| **SD** | 123.4 | 99.5 | 87.0 | 88.1 | 72.8 |
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_1, \psi_0)$ with $\chi_1 = 2\chi_0$ | | |
| **%missAL** | 95.6 | 94.4 | 3.2 | 3.5 | 0.7 |
| **ARL$_1$** | 180.7 | 115.8 | 51.4 | 54.6 | 39.3 |
| **SD** | 129.8 | 99.8 | 60.7 | 63.6 | 48.9 |
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_1, \psi_0)$ with $\chi_1 = 2.5\chi_0$ | | |
| **%missAL** | 95.3 | 94.1 | 0.5 | 0.7 | 0.1 |
| **ARL$_1$** | 165.7 | 107.4 | 29.8 | 31.4 | 23.5 |
| **SD** | 132.2 | 97.3 | 37.7 | 39.7 | 29.6 |

Table 4: Simulation results for changes in the parameter $\psi$

| | RS/P (location) | RS/P (scale) | MW | CvM | KS |
|---|---|---|---|---|---|
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_1)$ with $\psi_1 = 0.75\psi_0$ | | |
| **%missAL** | 95.7 | 94.5 | 16.0 | 15.9 | 7.3 |
| **ARL$_1$** | 202.0 | 116.6 | 123.3 | 122.9 | 105.8 |
| **SD** | 122.7 | 96.1 | 94.2 | 94.3 | 87.1 |
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_1)$ with $\psi_1 = 0.50\psi_0$ | | |
| **%missAL** | 95.9 | 94.5 | 15.2 | 15.2 | 6.2 |
| **ARL$_1$** | 205.4 | 112.5 | 120.6 | 121.1 | 102.5 |
| **SD** | 122.0 | 93.0 | 93.2 | 93.8 | 86.2 |
| | | | $X_1 \sim N^{-1}(\lambda_0, \chi_0, \psi_1)$ with $\psi_1 = 0.25\psi_0$ | | |
| **%missAL** | 95.7 | 94.1 | 14.3 | 14.5 | 5.7 |
| **ARL$_1$** | 208.1 | 107.7 | 117.4 | 118.5 | 98.2 |
| **SD** | 122.6 | 93.2 | 92.9 | 93.3 | 84.8 |

DRAFT

Table 5: Simulation results for changes in the parameter $\lambda$

| | RS/P (location) | RS/P (scale) | MW | CvM | KS |
|---|---|---|---|---|---|
| $X_1 \sim N^{-1}(\lambda_1, \chi_0, \psi_0)$ with $\lambda_1 = 0.75\lambda_0$ | | | | | |
| **%missAL** | 95.7 | 94.2 | 9.7 | 10.0 | 2.7 |
| **$ARL_1$** | 185.0 | 107.8 | 93.9 | 96.1 | 69.7 |
| **SD** | 125.3 | 93.1 | 86.9 | 87.9 | 72.3 |
| $X_1 \sim N^{-1}(\lambda_1, \chi_0, \psi_0)$ with $\lambda_1 = 0.50\lambda_0$ | | | | | |
| **%missAL** | 94.2 | 92.2 | 1.5 | 1.9 | 0.3 |
| **$ARL_1$** | 152.1 | 93.8 | 43.0 | 46.5 | 32.0 |
| **SD** | 131.4 | 93.3 | 52.6 | 56.2 | 40.2 |
| $X_1 \sim N^{-1}(\lambda_1, \chi_0, \psi_0)$ with $\lambda_1 = 0.25\lambda_0$ | | | | | |
| **%missAL** | 90.9 | 88.8 | 0.02 | 0.04 | 0.01 |
| **$ARL_1$** | 100.5 | 69.4 | 16.7 | 18.0 | 14.1 |
| **SD** | 122.4 | 85.0 | 17.4 | 19.5 | 15.2 |

In the second simulation study (Tables 3-5), where specific changes in the distribution parameters have been imposed, the KS change point model control chart has the best statistical properties both in terms of $ARL_1$ and percentage of missed alarms (values with a shaded background in Tables 3-5). Furthermore, the performance of the MW and CvM control charts is very similar across all change magnitudes. As an example, for the case where the parameter $\psi$ had been decreased by 75% (Table 4 for $\psi_1 = 0.25\psi_0$) the estimated $ARL_1$ for the KS monitoring algorithm is 98.2 (%missAL=5.7), whereas the $ARL_1$ for the CvM, MW, RS/P (scale) and RSP/location are 118.5 (%missAL=14.5), 117.4 (%missAL=14.3), 107.7 (%missAL=94.1) and 208.1 (%missAL=95.7), respectively. It can be noted that detecting changes in the distribution parameters seems a very difficult task for the RS/P control chart since for all the considered scenarios it has a very high percentage of missed alarms.

Summarising, the monitoring algorithm based on recursive segmentation and permutation has the best performance for detecting moderate shifts ($\delta$=1.5, 2.5) in the location. However, its statistical properties are not as good for the other scenarios examined. On the whole the Kolmogorov-Smirnov control chart provides the best results both in terms of out-of-control ARL and missed alarms.

DRAFT

## References


[1] Capizzi, G., Masarotto, G.: Phase I Distribution-Free Analysis of Univariate Data. J. Qual. Technol. **45**(3), 273--284 (2013) doi: 10.1080/00224065.2013.11917938

[2] Hawkins, D.M., Deng, Q.: A Nonparametric Change-Point Control Chart. J. Qual. Technol. **42**(2), 165--173(2010). Doi: 10.1080/00224065.2010.11917814

[3] Jørgensen, B: Statistical Properties of the Generalized Inverse Gaussian Distribution. In Lecture Notes in Statistics. Vol. 9. Springer-Verlag, New York, Berlin (1982)

[4] Ross, G.J., Adams, N.M.: Two Nonparametric Control Charts for Detecting Arbitrary Distribution Changes. J. Qal. Technol. **44**(12), 102--116(2012) doi: 10.1080/00224065.2012.11917887

[5] Scagliarini, M., Gualdi, R., Ottaviano, G., Rizzo, A., Padoani, F.: A Distribution-Free Approach for Detecting Radioxenon Anomalous Concentrations. In Perna, C., Salvati, N., Schirippa Spagnolo, F. (eds.) Book of Short Papers SIS 2021. Pearson, pp. 872-877. (2021). ISBN: 9788891927361.

# Characterizing Heterogeneity of Causal Effects in Air Pollution in Florida

Dafne Zorzetto[a]

[a]Department of Statistics, University of Padova, Italy; dafne.zorzetto@phd.unipd.it

**Abstract**

A main actual interest, in epidemiological studies, is environmental justice –i.e. to identify vulnerability/resilience to air pollution, taking into account the unfair difference among people with different races, national origin, and/or social-economic status [3]. In several studies, the causal link between long-term exposure to fine particulate and mortality risk is been studied and this work wants to identify how this causal effect changes among different population subgroups. This work analyses data on air quality in Florida with the goal of estimating the causal effects of fine particulate matter on mortality rates, leveraging the flexibility of a recently proposed Bayesian nonparametric mixture model specification. Five population subgroups, with different levels of vulnerability/resilience, are discovered and characterized by socio-economic status and ethnicity composition.

***Keywords:*** Air Pollution Epidemiology, Bayesian Nonparametrics, Causal Inference, Cluster-Specific Characteristics, Heterogeneous Causal Effects.

DRAFT

## 1. Introduction

Several epidemiological studies have provided significant evidence that long-term exposure to fine particulate matter ($PM_{2.5}$) increases mortality risk [see, e.g., 2; 5; 13; 16; 17]. However, it is now in the spotlight the importance of achieve the environmental justice, as the the Environmental Protection Agency (EPA) has declared [3]. This is mean, that health regulations and policies take into account the unfair difference in air pollution vulnerability among people with different race/ethnicity, national origin, and/or social-economic status.

Indeed, variable as race (e.g. white, black, hispanic, etc), national origin, age, sex, and/or social-economic status seem to play an explanatory role on identification of population subgroups, characterized by different level of vulnerability/resilience to air pollution [see, e.g., 4; 6]

In causal inference, the estimation of different causal effects, conditional to values of covariates or specific subgroups, is unified in the heterogeneous causal effect literature [see the reviws 1; 15]. In particular, the focus is on the estimation of the Conditional Average Treatment Effect (CATE), i.e. the expectation of the causal effect (defined as a function of the individual potential outcomes) for the units allocated in pre-specified groups of the population.

While, most of the approaches for estimating the CATE require defining covariates values *a priori*, with consequent limitations due to the subjective choices, [18] recently proposed a fully Bayesian model enabling us to identify heterogeneous and mutually exclusive population subgroups defined by shared CATEs, and discover the group-specific characteristics. The approach presented in [18] leverages on the flexibility of Bayesian Nonparametric approach, specifically using a Dependent Dirichlet Mixture Model [7; 9] specification.

In this work, we adopt the approach presented in [18] and analyse the causal effect of $PM_{2.5}$ on mortality in Florida.

## 2. Assumptions and model specification

The potential outcome framework [11] assumes that each observed subjects can potentially be assigned to a treatment $T$. Notably, in case of binary variable $T_i \in \{0, 1\}$, $T_i = 1$ if the unit $i$ is assigned to the treatment group, otherwise $T_i = 0$ when the unit $i$ is assigned to the control group. The potential outcome for unit $i$ is defined as $\{Y_i(0), Y_i(1)\} \in \mathbb{R}^2$, for $i = 1, \ldots, n$. The vector $\{Y_i(0), Y_i(1)\}$ represents the collection of the two potential outcomes, specifically $Y_i(0)$ is the outcome when the unit $i$ is assigned to the control group while $Y_i(1)$ is the outcome when it is assigned to the treatment group. In practice, however, for $i = 1, \ldots, n$, we observe only $y_i$, that is the realization of the random variable $Y_i$ defined as

$$Y_i := (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1).$$

Conversely, we can not observe the realization $y_i^{mis}$ of the random variable $Y_i^{mis} := T_i \cdot Y_i(0) + (1 - T_i) \cdot Y_i(1)$. Additionally, we define $x_i$ the $p$-dimensional vector of subject-specific background characteristics.

Some assumptions are necessary to identify and estimate the causal effect, according with the potential outcome literature [12]. Specifically, Stable Unit Treatment Value Assumption (SUTVA)—i.e. there are no different versions of the treatment levels assigned to each unit and no interference among the units—and strong ignorability—i.e all units, in each group conditional on some covariates values, have a positive chance of receiving the treatment.

We leverages the flexibility of the probit stick-breaking process [10] and adopt an infinite mixture specification for the probability density distribution of the outcomes. Details are reported in [18].

In the subsequent analysis we focus on estimating the Conditional Average Risk Ratio (CARR), defined as

$$\tau(x) := \frac{\mathbb{E}[Y_i(1) \mid X_i = x]}{\mathbb{E}[Y_i(0) \mid X_i = x]}, \tag{1}$$

where the value of $x$ are identified by the discovered groups.

DRAFT

## 3. Causal Effects of Air Pollution in Florida

The analysis used the ZIP codes level data recorded in Florida (USA). In particular, for each ZIP code (i.e. a statistical unit) the following variables are available: average of $PM_{2.5}$ levels during the years 2010, mortality rate in the 5 follow-up years, census variables such as the percentage of residents for different races/ethnicities (in particular, categorized as Hispanics, blacks, whites, and other races), percentage of men/women, age average among the Medicaid enrolees (people older than 65), population densities, and percentage of people who are eligible for Medicaid (this variable is a proxy of low social-economic status, abbreviate in S.E.S.).

The population density, in Florida during 2010, is represented in the first left map in Fig. 1 and it is clear how the population is concentrate around the main cities, as Jacksonville, Miami, Tampa, and Orlando. Consequently, we consider the ZIP codes with a population density significantly different to zero, and for them we reported in the second and third map in Fig. 1 the $PM_{2.5}$ level, recorded during 2010, and the mortality rate in the 5 follow-up years. The distribution of $PM_{2.5}$ among the state, seem to increase in the north particularly, due to the morphology of this peninsula, and close to the main cities, but with less intensively. While the mortality rate doesn't show any particular pattern.

### 3.1 Results

We define the exposure variable as $T = 1$ if the average $PM_{2.5}$ in 2010 is above the threshold $10\mu g/m^3$ and $T = 0$ otherwise. The choice of $10\mu g/m^3$ as a threshold aligns with the trend of National Ambient Air Quality Standard (NAAQS) established by the EPA [3].

The model estimation identifies 5 different subgroups: 3 with a negative effect of the exposure to $PM_{2.5}$ on the mortality rate, i.e. a high exposure to $PM_{2.5}$ increases the mortality rate, and 2 groups with a positive effect, i.e. a high exposure to $PM_{2.5}$ decrease the mortality rate.

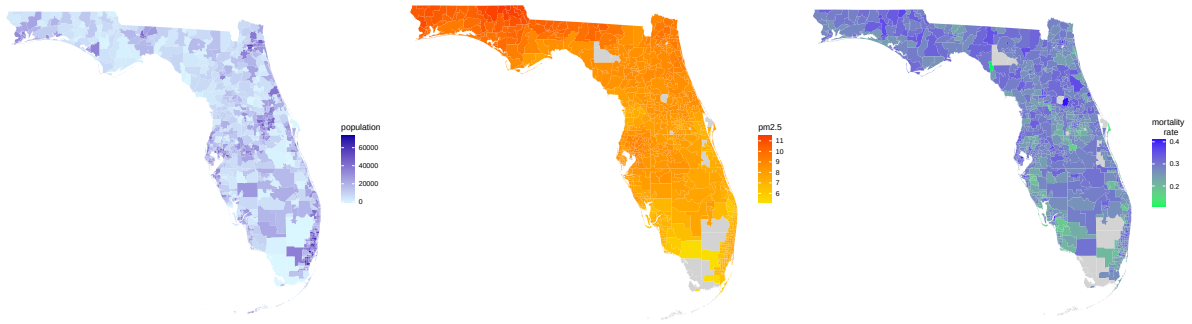
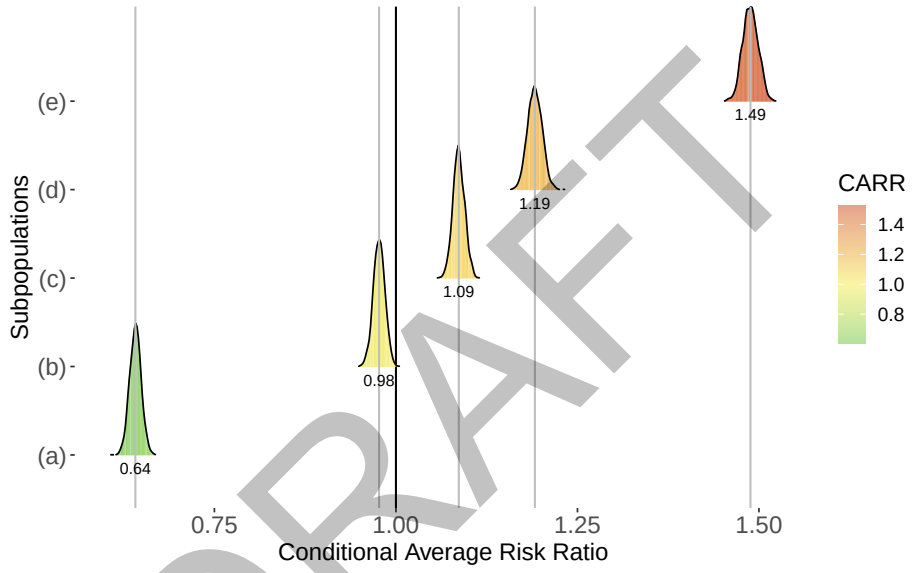



Figure 1: Florida information by ZIP code level: (i) Population in 2010; (ii) Recorded $PM_{2.5}$ during the years 2010; (iii) Mortality rate in the 5 follow-up years.




DRAFT

Figure 2: Posterior Conditional Average Risk Ratio (CARR) for the identified subgroups. The black line identifies the CARR = 1, i.e. the null causal effect. The gray lines are the mean of each posterior distribution.

Figure 2 presents the posterior distribution of the CARR for each identified subgroup. The vertical black line in the figure represents the null causal effect, which is indicated by a CARR equal to 1. The first three subgroups, which exhibit negative effects, have CARR values greater than 1 (in order from highest to lowest: (e) 1.49, (d) 1.19, (c) 1.09), while the two subgroups with positive effects have CARR values less than 1 (in order from lowest to highest: (a) 0.64, (b) 0.98).

By the definition of CARR—in (1)—, the causal effect of the $PM_{2.5}$ exposure varies among the groups, with a increment of up to $49\%$ of the mortality rate when the sub-population in (e) is exposed to high level of $PM_{2.5}$ instead of a lower level, and a decrement of $36\%$ of the mortality rate when the sub-population in (a) is exposed to high level of $PM_{2.5}$ instead of a lower level. However, these more extreme subgroups include both $1\%$ of the ZIP codes, while the majority of the population is included in the groups with less powerful effects: the group (d) includes the $37\%$ of the ZIP codes, where the mortality rate of the population increases by $19\%$ under high level of $PM_{2.5}$, the main group (c), with $52\%$ of the ZIP codes, has an increment of $9\%$, and the group (b) includes the $9\%$ of the ZIP codes, where the mortality rate of the population decrease by $2\%$ under high level of $PM_{2.5}$.

The different vulnerability/resilience to $PM_{2.5}$ across the identify subgroups, can be understood analysing the various compositions of the population among the ZIP codes. Indeed, the Fig. 3 shows

Figure 3: Group-specific covariates. Each spider plot reports in the colored area the group-specific characteristics (the mean of the analyses covariates) and in the gray area the collective characteristics (the mean of the covariates among the analyzed Californian ZIP codes).

DRAFT

that each group has different means of the analysed characteristics (reported with colours in the spider-plots) from the means of the same characteristics among all the consider ZIP codes (reported in gray in each spider-plot). Specifically, the spider-plots compare the following variables: sex (close to the center indicates a bigger percentage of women in the population of the ZIP codes, far to the center a bigger percentage of men), percentage of white, hispanic, black and other races (where smaller percentages are closer to the center), age (where the ZIP codes with the age mean close to 65 years for Medicaid enrollees are close to the center, and older population far from the center), and S.E.S. (close the center the population with high income and far from the center lower income the opposite).

Starting from the most vulnerable group, we can recognize that the group (e) is characterized by hispanic and black poor women, identify in literature as one of the most vulnerable population [4]. Additionally, the group (d) is distinguished by minorities as black and other races women. Actually, minorities, as hispanics, lack, or other no white races, are structurally exposed, over time, to higher levels of air pollution, becoming subjects more vulnerable and with more risk of mortality [4]. Moreover, the vulnerable group (c) is characterize by old men with a income lower than the average mean.

In juxtaposition, the groups characterized by positive effects, i.e. high exposure to $PM_{2.5}$ decrease the mortality rate, are mainly composed by rich young whites. In particular, the two groups differ in the sex: women in the group (a) and men in the group (b). This behavior is particular interesting, and it could be clarified by the potential survival bias [see, e.g., 8; 14], usually linked to younger people. Basically, cohort studies that start later leading to vulnerable individuals in certain subgroups dying before entering the cohort, therefore the individuals entering the cohort are the most resilient ones and might depict a decreasing mortality effect even when exposed to higher levels of pollutant.

## References

[1] Dominici, F., Bargagli-Stoffi, F. J., and Mealli, F.: From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework. Harvard Data Science Review (2021)

[2] Dominici, F., Greenstone, M., and Sunstein, C. R.: Particulate matter matters. Science **344**, 257-259 (2014)
[3] U.S. Environmental Protection Agency. Regulatory impact analysis for the proposed reconsideration of the national ambient air quality standards for particulate matter. In Technical Report: EPA-452/P-22-001 (2022)
[4] Jbaily, A., Zhou, X., Liu, J., Lee, T.-H., Kamareddine, L., Verguet, S., and Dominici, F.: Air pollution exposure disparities across us population and income groups. Nature **601**, 228-233 (2022)
[5] Lee, K., Small, D. S., and Dominici, F.: Discovering heterogeneous exposure effects using randomization inference in air pollution studies. Journal of the American Statistical Association **116**, 569-580 (2021)
[6] Liu, M., Saari, R. K., Zhou, G., Li, J., Han, L., and Liu, X.: Recent trends in premature mortality and health disparities attributable to ambient pm2. 5 exposure in china: 2005-2017. Environmental Pollution **279**, 116882 (2021)
[7] MacEachern, S. N.: Dependent dirichlet processes. technical report. Department of Statistics, The Ohio State University, Columbus, OH..(2000)
[8] Mayeda, E. R., Filshtein, T. J., Tripodis, Y., Glymour, M. M., and Gross, A. L.: Does selective survival before study enrolment attenuate estimated effects of education on rate of cognitive decline in older adults? a simulation approach for quantifying survival bias in life course epidemiology. International Journal of Epidemiology **47**, 1507-1517 (2018)
[9] Quintana, F. A., Mueller, P., Jara, A., and MacEachern, S. N.: The dependent dirichlet process and related models. arXiv preprint arXiv:2007.06129 (2020)
[10] Rodriguez, A. and Dunson, D. B.: Nonparametric Bayesian models through probit stick-breaking processes. Bayesian analysis **6**, (2011)
[11] Rubin, D. B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66**, 688, (1974)
[12] Rubin, D. B.: Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American statistical association **75**, 591-593 (1980)
[13] Rückerl, R., Schneider, A., Breitner, S., Cyrys, J., and Peters, A.: Health effects of particulate air pollution: a review of epidemiological evidence. Inhalation toxicology **23**, 555-592 (2011)
[14] Shaw, C., Hayes-Larson, E., Glymour, M. M., Dufouil, C., Hohman, T. J., Whitmer, R. A., Kobayashi, L. C., Brookmeyer, R., and Mayeda, E. R.: Evaluation of selective survival and sex/gender differences in dementia incidence using a simulation model. JAMA network open **4**, e211001-e211001 (2021)
[15] Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B.: Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. **37**, 3309-3324 (2018)
[16] Wu, X., Braun, D., Kioumourtzoglou,M.-A., Choirat, C., Di, Q., and Dominici, F.: Causal inference in the context of an error prone exposure: air pollution and mortality. The annals of applied statistics **13**, 520 (2019)
[17] Wu, X., Braun, D., Schwartz, J., Kioumourtzoglou, M., and Dominici, F.: Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. Science advances **6**, eaba5692 (2020)
[18] Zorzetto, D., Bargagli-Stoffi, F.J., Canale, A., and Dominici, F.: Confounder-Dependent Bayesian Mixture Model: Characterizing Heterogeneity of Causal Effects in Air Pollution Epidemiology. Technical report (2023)

DRAFT

# Comparing three robust procedures for CANDECOMP/PARAFAC estimation

Valentin Todorov[a], Violetta Simonacci[b], Michele Gallo[c], and Nikolay Trendafilov[c]

[a]UNIDO, Vienna, Austria; valentin@todorov.at
[b]University of Naples Federico II, Naples, Italy; violetta.simonacci@unina.it
[c]University of Naples-L'Orientale, Naples, Italy; mgallo@unior.it, ntrendafilov@unior.it

DRAFT

**Abstract**

CANDECOMP/PARAFAC aims to identify the true components underlying data with a trilinear configuration. The search for a unique solution is not always an easy task, as degeneracies may occur. The presence of outlier contamination further complicates the matter by requiring the implementation of robust procedures. The most used robust approach R-ALS is based on the iterative repetition of the standard alternating least squares algorithm, which is known to be slow and vulnerable to over-factoring, collinearity, and bad initial values. Here the faster and stable robust alternative R-INT1, based on the SWATLD-ALS integrated scheme INT-1, is implemented. Its performance is tested against ALS, R-ALS, and R-INT2 (built on INT-2, an ATLD-ALS procedure already proposed in the literature). Performance is assessed in a simulation study with varied levels of outlier contamination.

***Keywords:*** ALS, ATLD-ALS, SWATLD-ALS, outliers, computational efficiency

## 1. Introduction

Three-way data sets collect observations on a set of variables measured over several occasions (locations, times, conditions) and are represented as three-dimensional arrays rather than data matrices, as it is usual in the multivariate data analysis. Different techniques exist to analyze such three-way data but CANDECOMP/PARAFAC (CP), which can be seen as a generalization of principal component analysis (PCA) to higher-order tensors is one of the most popular. The idea of CP is to find a given number of components that jointly represent the data well.

The usual way of parameter estimation in CP is an alternating least squares (ALS) procedure which yields least-squares solutions and provides consistent outcomes. It is well-known that algorithms which rely on least squares easily break down in the presence of outliers. This is well recognized for PCA and a number of robust alternatives were proposed in the literature. In the multivariate case where two-way data are analyzed, the outliers are assumed to be rows (observations, objects, subjects, etc.) in the data set which lie significantly far from the other observations. Similarly, in the three-way case, we can assume that outliers are matrices (slices) that have a profile strongly deviating from the rest. (3) have demonstrated the influence of outlying samples on the classical ALS. They split the observations into four groups: regular observations, good leverage points, bad leverage points and residual outliers constructing a plot, the outlier map, similar to the one in robust regression or in robust PCA on two-way

data. To cope with the presence of outlying samples they have also proposed a robust version of the ALS algorithm, R-ALS, which relies on the robust PCA method ROBPCA (5).

Apart from its proneness to outlying samples, the standard ALS-PARAFAC procedure suffers several major flaws which might be particularly problematic for large-scale problems: slow convergence and sensitiveness to degeneracy conditions such as over-factoring, collinearity, bad initialization and local minima. A lot of research was invested to find a solution to these problems and a number of improved versions of ALS were created. Several alternatives to ALS were developed, like the alternating trilinear decomposition (ATLD) (10), self-weighted trilinear decomposition (SWATLD) (2) and their properties and comparative performances have been studied in several works, e.g. (9). These alternative procedures resist temporary degeneracies and over-factoring problems while ensuring a much speedier estimation process than ALS, thanks to a steeper convergence curve. These advantages are obtained at the cost of losing the stability of results and obtaining non-least-squares solutions. As a possible fix, an algorithm integration strategy was introduced in (6; 7) to combine the benefits of faster procedures with ALS stability.

The robust version of ALS proposed by (3) also shares these issues. In this approach, standard ALS is iteratively executed; thus, the effect of these disadvantages will be multiplied causing the whole procedure to become very slow, especially in the case of large data sets. The ATLD-based procedure INT2 (7) was extended to a robust version R-INT2 by (8) and its superior performance was demonstrated in an extensive simulation study and experimental data. Following these lines, in this paper we propose to robustify also the INT1 procedure based on SWATLD and call it R-INT1. We study its performance in the case of data without contamination or with increasing levels of contamination and compare to the other two known robust procedures.

DRAFT

## 2. The CP model, the ALS algorithm and its integrated versions

The CP model (1; 4) decomposes the 3-way data array $\underline{\boldsymbol{X}}(I \times J \times K)$ with a generic element $x_{ijk}$ into the three loading matrices $\boldsymbol{A}$ $(I \times R)$, $\boldsymbol{B}$ $(J \times R)$, $\boldsymbol{C}$ $(K \times R)$ with $R$ components (using the same number for each mode). The CP model can be written formally as

$$\boldsymbol{X}_A = \boldsymbol{A}(\boldsymbol{C} \otimes \boldsymbol{B})^\top + \boldsymbol{E}_A, \tag{1}$$

where $\boldsymbol{X}_A$ and $\boldsymbol{E}_A$ are the original array and the error array unfolded with respect to mode A and the symbol $\otimes$ represents the *Kronecker product* between two matrices. To estimate the optimal component matrices the residual sum of squares

$$||\boldsymbol{E}_A||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2 = \sum_{i=1}^{I}||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2 = \sum_{i=1}^{I} RD_i^2 \tag{2}$$

is minimized. The residual distance (RD) for observation $i$ is thus given by

$$RD_i = ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i|| = \sqrt{\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2} \tag{3}$$

and the estimation is equivalent to the minimization of the sum of the squared distances. With ALS the component matrices are estimated one at a time, keeping the estimates of the other component matrices fixed, i.e. we start with initial estimates of $\boldsymbol{B}$ and $\boldsymbol{C}$ and find an estimate for $\boldsymbol{A}$ conditional on $\boldsymbol{B}$ and $\boldsymbol{C}$ by minimizing the objective function. Estimates for $\boldsymbol{B}$ and $\boldsymbol{C}$ are found analogously. The iteration continues until the relative change in the model fit is smaller than a predefined constant.

As already mentioned in the introduction, while ALS is still the algorithm of choice because of its many desirable characteristics, its use can be problematic especially in cases of large data sets due to slow convergence and sensitiveness to degeneracy conditions such as over-factoring, collinearity, bad initialization and local minima. The alternating trilinear decomposition (ATLD) proposed by (10) seems

to be the most efficient method among the proposed alternatives to ALS and it is claimed to be less sensitive than ALS to over-factoring. It is based on the use of three loss functions with different response surfaces. SWATLD (2) does not attempt to find the minimum of (2), instead, it alternates between minimizing three different (non-least squares) loss functions, one per each of the loading matrices. It is important that the three matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ must have full column rank in order for the algorithm to resolve uniquely the components of interest. SWATLD seems to be the best algorithm in terms of recovery capability (factor congruence) (9).

However, these advantages of ATLD and SWATLD are obtained at the cost of unstable results and non-least squares solutions. To cope with these disadvantages a recent research development demonstrated that an integrated approach, combining algorithms with complementary points of strength, could provide a suitable solution. Two integrated algorithms INT-1 (6) and INT-2 (7) were proposed which combine SWATLD and ATLD steps with ALS, respectively, to ensure faster convergence, stability, and insensitivity to wrong model specification. For both integrated procedures the authors demonstrated the gain in performance in terms of computational efficiency and resistance to different undesirable effects. However, it is not known which of them is to be preferred in different situations since no comparison between them was conducted.

## 3. The robust alternatives to ALS

The idea of a robust version of CP proposed by (3) is to identify enough "good" observations and to perform the classical ALS on these observations. This is repeated until no significant change is observed. Finally, a reweighting step is carried out to improve the efficiency of the estimators. To identify the "good" observations a robust version of PCA, e.g. ROBPCA (5), is used on the unfolded array. We will call this procedure R-ALS in the rest of the paper. It is obvious that the robust procedure will be much more time-consuming than the classical one, repeating many times the ALS optimization. Therefore, any improvement of the parameter estimation procedure will contribute to the improvement of the performance of the complete robust procedure. R-ALS is entirely based on ALS and thus suffers the slow convergence and other disadvantages of this algorithm. (8) proposed to replace ALS by INT2 thus obtaining a new robust estimation procedure which they called R-INT2. As in R-ALS, it starts with robust principal components to identify any outlying points and then iterates using the INT2 algorithm until no significant change is observed. After convergence, a reweighting step with INT2 is conducted which produces the final solution.

Since the integrated procedure (6) based on SWATLD was demonstrated to have also very good performance as an alternative to ALS, we suggest extending it to a robust version in the same way as it was done for R-INT2 in (8). To verify that it can cope with outlying samples in the data and outperform R-ALS we conduct a simulation survey which is presented in Section 4. The purpose of this simulation study is also to find out which of the two integrated robust procedures performs better.

## 4. Simulation study

We will study the performance of the newly proposed procedure R-INT1 for robust estimation of trilinear CP models and will compare it to the classical ALS and the other two known robust procedures R-ALS and R-INT2 on a detailed simulation platform. Similarly as with the other integrated procedures the performance of the two-stage procedure R-INT1 will depend significantly on the transition parameters for switching from the initialization stage to the refinement stage. For this reason, the preliminary part of this simulation study is dedicated to the empirical estimation of these parameters, but due to the space limitation we will only state the final result - the values $10^{-2}$ and $10^{-3}$ seem to be most favorable and $10^{-2}$ will be used in all further computations. Successively we will compare the classical CP, the robust version based on ALS as proposed by Engelen and Hubert (3) R-ALS, the integrated robust version proposed by Todorov et al (8) R-INT2 and the newly proposed integrated procedure based on SWATLD R-INT1. First of all, we want to verify that R-INT1 and R-INT2 work well on data sets with

DRAFT

and without contamination by identifying the outliers at least as well as R-ALS, retrieving solutions with good statistical quality. At the same time, we want to verify that the convergence of these two procedures is improved significantly and thus the computational time is reduced. And finally, we want to compare the computational performance of R-INT1 and R-INT2 in different scenarios.

These aspects will be illustrated on three-way data generated as in (7), (9), (3) and (8). The three-way arrays have $I = 50$ observations, $J = 100$ variables and $K = 10$ occasions and the number of factors is $R = 3$ or $R = 5$. For each data set random matrices $\boldsymbol{A} \in \mathbb{R}^{I \times R}$, $\boldsymbol{B} \in \mathbb{R}^{J \times R}$ and $\boldsymbol{C} \in \mathbb{R}^{K \times R}$ are generated from a uniform distribution. With the so defined loadings matrices a three-way data set can be constructed according to Equation 1 with different levels of homoscedastic (HO) and heteroscedastic (HE) noise $\boldsymbol{E}_A^{HO}$ and $\boldsymbol{E}_A^{HE}$ respectively defined as in (9).

Different types of outliers were then added following the scheme proposed by (3) and used by (8). First of all, we want to study the behavior of the four procedures on clean data and thus, the first setup is created by not including any outliers. In the other configurations 10% and 20% of the observations are modified to contain *bad leverage points*. For each setup, in order to account for minor statistical fluctuations 100 replicates were conducted. For evaluating efficiency performance CPU time, iterations, and incidence of temporary degeneracies (swamps) are considered. For accuracy, the computed diagnostics include the value of the objective function (FIT), the occurrence of fault recoveries (FR), the mean square error (MSE), and the angle between the estimated and original subspaces of the second and third mode. See (9) and (8) for a full description.

Overall, at 20% contamination with bad leverage points, the difference between R-ALS FIT and R-INT1 FIT is higher than $1e^{-4}$ in less than 1% of the cases. In more than half of the cases (55.9%) the fit of R-INT1 is better than that of R-ALS which demonstrates that R-INT1 is capable of identifying the best low rank approximation as well as R-ALS. The results for R-INT2 are similar, as shown in (8).

It is important to verify that the known instabilities of ATLD and SWATLD are not passed to the integrated procedures. We can check this by looking at the percentage of fault recoveries reported for all three algorithms for different levels of contamination and different ranks in Table 1. For $R = 3$, both in the correct rank estimation case and when over-factoring, the percentage of fault recoveries for all robust methods is not higher than 1%. Only for the classical estimates on data with 10% and 20% contamination the percentage increases drastically, coming close to 100%. For R = 5 all percentages are slightly higher (not shown here).

DRAFT

| | | $F = R = 3$ | | | | $F = R + 1 = 4$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | C | RALS | RINT1 | RINT2 | C | RALS | RINT1 | RINT2 |
| FR | 0% | 0.0 | 0.0 | 0.0 | 0.1 | 1.2 | 1.1 | 0.0 | 0.0 |
| | 10% | 98.9 | 0.0 | 0.0 | 0.1 | 0.0 | 1.0 | 0.0 | 0.0 |
| | 20% | 99.4 | 0.0 | 0.2 | 0.1 | 0.0 | 0.9 | 0.0 | 0.0 |
| SWAMPS | 0% | 0 | 0 | 0 | 0 | 22 | 18 | 0 | 0 |
| | 10% | 0 | 0 | 0 | 0 | 1 | 25 | 0 | 1 |
| | 20% | 0 | 0 | 0 | 0 | 2 | 15 | 0 | 0 |

Table 1: Total percentages of FR and number of swamps (out of 4500 repetitions) by rank and number of factors for different levels of contamination with bad leverage points.

The next measure of accuracy to look at is the MSE. The results for the four estimators and three types of data (no outliers, 10% bad leverage points and 20% bad leverage points) are presented in the box plots in the left panel of Figure 1. The right panel of the same figure presents the angles of the B-loadings for the four methods and the three contamination types.

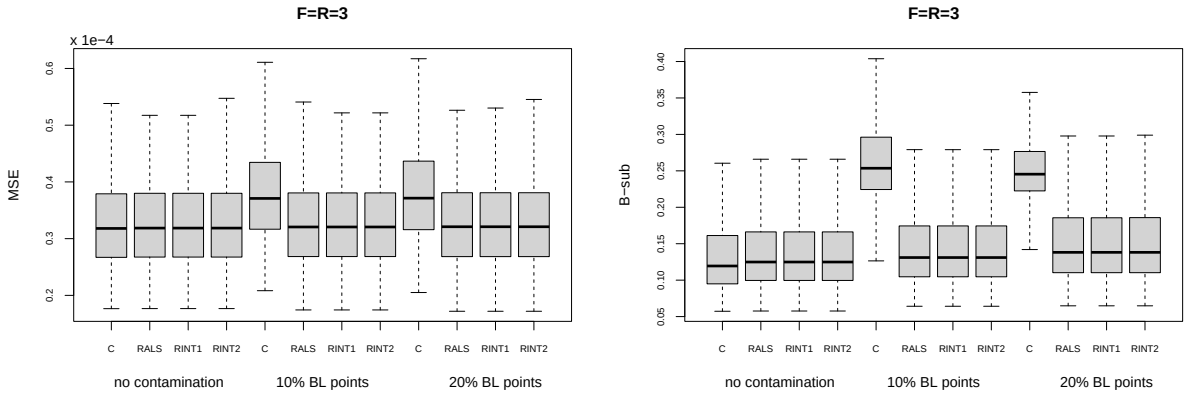



Figure 1: MSE values and angle of B-loadings of classical CP (C), robust CP with ALS (R-ALS), robust CP with INT1 (R-INT1) and robust CP with INT2 (R-INT2) on data sets without contamination, and with 10% and 20% bad leverage points respectively.

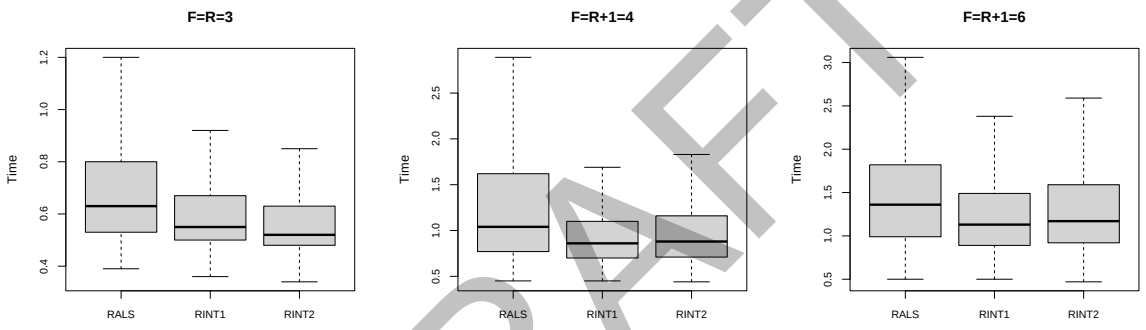



Figure 2: CPU time in seconds, robust CP with ALS (R-ALS), robust CP with INT1 (R-INT1) and robust CP with INT2 (R-INT2) on data sets 20% bad leverage points for different number of factors.

DRAFT

All four estimators perform equally well on clean data both in terms of MSE and maxsub, however, when outliers are added to the data (10% and 20%) the classical CP is influenced - the MSE increases and the quality of the fit of the loadings decreases. There is non much difference in the performance of the three robust methods in terms of MSE and maxsub. However, if we look at Fig. 2 which presents their performance in terms of computational time the gain in performance in the two integrated procedures is obvious. The integrated procedures R-INT1 and R-INT2 perform better than R-ALS both in terms of median time and variance, with R-INT2 being slightly better in the case of correct factor decomposition ($F = R = 3$, left panel). In over-factoring ($F = R + 1 = 4$ and $F = R + 1 = 6$, middle and right panels) the roles change and R-INT1 becomes better.

The computational efficiency can also be judged by counting the number of swamps, i.e. the temporary degeneracies which continue for more than 10 iterations and thus slowdown the procedure. This problem was not significantly manifested in our simulation. As seen in the lower part of Table 1, no swamp cases are observed when estimating the correct rank and when $R = 3$, for none of the estimators and for none of the contamination levels. Only several cases were observed when over-factoring, for the classical ALS and the robust R-ALS, however the number of these cases is insignificant when compared to the total number of 4500 repetitions.

## 5. Summary and conclusions

The simulation study shows that R-INT1 is a viable alternative to R-ALS for dealing with outliers in a three-way setting. Throughout scenarios, it is stable in converging to the correct parameters and does not appear to model excessive noise, reaching least squares solutions. The integrated strategy ensures higher efficiency and also proves useful in terms of accuracy when over-factoring. As far as the differences between R-INT1 and R-INT2 go, they are both solid procedures. R-INT2 is slightly more efficient in terms of median values, however, in presence of over-factoring, R-INT1 is the top performer. A thorough presentation of the resulting diagnostics, as well as an analysis carried out on an application will be provided to further strengthen the comparison and complete this work. Future work should also study the possibilities for combination with other computational algorithms. The behavior of the algorithms if collinearity is present will be of great interest as well as their extension with additional constraints.

## References


[1] Carroll J, Chang J (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrica 35(3):283–319

[2] Chen ZP, Wu HL, Jiang JH, Li Y and Yu RQ (2000) A novel trilinear decomposition algorithm for second-order linear calibration. Chemometrics and Intelligent Laboratory Systems 52 75–86.

[3] Engelen S, Hubert M (2011) Detecting outlying samples in a parallel factor analysis model. Analytica Chemica Acta 705:155–165

[4] Harshman RA (1970) Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. Tech. Rep. 10, UCLA

[5] Hubert M, Rousseeuw PJ and Vanden Branden K (2011) ROBPCA: A new approach to robust principal component analysis. Technometrics 47 64–79

[6] Simonacci V, Gallo M (2019) Improving PARAFAC-ALS estimates with a double optimization procedure. Chemometrics and Intelligent Laboratory Systems 192:103822.

[7] Simonacci V, Gallo M (2020) An ATLD—ALS method for the trilinear decomposition of large third-order tensors. Soft Computing 24:13535–13546

[8] Todorov V, Simonacci V, Gallo M and Trendafilov N (2023) A novel estimation procedure for robust CANDECOMP/PARAFAC model fitting. Submitted for publication.

[9] Tomasi G, Bro R (2006) A comparison of algorithms for fitting the PARAFAC model. Computational Statistics & Data Analysis 50(7):1700–1734

[10] Wu HL, Shibukawa M, Oguma K (1998) An alternating trilinear decomposition algorithm with application to calibration of HPLC-DAD for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. Journal of Chemometrics 12 1–26


DRAFT

# How active is a genetic pathway? Comparative analysis of post-hoc permutation-based methods

Anna Vesely[a] and Angela Andreella[b]

[a]Institute for Statistics, University of Bremen, Germany; vesely@uni-bremen.de
[b]Department of Economics, Ca' Foscari University of Venice, Italy;
angela.andreella@unive.it

**Abstract**

Procedures with true discovery guarantee, i.e., methods for simultaneous inference on the True Discovery Proportion (TDP), have become widely popular in many applications. They permit addressing the multiplicity problem while at the same time solving the spatial specificity paradox. Here we propose a comparative analysis of some of the most widely used permutation-based procedures: *sumSome*, *pARI*, *sansSouci* and *Notip*. We compare their performance on differential gene expression data analysis, where the interest lies in quantifying levels of activation in different pathways.

***Keywords:*** true discovery proportion, permutation testing, differential gene expression

DRAFT

## 1. Introduction

In recent years, simultaneous inference on the True Discovery Proportion (TDP) has become a popular inferential method. Its wide use is substantial since several application fields, including neuroscience and genetics, have to cope with the problem of multiple testing and the spatial specificity paradox. Indeed, in most analyses the inference is made not at the level of individual features but of groups of features (e.g., biological pathways in genetics and brain regions in neuroimaging) to have a less conservative correction for multiplicity. However, this leads to the so-called spatial specificity paradox: a significant p-value for a group only denotes that there is at least one significant feature, but does not give any information on the number of significant features nor their localization. Therefore, as the size of the group becomes larger, the finding becomes weaker (9).

Simultaneous inference on the TDP addresses the multiplicity problem while overcoming this paradox. The main idea is to define procedures with true discovery guarantee, i.e., methods that give lower confidence bounds for the TDP simultaneously over all possible groups. Simultaneity guarantees that the confidence bounds remain valid when the group of interest is chosen post-hoc and when doing follow-up inference inside a group (6). Several methods have been proposed in the literature; here we focus on the permutation framework, as it relies on minimal assumptions and often offers an improvement over the parametric approach (7).

This paper follows the idea presented in (12), where two permutation-based procedures with true discovery guarantee, *sumSome* (13) and *pARI* (2), are compared on brain imaging data. Here we provide a more comprehensive comparison, studying the performance of the above-mentioned methods, as well as other recent proposals: *sansSouci* (4) and *Notip* (3). We analyze the differential gene expression dataset studied in the supplementary material of (13). This type of data contains expression levels for different genes; interest lies in assessing differences between two sub-populations at the level of pathways, collections of genes associated with a specific biological process that interact with each other.

The manuscript is organized as follows. In Sect. 2, we formalize the concept of true discovery guarantee and briefly review the methods that will be used in the analysis. Then we discuss the dataset in Sect. 3. and present results in Sect. 4.

## 2. Permutation-based true discovery guarantee

Suppose that we are interested in studying $m$ univariate hypotheses $H_i$ with $i \in M = \{1, \dots, m\}$, with significance level $\alpha \in [0, 1)$. Denote with $T \subseteq M$ the subset of false hypotheses. Subsequently, for any non-empty subset $S \subseteq M$ define its TDP as $\pi(S) = |S \cap T|/|S|$, where $|\cdot|$ denotes the size of a set. Hence $\pi(S)$ represents the proportion of false hypotheses (true discoveries) in $S$. The set $T$ is unknown, and so is $\pi(S)$. In the case of differential gene expression data, $M$ represents the whole set of genes. A hypothesis $H_i$ is false when gene $i$ is active, i.e., differentially expressed between two populations. Finally, each subset $S$ represents a biological pathway, and $\pi(S)$ is the proportion of truly active genes inside.

To make inference on $\pi(S)$, we consider procedures with true discovery guarantee. They are defined as random functions $\bar{\pi} : 2^M \to \mathbb{R}$, where $2^M$ is the power set of $M$, such that

$$P\left(\pi(S) \geq \bar{\pi}(S) \text{ for each } S \subseteq M\right) \geq 1 - \alpha.$$

Hence $\bar{\pi}(S)$ is a lower $(1 - \alpha)$-confidence bound for $\pi(S)$, simultaneously over all possible subsets $S$; simultaneity ensures the validity of the confidence bounds even under post-hoc selection (6).

As argued in Sect. 1, in this paper we focus on the permutation framework. In this context, recent proposals for methods with true discovery guarantee are the following. *sumSome* (13) is an iterative method that relies on sum-based global tests, such as most p-value combinations, and converges to an admissible procedure after a finite, but possibly exponential in $m$, number of iterations. *pARI* (2) and *sansSouci* (4) both use critical vectors of ordered p-values, e.g., based on higher criticism (5) or Simes inequality (10). *pARI* is presented in two ways: a fast single-step version, similar to *sansSouci*, and an iterative version that uniformly improves both the single-step version and *sansSouci*. Since the gain in power is generally small, here we will consider the single-step version. Finally, *Notip* (3) further improves *pARI* and *sansSouci* through a modification of the critical vector that is data-dependent. The default version of the method requires to divide the data into two subsets, employing the first (training set) to define a family of data-dependent critical vectors, and the second (inference set) to select a suitable critical vector within the family and use it for inference. The Authors show through simulations that the method remains valid and is even more powerful if the same data is used for training and inference, provided that two independent rounds of randomization are performed; here we will focus on this second version.

With the exception of *Notip*, all the above-mentioned procedures are not single methods but families of methods, allowing for different choices of the setting: the sum test in *sumSome* and the family of critical vectors in the others. Different choices have different power properties, depending on many aspects of the problem, such as intensity of the signal, density, homogeneity, structure, etc. In the next sections, we will apply the methods to differential gene expression data. For each method, we choose the setting that we expect to be more powerful according to simulation results from the corresponding original paper.

Finally, we underline that each particular choice for the method/setting gives statements for all possible subsets of genes; so it may be more powerful than other choices for some pathways and less powerful for others. Depending on which pathways we are interested in, different choices may result to be preferable. Here we will comment on the overall behavior on all pathways.

## 3. Analysis of the breast invasive carcinoma data

In this section, we analyze the breast invasive carcinoma dataset from The Cancer Genome Atlas (TCGA) research network (https://www.cancer.gov/tcga). It contains gene expression data,

DRAFT

i.e., levels of each gene's product, for patients with different types of breast tumor. The analysis aims to assess differences between two histological types of primary solid tumor: infiltrating lobular carcinoma and infiltrating ductal carcinoma. After selecting patients in the two sub-populations of interest, we filter out the $10\%$ of genes with the lowest mean expression among subjects, obtaining expression values for $985$ subjects and $15{,}678$ genes. We study the whole set of genes as well as the $352$ pathways contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (8). The database contains a large pathway including $1{,}339$ genes, while the size of the others ranges from $3$ to $501$.

The analysis is carried out as following. For each gene, we consider the null hypothesis that the mean expression is the same in the two sub-populations, and we compute a p-value through a two-sided two-sample t-test. Then we use the p-values as input for the different procedures with true discovery guarantee, taking $\alpha = 0.05$ as significance level. Hence each procedure will return lower $0.95$-confidence bounds for the TDP of the considered gene sets.

For each method, the analysis is performed using the code provided by the Authors: the R packages (11) and (1) for *sumSome* and *pARI*, respectively, and the Python code available at `https://github.com/sanssouci-org/sanssouci.python` for the remaining methods. Each setting is chosen considering that gene expression data is generally characterized by dense signal and medium-high correlation between genes. Therefore in *sumSome* we use the harmonic mean p-value, with truncation of all p-values greater than $\alpha$ to $0.5$; the algorithm is run for at most $50$ iterations. In *pARI* and *sansSouci* we use critical vectors based on higher criticism (5) and Simes (10), respectively. The different choice is due to the fact that higher criticism is not implemented for *sansSouci*. Finally, the methods that rely on critical vectors require choosing the vector size (parameter $K$ in (4), and $k_{\max}$ in (3)). As there is no choice that is optimal in all scenarios and there are no guidelines for the particular case of gene expression data, we opt for the most intuitive alternative and take the size as the number of all genes. However, smaller sizes could lead to more powerful procedures.

Regarding the number $B$ of permutations, *Notip* requires a high number to obtain a suitable resolution when computing the data-driven critical vector. Therefore we take $B = 1000$ for the main analysis. For the other methods, however, we run additional analyses with $B = 200$ permutations to evaluate to which extent results are robust to this choice.

DRAFT

## 4. Results

First, we consider the main analysis, carried out with $1000$ permutations. All methods find activation, i.e., non-null TDPs, in the whole set of genes and in the same $344$ pathways out of $352$. Table 1 contains results for all genes and for some of the most active pathways. In particular, for each method we report the ten pathways having the highest TDP. Notice that there is a substantial, but not complete, overlap between the pathways that result to be most active in different methods; as a consequence, the table shows thirteen pathways in total.

Now we can compare the results given by different methods. For each pair of methods, Fig. 1 displays the relationship with an highlight on the size of the pathways, while Fig. 2 shows the boxplot of the differences in results. Results are generally homogeneous, with differences most likely dominated by the variability due to the random permutations. The greatest differences, with a maximum of $9.94\%$, tend to correspond to the smallest pathways, but there is not a clear pattern regarding size.

Finally, we examine results for *sumSome*, *pARI*, and *sansSouci* from the additional analyses that make use of $B = 200$ permutations instead of $1000$. In general, the only requirement on $B$ to have non-zero power in permutation-based tests is that $B \geq 1/\alpha$. However, low values of $B$ give lower mean power and more variable results due to the randomness of the permutations (7). The role of the choice of $B$ has been investigated for *sumSome* (13), for which the Authors suggest that $B = 200$ is generally sufficient to have suitable power. The outcome of this analysis seems to point in the same direction for all considered methods, as the TDP does not change significantly for the two choices of $B$ (Fig. 3).

To summarize, we have compared four permutation-based procedures with true discovery guarantee in the analysis of a gene expression dataset. The methods under study exhibit comparable powers, with differences primarily concentrated in the smallest pathways. We underline that we considered only one

Table 1: Analysis of pathways: name, size, and lower confidence bound for the TDP (%) for different methods. Results are shown for the whole gene set and for the ten most active pathways in each method

| $S$ | $\|S\|$ | $\bar{\pi}(S)$ (%) | | | |
|---|---|---|---|---|---|
| | | sumSome | pARI | sansSouci | Notip |
| all genes | 15,678 | 32.86 | 53.68 | 29.09 | 45.87 |
| hsa03450: non-homologous end-joining | 11 | 72.73 | 72.73 | 72.73 | 72.73 |
| hsa03050: proteasome | 44 | 70.46 | 68.18 | 68.18 | 70.45 |
| hsa04110: cell cycle | 120 | 55.83 | 55.83 | 55.00 | 55.83 |
| hsa03030: DNA replication | 36 | 55.56 | 58.33 | 55.56 | 58.33 |
| hsa03013: nucleocytoplasmic transport | 100 | 51.00 | 53.00 | 49.00 | 54.00 |
| hsa00900: terpenoid backbone biosynthesis | 22 | 50.00 | 50.00 | 50.00 | 50.00 |
| hsa03267: virion | 4 | 50.00 | 50.00 | 50.00 | 50.00 |
| hsa03008: ribosome biogenesis in Eukaryotes | 69 | 49.28 | 50.72 | 44.93 | 52.17 |
| hsa01210: oxocarboxylic acid metabolism | 17 | 47.06 | 47.06 | 47.06 | 47.06 |
| hsa00450: selenocompound metabolism | 13 | 46.15 | 46.15 | 46.15 | 46.15 |
| hsa03460: Fanconi anemia | 44 | 45.45 | 47.73 | 40.91 | 50.00 |
| hsa03060: protein export | 22 | 45.45 | 45.45 | 45.45 | 45.45 |
| hsa04141: protein processing | 159 | 45.91 | 46.54 | 44.65 | 47.17 |

DRAFT

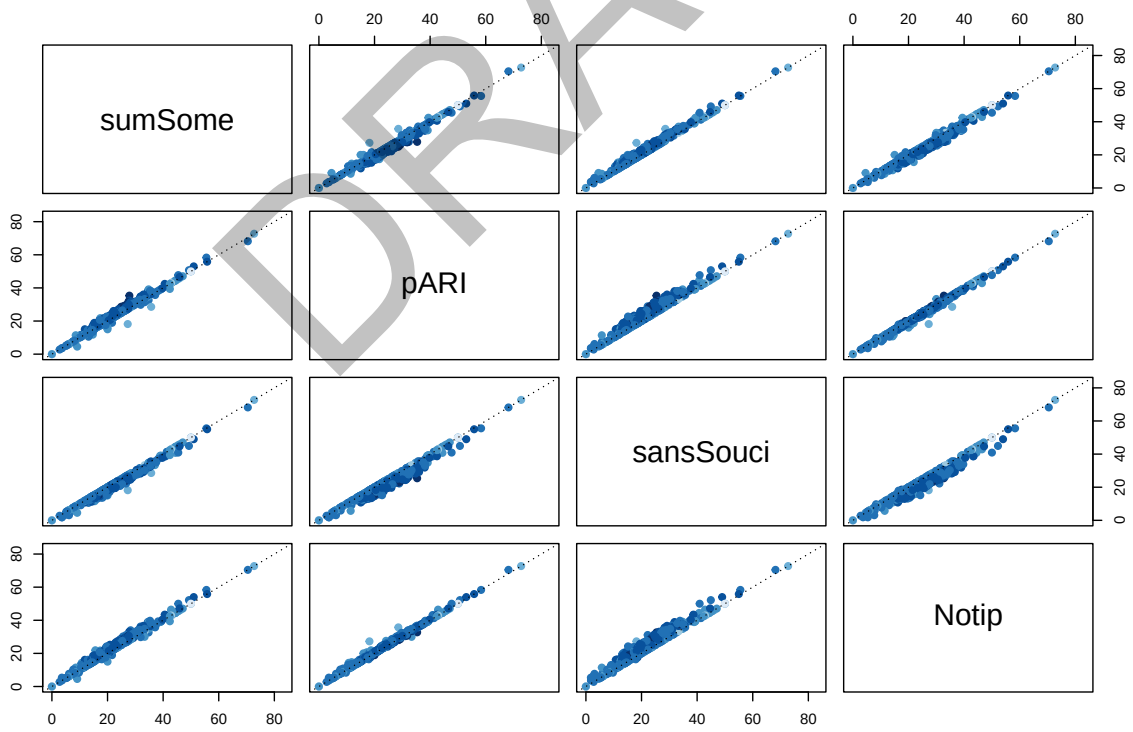



Figure 1: Lower confidence bounds for the TDP (%) of all pathways for different methods. The color of the points corresponds to the pathway size: a darker color indicates a higher number of genes

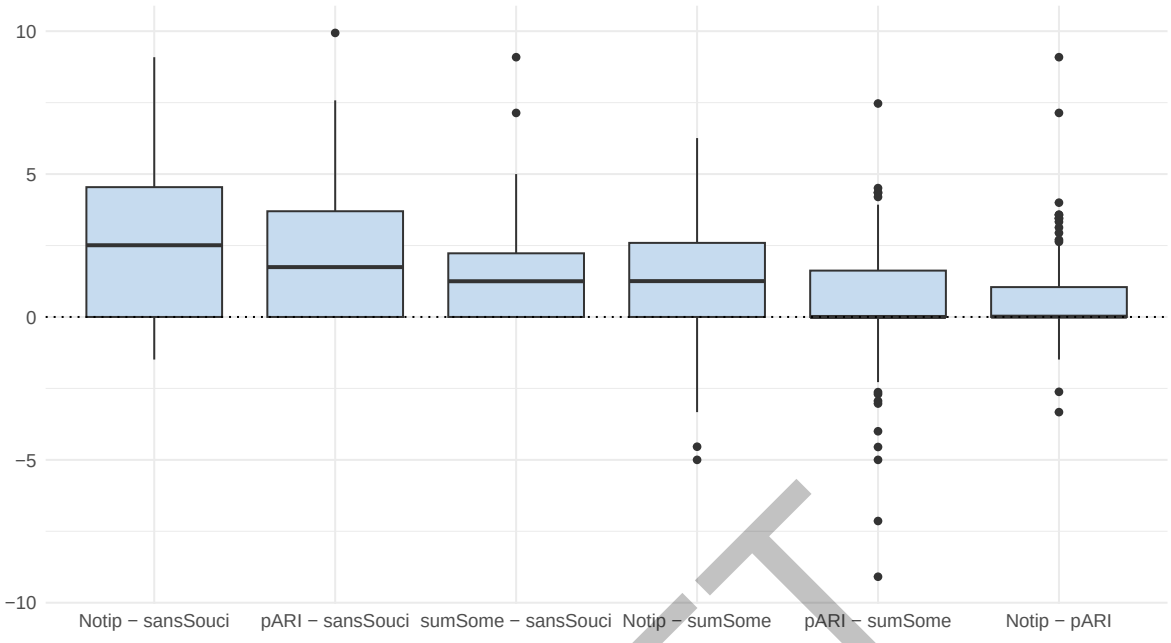



Figure 2: Differences between the lower confidence bounds for the TDP (%) given by different methods for the same pathways

DRAFT

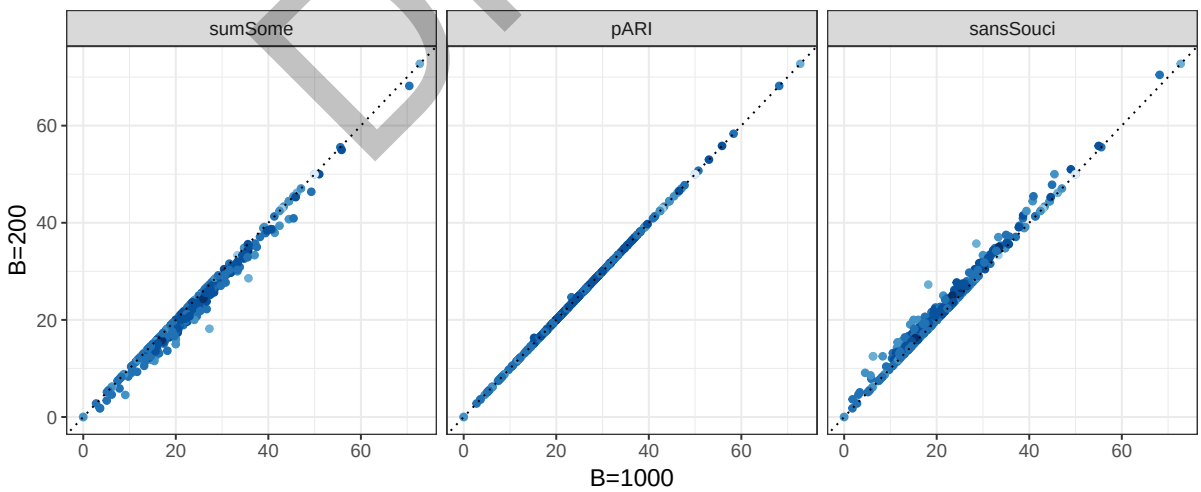



Figure 3: Lower confidence bounds for the TDP (%) of all pathways for different methods and different numbers of permutations $B$. The color of the points corresponds to the pathway size: a darker color indicates a higher number of genes

setting for each procedure (sum test for *sumSome*, family of critical vectors for *pARI* and *sansSouci*), but different choices could have displayed different power properties. Furthermore, we showed that *sumSome*, *pARI* and *sansSouci* can be used with fewer permutations (200 instead of the usual value 1000). This reduces the computational workload without leading to a significant loss of power.

## Acknowledgements

We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG) via Grant No. DI 1723/5-3, and by Ca' Foscari University of Venice via Grant No. PON 2014-2020/DM 1062. The analysis was carried out using the University of Padua Strategic Research Infrastructure Grant 2017: CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione, https://capri.dei.unipd.it.

## References

[1] Andreella, A.: pARI: permutation-based All-Resolutions Inference method. R package (2022) https://CRAN.R-project.org/package=pARI
[2] Andreella, A., Hemerik, J., Weeda W.D., Finos, L., Goeman, J.J.: Permutation-based true discovery proportions for fMRI cluster analysis. Statistics in Medicine, in press (2023)
[3] Blain, A., Thirion, B., Neuvial, P.: Notip: Non-parametric true discovery proportion control for brain imaging. NeuroImage **260**, 119492 (2022)
[4] Blanchard, G., Neuvial, P., Roquain, E.: Post hoc confidence bounds on false positives using reference families. Ann. Stat. **48**(3), 1281–1303 (2020)
[5] Donoho, D., Jin, J.: Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat. **32**(3), 962–994 (2004)
[6] Goeman, J.J., Solari, A.: Multiple testing for exploratory research. Stat. Sci. **26**(4), 584–597 (2011)
[7] Hemerik, J., Goeman, J.J.: False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. J. R. Stat. Soc. Series B Stat. Methodol. **80**(1), 137–155 (2018)
[8] Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **28**, 27–30 (2000)
[9] Rosenblatt, J.D., Finos, L., Weeda, W.D., Solari, A., Goeman, J.J.: All-Resolutions Inference for brain imaging. NeuroImage **181**, 786–796 (2018)
[10] Simes, J.R.: An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**(3), 751–754 (1986)
[11] Vesely, A.: sumSome: permutation true discovery guarantee by sum-based tests (2021) https://CRAN.R-project.org/package=sumSome
[12] Vesely, A., Finos, L., Goeman, J.J., Andreella, A.: Valid double-dipping via permutation-based closed testing. In: Perna, C., Salvati, N., Spagnolo, F.S. (eds.) Book of Short Papers SIS 2021, pp. 776–781. Pearson (2021)
[13] Vesely, A., Finos, L, Goeman, J.J.: Permutation-based true discovery guarantee by sum tests. J. R. Stat. Soc. Series B Stat. Methodol., qkad019 (2023)

DRAFT

# Non Parametric Combination methodology: a literature review on recent developments

Elena Barzizza[a], Nicolo' Biasetton[a], and Riccardo Ceccato[a]

[a]University of Padova, Department of Management Engineering;
elena.barzizza@phd.unipd.it, nicolo.biasetton@phd.unipd.it,
riccardo.ceccato.1@unipd.it

**Abstract**

The Non Parametric Combination (NPC) methodology is a permutation-based solution which can be successfully applied in many statistical analysis. It has many advantages including the fact that it does not require any distributional assumptions, it has few requirements (e.g. exchangeability), it can deal with different types of variables both for repeated measures and paired data, it is flexible in terms of choice of test statistic, and it implicitly takes into account dependency among variables. All these features make it suitable for application in many challenging contexts. The aim of this paper is to prove a literature review on recent developments on this topic since 2010.

***Keywords:*** Non Parametric combination methodology, literature review, applications, developments

DRAFT

## 1. Introduction

The Non Parametric Combination (NPC) methodology is a permutation-based technique. Being permutation-based, it does not require any distributional assumptions. This feature makes it quite competitive when compared to the parametric methodologies presented in the literature, i.e. it can be considered a good alternative to the parametric F-test or t-test which require an assumption of normality. In fact, in real applications data do not follow a known distribution and in the multivariate framework the assumption of normality is particularly hard to justify. This is true in many application fields, such as medicine, psychology and biology, both for survey data and observational studies. Additionally, NPC has multiple advantages that make it particularly exploitable in multivariate business contexts: it has few requirements, such as the exchangeability property, it can handle different types of data, such as continuous, categorical or mixed variables, both for repeated measures and paired data, it can be successfully applied in situations characterized by small sample size or high dimensional data, it can handle restricted alternative hypotheses or datasets characterized by missing values. NPC shows low type II error and good power performances even with small sample sizes (2). Furthermore, thanks to a permutation approach, the NPC methodology implicitly takes dependency among variables into account, both for dependent and independent samples. Finally, it is flexible: the most suitable system of hypotheses can be chosen depending on the nature of the problem. For more detail on technical aspects of the methodology, see Pesarin and Salmaso (30). Given all these promising characteristics, the NPC methodology appears to be an instrument of interest not only from a statistical point of view but also for practitioners. The main aim of this short paper, therefore, is to understand what developmental trends have been introduced for the NPC methodology since 2010. In particular, we want to understand: the main directions of development reported in the literature; the main content of the developments introduced; the methodology's main

fields of application; and what suggestions have been made for interesting future directions of study. To address the aforementioned points, we perform a systematic literature review of the Non Parametric Combination methodology with specific focus on most recent developments since 2010.

The paper is set out as follows: the next section explains the methodology used to conduct the literature review (Section 2); Section 3. presents the results; conclusions and final remarks are provided in Section 4.

## 2. Methodology

The review was carried out on two databases, namely Scopus and Web of Science, using a query made up of three main blocks: the first concerned the NPC methodology or variations of the term; the second concerned permutation tests; the third concerned a multivariate setting. All keywords were searched for in titles, abstracts and keywords. Filters were then used to select the papers: only papers in English, only papers published in the period 2010-2023 (for methodology prior to 2010, see Pesarin and Salmaso (2010) (30)) and only articles were considered. Papers from the two databases were merged and duplicates removed. We then read the abstracts and made an initial selection based on the information reported there. The resulting 27 papers were read in full and analysed.

## 3. Main findings

Of the 27 papers considered, 18 focus on theoretical developments (67%) and 9 focus on the application of the methodology in real contexts (33%). With regard to the papers on application of the methodology, the majority are linked to the field of medicine, confirming the assertion of Racioppi et al. (2015) that NPC is particularly suited to the biomedical field (33). The importance of NPC's applicability is further confirmed by the structure of the papers focused on the development of knowledge from a theoretical point of view. Of the 18 theoretical papers, 15 (which is approximately 83%) contain one or more case studies from datasets presented in the literature or from datasets collected by the authors. The presence of a case study in almost all the publications highlights the importance of the applicability of the methodology in real contexts. As such, the methodology is not only appealing from a theoretical and statistical perspective, but is also appealing to practitioners. The NPC methodology also appears to be particularly suited to analyzing survey data. Indeed, 4 of the 9 application papers deal with this scenario.

As stated before, the NPC methodology appears to be particularly suited to many application fields, such as biomedical, or in general to both observational and experimental studies (33) as well as to survey data (11). In all these scenarios the number of observational units is commonly lower than the number of considered variables. Two considerations of particular interest arise from this field. Firstly, in Pesarin et. al (2010) (31), the authors described a particular property of the NPC methodology, namely finite sample consistency. According to this property, by adding a variable while the sample size remains the same (preferably small), the power of the NPC methodology can monotonically increase. As a result, NPC is seen to be particularly suited to application in datasets characterized by a number of variables higher than the number of observational units: this type of configuration is known as thick data and is challenging to deal with. Langthaler et al. (2022) (24) further investigate the analysis of thick data by considering the most popular methods used to analyze this type of dataset, including the NPC methodology, and compare them by running a simulation study. Secondly, medical datasets are very frequently characterized by a high number of zero values and this can represent a challenge when it comes to performing a statistical analysis. The reason for this high presence of zero data is the fact that the field often includes censored data or data with measurements below the limit of the instruments used for the measurement. Arboretti et al. (2020) (8) focused on the application of NPC in this particular circumstance taking into consideration three different test statistics (Anderson-Darling, Mann-Whitney and difference of means) in a two-sample multivariate setting.

The property of finite sample consistency leads to another property, namely equipower, introduced

DRAFT

by Salmaso (2015) (34). This property states that, by adding informative variables while the sample size remains fixed, the power of the NPC-based test increases. This is also a very important feature of the NPC methodology from a computational point of view: it allows us to apply NPC in situations characterized by high dimensional datasets, which are typical in, for example, healthcare or genetic studies. Neuroimage data is a concrete example of a high dimensional dataset, particularly if we consider the high-resolution imaging of today. Winkler et al. (2016) (36) introduce a modified NPC solution to deal with this type of data with the aim of obtaining an improvement in power performances compared to traditional tools. Another feature of the NPC methodology is the impact of the presence of dependency structures in a dataset, in other words the presence of correlated variables. Salmaso (2015) (34) used a simulation study to analyze the impact of correlation in terms of power performances of NPC: the results reveal that the presence of correlation causes a loss in power performance. Similarly, Arboretti Giancristofaro et. al (2016) (4) investigated this aspect specifically for ordered categorical outcomes with the same results. The power performance of the NPC methodology is also influenced by the choice of combining function as studied in Arboretti Giancristofaro et. al (2016) (4). On a similar theme, Langthaler et al. (2022) (24) provide some recommendations for choosing the most suitable combining function. Alfieri et al. (2012) (1) proposed an NPC-based methodology to perform an iterated procedure on the choice of most appropriate combining function.

As said at the beginning, the NPC methodology appears to be particularly suitable for many application fields. Specifically, it seems to be regularly adopted in all statistical problems related to comparisons among $C > 2$ populations, from different points of view. Some examples of its application to compare $C > 2$ groups or populations are presented in Lanfranchi et al. (2020) (23), Barisan et al. (2015) (11), Fasolato et al. (2010) (20) and Alibrandi et al. (2022) (2). Corain et. al (2019) (19) applied NPC as a ranking procedure as did Montelli et. al (2016) (28). This represents a common problem in statistics which deal with the ability to rank $C$ populations according to $p > 2$ variables of interest. Arboretti et al. (2014) (5) and Corain et al. (2014) (17) focused on ordered categorical response variables and provided a solution within the NPC approach. Although the problem is widely studied in the literature, the solution provided by Arboretti et al. (2014) (5) is essentially different to all the other solutions, particularly in relation to its purpose, methodology and inference. This NPC-based ranking solution can be applied to different mixes of response variables as well as to thick data. Another way to study the ordering of $C$ is to use the stochastic ordering problem, which is slightly different. In this case the aim is to test the equality of distributions against a one-sided ordered alternative with at least one strict inequality. An example of the application of the NPC methodology to this problem is presented in Brombin et. al (2016) (13). For a theoretical application of the NPC methodology to the stochastic ordering problem, we refer the readers to contributions by Bazyari et al. (2013) (12) and Arboretti et al. (2021) (9), the latter in particular for the goodness-of-fit problem. The problem of comparing populations can also be approached from different angles. Comparing populations also means dealing with location, scale or symmetry testing in a multivariate scenario. For example, Antonucci et al. (2019) (3) applied the NPC methodology to test the mean difference and this is particularly true if the symmetry of distribution is around zero. Many authors have contributed to the development of methodologies to address these issues. The most popular multivariate parametric test in the literature for the location problem is Hotelling's $T^2$ test. Some authors also provided non-parametric solutions based on ranks (see for example Oja et al. (2004) (29)). In a multivariate non-parametric setting, other solutions were provided based on the notion of data depth - see for example Liu et al. (1993) (27), the T-based and M-based test by Li et al. (2004) (25), Chanvan et al. (2016) (14), Chenouri et al. (2012) (16), and Li et al. (2016) (26). However, even solutions within the NPC methodology framework are developed and presented in the literature. Chanvan et al. (2019) (15) presented a solution to simultaneously deal with location and scale by considering an NPC-based solution based on the notion of data depth. This is not the only development introduced in the last decade. Corain et al. (2015) (18) proposed some improvements to NPC to test the equality of mean vectors with a linear additive model by exploiting a multi-aspect strategy, an adding-variable strategy, and an iterated combination technique. Arboretti et al. (2017) (6) faced another peculiar problem: testing location shift among populations, which is useful for comparing, for example, two treatments. To do so they exploited an NPC-based methodology and extended the solution provided by Pesarin et al. (2016) (32) for a multivariate setting taking the union-intersection (UI) solution into account. A natural extension of

DRAFT

the multivariate location problem is the MANOVA problem; see, for example, Racioppi et al. 2015 (33). NPC methodology developments are also proposed in the literature to deal with this task; see Giancristofaro et al. (2012) (21) and Arboretti et al. (2018) (7). In the first, an extension was introduced to compare two or more treatments in a randomized complete block (RCB) design dealing with categorical response variables. In the second, the authors suggested an NPC-based methodology for a two-way design for the specific scenario of small sample size and high number of variables. In the literature there are also solutions to test specific symmetries among populations (e.g. bivariate, angular, spherical and central). Kalina (2021) (22) presents a test for general symmetry. The paper considered the NPC methodology for testing symmetry among multivariate populations after providing a list of robust multivariate estimators.

By exploiting the multi-aspect strategy mentioned above, the NPC methodology can be successfully applied in all situations which require the testing of a multidimensional hypothesis. Again this could be useful to resolve the issue of comparing $C > 2$ populations under different aspects of interest. To this end, Arboretti et al. (2022) (10) exploited the approach to resolve the goodness-of-fit problems in Arboretti et al. (2021) (9) and exploited multi-aspect testing to provide an NPC-based solution to test three population aspects: location, variability and cumulative distribution function.

In summary, from the literature review we have identified three main strands of development: insights into NPC properties; particular types of data dealt with by NPC; and comparisons among populations using NPC. If we consider the percentage of papers for each principle direction, we observe that development of knowledge in support of comparisons among populations is the one with the highest percentage (61.1%), i.e. this is the topic on which authors have concentrated efforts to propose solutions in the NPC field, either to solve some challenging tasks or to propose a better alternative to the solutions already present in the literature. The importance of this field is also confirmed by papers on the practical applications of the NPC methodology; all of them focus on the problem of comparing populations in a multivariate permutation framework. This is found to be quite a common problem in many different application fields (medical, sport, agricultural, biological, food, psychological, and so on). Therefore, we recommend further investigation of the application of the NPC methodology to solve the task of comparing multivariate populations and implementing appropriate changes. Another important aspect of interest is the fact this methodology can handle not only numerical variables, but also categorical or mixed variables. As mentioned above, some authors have focused their attention over the past decade on these types of variables in order to provide some solutions. Furthermore, the NPC methodology can be successfully applied in situations characterized by small sample size or high dimensional dataset. Given the development of information technologies in recent years which allow the collection of huge amounts of data, and given the higher quality and complexity of, for example, imaging data, the application of NPC may be a suitable solution which needs more attention in future years. Particular attention should be paid to thick data as they become more and more popular in many industrial contexts and the need grows for the right instruments to carry out their statistical analysis. Langthaler et al. (2022) (24) suggest an in-depth investigation of the ratio between informative and uninformative variables within the framework of thick data in order to investigate NPC's power performance. Additionally, the multi-aspect strategy which is possibly applicable in NPC methodology can be further exploited in common application contexts. For example, Arboretti et al. (2020) (8) suggested providing a multi-aspect solution within the NPC-based framework in situations characterized by zero-inflated data. Finally, with regard to properties of the NPC methodology, some aspects were investigated, such as finite sample consistency, equipower, and the impact of correlation and combining functions, however further aspects could be investigated especially in cases where exchangeability might not be fully assumed (e.g. second order exchangeability (35)).

DRAFT

## 4. Conclusion

This short paper looks at the Non Parametric Combination (NPC) methodology and in particular recent developments in the multivariate framework, the main directions of development since 2010, and its main applications to real contexts. To the best of our knowledge, a comprehensive literature review which explains and summarizes the most important recent developments and possible future developments of the methodology has not been carried out therefore the review presented in this short paper fills

this gap. The analysis highlighted three main directions of development, namely development of NPC properties, development to treat non-traditional types of data, and development to conduct statistical comparisons among populations. The large number of case studies in papers devoted to the theoretical development of NPC methodology together with analysis of the NPC application in real contexts emphasize its suitability to deal with real case studies. NPC appears to be particularly suitable for performing statistical analysis both in observational studies and on survey data, particularly in the biomedical field. Our literature review highlights also emerging fields of application as well as some possible directions of development.

## References


[1] Alfieri, R., Bonnini, S., Brombin, C., Castoro, C. & Salmaso, L. Iterated combination-based paired permutation tests to determine shape effects of chemotherapy in patients with esophageal cancer. *Statistical Methods In Medical Research*. **25**, 598-614 (2016)

[2] Alibrandi, A., Giacalone, M. & Zirilli, A. Psychological stress in nurses assisting Amyotrophic Lateral Sclerosis patients: a statistical analysis based on Non-Parametric Combination test. *Mediterranean Journal Of Clinical Psychology*. **10** (2022)

[3] Antonucci, L., Bolzan, M., Carrozzo, E., Crocetta, C., Di Gioia, L., Manacorda, M., Mastrangelo, F., Russo, M. & Salmaso, L. Accuracy of computer guided implant dentistry: A permutation testing approach. *Electronic Journal Of Applied Statistical Analysis*. **12**, 542-551 (2019)

[4] Arboretti Giancristofaro, R., Bonnini, S., Corain, L. & Salmaso, L. Dependency and truncated forms of combinations in multivariate combination-based permutation tests and ordered categorical variables. *Journal Of Statistical Computation And Simulation*. **86**, 3608-3619 (2016)

[5] Arboretti, R., Bonnini, S., Corain, L. & Salmaso, L. A permutation approach for ranking of multivariate populations. *Journal Of Multivariate Analysis*. **132** pp. 39-57 (2014)

[6] Arboretti, R., Carrozzo, E., Pesarin, F. & Salmaso, L. A multivariate extension of union-intersection permutation solution for two-sample testing. *Journal Of Statistical Theory And Practice*. **11**, 436-448 (2017)

[7] Arboretti, R., Ceccato, R., Corain, L., Ronchi, F. & Salmaso, L. Multivariate small sample tests for two-way designs with applications to industrial statistics. *Statistical Papers*. **59**, 1483-1503 (2018)

[8] Arboretti, R., Bathke, A., Carrozzo, E., Pesarin, F. & Salmaso, L. Multivariate permutation tests for two sample testing in presence of nondetects with application to microarray data. *Statistical Methods In Medical Research*. **29**, 258-271 (2020)

[9] Arboretti, R., Ceccato, R. & Salmaso, L. Permutation testing for goodness-of-fit and stochastic ordering with multivariate mixed variables. *Journal Of Statistical Computation And Simulation*. **91**, 876-896 (2021)

[10] Arboretti, R., Barzizza, E., Biasetton, N., Ceccato, R., Corain, L. & Salmaso, L. A Multi-Aspect Permutation Test for Goodness-of-Fit Problems. *Stats*. **5**, 572-582 (2022)

[11] Barisan, L., Boatto, V., Rossetto, L. & Salmaso, L. The knowledge of Italian wines on export markets: A nonparametric methodology to analyze promotional actions. *British Food Journal*. **117**, 117-138 (2015)

[12] Bazyari, A. & Pesarin, F. Parametric and permutation testing for multivariate monotonic alternatives. *Statistics And Computing*. **23**, 639-652 (2013)

[13] Brombin, C. & Di Serio, C. Evaluating treatment effect within a multivariate stochastic ordering framework: Nonparametric combination methodology applied to a study on multiple sclerosis. *Statistical Methods In Medical Research*. **25**, 366-384 (2016)

[14] Chavan, A. & Shirke, D. Nonparametric tests for testing equality of location parameters of two multivariate distributions. *Electronic Journal Of Applied Statistical Analysis*. **9**, 417-432 (2016)

[15] Chavan, A. & Shirke, D. Simultaneously testing for location and scale parameters of two multivariate distributions. *Revista Colombiana De Estadistica*. **42**, 185-208 (2019)

[16] Chenouri, S. & Small, C. A nonparametric multivariate multisample test based on data depth. *Electronic Journal Of Statistics*. **6** pp. 760-782 (2012)


DRAFT

[17] Corain, L., Giuli, V. & Zecchin, R. A Permutation and Combination-Based Solution on the Ranking of Multivariate Populations in the Case of Ordered Categorical Responses with Application to the Evaluation of the Indoor Environment. *Communications In Statistics-Theory And Methods*. **43**, 879-890 (2014)
[18] Corain, L. & Salmaso, L. Improving power of multivariate combination-based permutation tests. *Statistics And Computing*. **25**, 203-214 (2015)
[19] Corain, L., Arboretti, R., Ceccato, R., Ronchi, F. & Salmaso, L. Testing and ranking on round-robin design for data sport analytics with application to basketball. *Statistical Modelling*. **19**, 5-27 (2019)
[20] Fasolato, L., Novelli, E., Salmaso, L., Corain, L., Camin, F., Perini, M., Antonetti, P. & Balzan, S. Application of nonparametric multivariate analyses to the authentication of wild and farmed European sea bass (Dicentrarchus labrax). Results of a survey on fish sampled in the retail trade. *Journal Of Agricultural And Food Chemistry*. **58**, 10979-10988 (2010)
[21] Giancristofaro, R., Corain, L. & Ragazzi, S. The multivariate randomized complete block design: A novel permutation solution in case of ordered categorical variables. *Communications In Statistics-Theory And Methods*. **41**, 3094-3109 (2012)
[22] Kalina, J. Common multivariate estimators of location and scatter capture the symmetry of the underlying distribution. *Communications In Statistics-Simulation And Computation*. **50**, 2845-2857 (2021)
[23] Lanfranchi, M., Zirilli, A., Alibrandi, A. & Giannetto, C. The behaviour of wine consumers towards organic wine: a statistical analysis through the non-parametric combination test. *International Journal Of Wine Business Research*. (2020)
[24] Langthaler, P., Ceccato, R., Salmaso, L., Arboretti, R. & Bathke, A. Permutation testing for thick data when the number of variables is much greater than the sample size: recent developments and some recommendations. *Computational Statistics*. pp. 1-32 (2022)
[25] Li, J. & Liu, R. New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*. **19**, 686-696 (2004)
[26] Li, J. & Liu, R. New nonparametric tests for comparing multivariate scales using data depth. *Robust Rank-based And Nonparametric Methods*. pp. 209-226 (2016)
[27] Liu, R. & Singh, K. A quality index based on data depth and multivariate rank tests. *Journal Of The American Statistical Association*. **88**, 252-260 (1993)
[28] Montelli, S., Suman, M., Corain, L., Cozzi, B. & Peruffo, A. Sexually diergic trophic effects of estradiol exposure on developing bovine cerebellar granule cells. *Neuroendocrinology*. **104**, 51-71 (2017)
[29] Oja, H. & Randles, R. Multivariate nonparametric tests. *Statistical Science*. **19**, 598-605 (2004)
[30] Salmaso, L. & Pesarin, F. Permutation tests for complex data: theory, applications and software. (John Wiley & Sons,2010)
[31] Pesarin, F. & Salmaso, L. Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal Of Nonparametric Statistics*. **22**, 669-684 (2010)
[32] Pesarin, F., Salmaso, L., Carrozzo, E. & Arboretti, R. Union-intersection permutation solution for two-sample equivalence testing. *Statistics And Computing*. **26**, 693-701 (2016)
[33] Racioppi, M., Salmaso, L., Brombin, C., Arboretti, R., Colombo, R., Serretta, V., Brausi, M., Casetta, G., Gontero, P., Hurle, R. & Others The clinical use of statistical permutation test methodology: a tool for identifying predictive variables of outcome. *Urologia Internationalis*. **94**, 262-269 (2015)
[34] Salmaso, L. Combination-based permutation tests: Equipower property and power behavior in presence of correlation. *Communications In Statistics-Theory And Methods*. **44**, 5225-5239 (2015)
[35] Wedlin, A. On the notion of second-order exchangeability. *Probability, Dynamics And Causality: Essays In Honour Of Richard C. Jeffrey*. pp. 37-54 (1997)
[36] Winkler, A., Webster, M., Brooks, J., Tracey, I., Smith, S. & Nichols, T. Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping*. **37**, 1486-1511 (2016)

DRAFT

# A Quantile Regression Model to Evaluate the Performance of the Italian Courts of Law

C. Cusatelli[a], M. Giacalone[b], E. Nissi[c]

[a] University of Bari "Aldo Moro"; carlo.cusatelli@uniba.it
[b] University of Campania "Luigi Vanvitelli"; massimiliano.giacalone@unicampania.it
[c] University of Chieti-Pescara "Gabriele D'Annunzio"; eugenia.nissi@unich.it

**Abstract**

The efficiency of the judicial system should be one of the main objectives of every democratic state, mainly to prevent the continuation of uncertainty, from a legal and procedural point of view, which can cause prejudice and harm to those who are subjected to a trial. Among the institutions that have the greatest impact on economic performance, the legal and judicial system plays a prominent role. Understanding how laws and regulations affect economic behaviour is fundamental in modern economies in order to verify the distributive impact of the different legal and judicial systems and what features they should have to encourage economic growth. Our analysis, using data disaggregated at district level, has the goal of measuring the efficiency of the Italian judicial offices, empirically estimating a quantile regression model to evaluate the performance of the courts. Data referred to 2021 about judges, pending cases on, incoming cases, as well as resolved cases are extracted from the annual reports of the Italian Ministry of Justice.

**Key words:** quantile regression, judicial system, efficiency evaluation

DRAFT

## 1. Introduction

Efficiency in the general context of modern society is the ability to achieve a predetermined goal, which takes the form of keeping the level of resource productivity. The objective is relevant both in the private sector and in the public one and is achieved through the definition of defined cost standards ex-ante. In the private sector, the issue is based on available resources and therefore often not related to ethical constraints, although these are operational both in the public and in the private. In several countries, economic development has boosted the mobility of population, changing the distribution of litigation. Hence, the increasing difference between the new demand of legal services and the old judicial maps has increased processing time and backlog, therefore, badly affecting judiciary efficiency. According to the Organization for Economic Co-operation and Development [11], the European Union [1], as well as the International Monetary Fund [3], the Italian judicial system is one of the less efficient in Europe, and its inefficiency differs within Italy also respect to judgment Court grade [2] and geographical areas [10].

In the current debate we often discuss of the public sector, being widely believed that an administration that has high levels of spending is inefficient. In fact, the concept of efficiency applied to the functions performed by government institutions should take on a meaning of productivity in relative terms, similarly to the activities of companies: an institution can be defined more or less efficient on the basis of a benchmarking analysis with other units involved in the same production process, and it is correct to speak of greater efficiency only when this comparison shows that a unit is able to achieve the same final results in terms of goods and services delivered (output) using a smaller amount of resources (input), or higher results on a parity of resources used. Identifying efficient performance cannot therefore

result from the mere observation of lower levels of expenditure but must result from an assessment of the institution's ability to allocate to the better resources than the objectives pursued.

## 2. Statistical methodologies for efficiency assessment

The purpose of our paper is to provide new information in judicial context. The performance of frontier estimation methods using data from Italian Courts is analysed in the paper. Quantile Regression (QR) is also suggested as solution to frontier production function estimation [6]. The choice of estimation methods among conventional techniques significantly affects the juridical evaluation. Consequently, QR furnishes valuable new information by estimating different levels of production functions corresponding to different efficiency scores. In addition, the method analyses the performance of the conventional methods.

For the median regression line, we consider initially the case in which it is required that the straight line passes through a point and, subsequently, the case without restrictions. The presentation of the methodology for the determination of the parameters of the straight-line results, especially in the constrained case, in this case the determination of the regression coefficients is obtained by leading, through an appropriate decomposition of the regression residues, bringing the minimum problem back to a linear programming problem.

This methodology, originally introduced and appreciated for its characteristics of robustness, has found great interest in the literature and success in various application areas for the greater completeness of analysis that it is able to offer compared to the classical linear model of regression. Linear quantile regression modelling is introduced using loss functions: in this case it is illustrated how the use of several loss functions lead to least squares regression, absolute minimums and, for the asymmetric absolute loss function, QR. In the application on judicial efficiency, we show some characterizations and properties of the estimators obtained by applying QR. Modelling for economic phenomena, typically not negative, such as income and consumption or for efficiency in a general sense, makes us understand the importance of this technique [14].

Since the Seventies, QR has gained considerable popularity in different fields of research. In the economic field it has been applied, for example, for the study of changes in incomes and wages, in the study of market strategies and in market analysis real estate [6]. QR can be seen as an extension of the linear regression model as it characterizes the entire distribution of a response variable, conditioned on a set of covariates through the estimation of its quantiles. In other words, the estimate of the conditional average is replaced from estimates of conditioned quantiles (e.g., the median), which allow to define more completely and exhaustive the relationships between the variables. Unlike regression on the mean, for QR it is not no assumption is needed about the distributional form of errors, and this makes this method robust [5].

A possible application is based on estimating the efficiency frontier through QR [8]. Moutinho et al. [9] proposed a Data Envelopment Analysis (DEA) and QR approach to evaluate the economic and environmental efficiency assessment in EU country. Tsionas [12] proposed a novel quantile Stochastic Frontier Model (SFM) and develop Markov Chain Monte Carlo techniques for numerical Bayesian inference. In an empirical application to US large banks, he documents important differences between the QR and the traditional SFM, in terms of several aspects of the data and considerable heterogeneity among different quantiles in terms of returns to scale, technical change, efficiency as well as productivity growth.

Economic analysis of productive system is fundamentally based on production function specification and estimation. The production function is a frontier function since it describes the maximum output attainable by the firms that implements efficient production processes. In particular if we denote with $y_n^{\delta}$ the maximum output given the input vector $\boldsymbol{x}_n$ for the n.th firm, in this case the theoretical frontier production is represented by the equation

$$y_n^{\delta} = f(\boldsymbol{x}_n) \tag{1}$$

which correspond to the technology upper limiting the production possibility set. At the same time, the observable production process could be at most equal to its maximum, that is

DRAFT

$$y_n \leq y_n^{\delta} \tag{2}$$

In this context, the frontier estimation can be made in parametric and non-parametric way.

Following the parametric approach, the frontier is represented by a probabilistic relation $f(\mathrm{x}, \theta)$ deriving by means of econometric techniques. This approach is better suited for frontier estimation than standard techniques since it explicitly attempts to estimate the parameters of production function on the frontier. To this purpose the function error term is supposed to consist of a Gaussian random component $\boldsymbol{v}_n$ and of a negative random component which explains the inefficiency $\boldsymbol{u}_n$

$$y_n^{\delta} = f(\boldsymbol{x}_n) + \boldsymbol{v}_n - \boldsymbol{u}_n \tag{3}$$

Empirical implementation of the data provides estimates of economic parameters and efficiency scores for each firm by mean of distance measured obtained separating the (in)efficiency component from the overall error term. In order to improve the implicit neutrality of efficiency distribution it is possible to reformulate the stochastic frontier model substituting the unconditional mean with a conditional mean dependent on a set of explanatory factors. The quantile function $Q_y(\tau|\mathbf{x})$ is defined as $\mathrm{F}^{-1}(\tau) = inf\ \{y: F(y|x) \geq \tau\}$ where $F(y|x)$is the conditional distribution of y given x. The conditional quantile can be determined through

$$\rho_\tau(\epsilon) = \epsilon(\tau - 1) \tag{4}$$

with $\epsilon < 0$ and where the loss function $\rho_\tau(\epsilon)$ is known as the check function.

For a traditional linear model $Q_y(\tau|\mathbf{x}) = x_i'\beta(\tau)$ and the quantile estimators can be obtained solving

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta(\tau)) \tag{5}$$

for a given $\tau$.

## 3. Analysis results

During the last two decades, greater attention was paid to improved performance in the public sector. In this context, the importance of a good judicial system is acknowledged, since it allows to maintain a peaceful coexistence among the citizens of a nation and, above all, the rights and duties that are necessary for each of them [13]. This is even more relevant if we consider the Italian judicial system, which is considered one of the most inefficient in Europe according to the Organization for Economic Co-operation and Development [11] and the European Union [1], as well as the International Monetary Fund [3]. Moreover, inefficiency differs within Italy also respect to the diverse geographical areas [4].

Data referred to 2021 about judges, pending cases on, incoming cases, as well as resolved cases are extracted from the annual reports of the Italian Ministry of Justice.

A Cobb-Douglas frontier function, for which factors returns are equal to parameters is specified.

The number *J* of judges, the incoming cases *I*, and the average duration (in days) *D* are considered as input. The output is represented by the resolved cases. The analysis is conducted on 140 Italian courts. Therefore, the general production function equation is:

$$lny_n = \beta_0 + \beta_1 lnJ + \beta_2 lnI + \beta_3 lnD + \boldsymbol{v}_n - \boldsymbol{u}_n \tag{6}$$

with $\boldsymbol{v}_n \sim N(0, \sigma_v^2)$

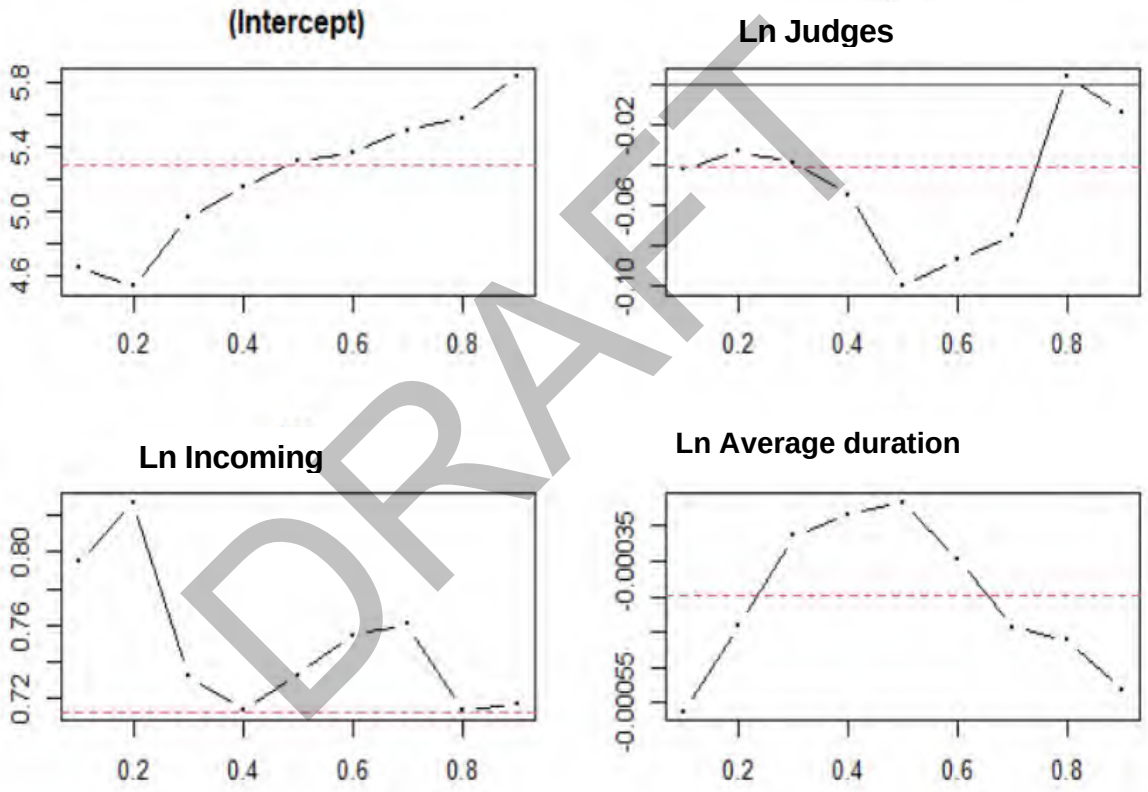
In the model all the Courts are assumed to operate efficiently. (In)Efficiency among Courts is admitted by subtracting a truncated normal stochastic component from the error component of the classical regression model.

DRAFT

Table 1: Stochastic Frontier parameters

| | Estimate | Std. error | Z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.1734 | 0.6341 | -0.2735 | 0.7845 |
| log(Judges) | 0.0299 | 0.0535 | 0.5598 | 0.5756 |
| log(Incoming) | 0.8876 | 0.0418 | 212.3380 | < 2.2e-16 |
| log(the average duration in days) | 0.1552 | 0.0961 | 16.1530 | 0.1063 |
| sigmaSq | 0.6238 | 0.1042 | 59.8670 | < 2.2e-6 |
| Gamma | 0.8793 | 0.0496 | 177.3070 | < 2.2e-16 |

An alternative approach can be obtained using QR which provides a description of a response variable as a conditional function of a set of covariates broader than the methods based on conditional means. The interesting feature consists of extending the analysis from mean or median values to the full range of other conditional quantile functions, providing an analytical description of an ordered set of technological relationships corresponding to different level of efficiency.




DRAFT

Figure 1: Estimate quantile regression

On Fig. 1 we can observe a strong variation in estimated coefficient along the percentile axis, and an increasing intercept along the quantiles. The change of coefficient is more noticeable for the incoming cases but also considered for the Judges.

The dotted line in each figure shows the least squares estimate of the conditional mean effect. A considerable dispersion is observed for the coefficient of explicative variables at different quantiles. The estimated value of $\beta_{1\tau}$ varied between -0,10 to -0,02. The estimated value of $\beta_{2\tau}$ varied 0,72 to 0,80 while $\beta_{3\tau}$varied from -0,00055 to -0,00035 and it reached it's the maximum at 50$^{th}$ quantile.

## 4. Final remarks

The purpose of the present paper is to provide new information in Judicial context. The performance of frontier estimation methods using data from Italian Courts is analysed in the application. QR is also suggested as solution to frontier production function estimation [6]. The choice of estimation methods among conventional techniques significantly affects the juridical evaluation. Consequently, QR furnishes valuable new information by estimating different levels of production functions corresponding to different efficiency scores. In addition, the method analyses the performance of the conventional methods.

Civil justice has been the subject of numerous legislative interventions over the recent years, also because of growing attention that the institutions European countries have attributed to this area reform. Since 2012 the improvement of quality, independence and efficiency of judicial systems has become, in fact, a priority of the European system. An inefficient judicial system harms the fundamental rights of citizens and penalizes businesses and investors. The European Commission, in the country-specific recommendations it presents each year, has for long been asking Italy for a reform of justice that eliminates delays and inefficiencies.

From the analyses we carried out concerning performance of the ordinary courts, the existence of territorial differences emerges significantly, between the Northern and Southern regions of the country. There are also evident greater difficulties of the smaller courts, respect to those of larger dimensions. Small courts seem, in fact, to suffer of critical issues such as the impact on the ability to cope with the annual flow of new practices, resulting increased gradients, both on duration of the proceedings.

DRAFT

## References


[1] CEPEJ: Report on the evaluation of the judicial system. Council of Europe's European Commission for the Efficiency of Justice Report (2012)

[2] Cusatelli, C., Giacalone, M.: Evaluating the judicial activity: a proposal of indicators and analyses of criminal burden. Social Indic. Res. 138(2), 725--746 (2018)

[3] Esposito, G., Lanau, S., Pompe, S.: Judicial system reform in Italy: a key to growth. IMF Work. Pap. 14(32), (2014)

[4] Giacalone, M., Nissi, E., Cusatelli C.: Dynamic efficiency evaluation of Italian judicial system using DEA based Malmquist productivity indexes. Socio-Econ. Plann. Sci. 72, (2020)

[5] Hao, K., Naiman D.Q.: Quantile Regression. SAGE, London (2007)

[6] Jradi, S., Parmeter, C.F., Ruggiero, J.: Quantile estimation of the stochastic frontier model. Econ. Lett. 182, 15--18 (2019)

[7] Koenker R.: Quantile Regression. Cambridge University Press, New York (2005)

[8] Liu C., Laporte A., Ferguson B.S.: The quantile regression approach to efficiency measurement: insights from Monte Carlo simulations. Health Econ. 17(9), 1073—1087 (2008)

[9] Moutinho, V., Madaleno, M., Robaina, M.: The economic and environmental efficiency assessment in EU cross-country: evidence from DEA and quantile regression approach. Ecol. Indic. 78, 85--97 (2017)

[10] Nissi E., Rapposelli A.: A data envelopment analysis of Italian courts efficiency. Stat. Appl. - Ital. J. Appl. Stat. 22(2), 199--210 (2010)

[11] OECD: What makes civil justice effective? OECD Econ. Department Policy Notes, 18 (2013)

[12] Tsionas, M.G.: Quantile stochastic frontiers. Eur. J. Oper. Res. 282(3), 1177--1184 (2020)

[13] Voigt S.: Determinants of judicial efficiency: a survey. Eur. J. Law Econ. 42(2), 183--208 (2016)

[14] Zhang, W., Chiu, Y.B.: Country risks, government subsidies, and Chinese renewable energy firm performance: new evidence from a quantile regression. Energy Econ. 106540 (2023)

# A variable selection procedure based on predictive ability: a preliminary study on logistic regression

Rosaria Simone[a] and Mariarosaria Coppola[a]

[a]Dipartimento di Scienze Politiche, Università degli Studi di Napoli Federico II;
rosaria.simone@unina.it, m.coppola@unina.it

**Abstract**

The contribution is meant as a pilot exercise on the development of a variable selection technique that combines fitting and prediction performance. The simplest step towards this goal involves a forward selection algorithm for binary classification achieved via logistic regression. At each step, the algorithm selects the predictor, among the covariates that are significantly associated with the outcome, that entails the significantly largest AUC increment with respect to the previous model. For the sake of illustration, we present some examples relative to the search of the most predictive factors of the propensity to pension planning, health insurance subscription, and the use of Financial Technology services like home banking, taken from the Survey on Household Income and Wealth 2020 run by Bank of Italy. Concluding remarks are provided on further developments.

***Keywords:*** Logistic regression, Forward variable selection, AUC index

DRAFT

## 1. The framework

The paper investigates the performance of a forward variable selection for (possibly weighted) logistic regression based on the area under the ROC[1] curve (AUC) (1; 5) with the aim of determining a parsimonious model with both good fitting and predictive abilities. This goal is strategic for diagnostic procedures in the medical field, but also for socio-economic studies, as several indicators of consumers' behavior are difficult to predict. The background idea is not new, as the challenge of combining explanatory and predictive goals in regression models has solicited several works in the literature (13). For instance, (11) provided an estimation procedure based on the optimization of the empirical area under the ROC curve, which has been improved in (14). When approaching this challenge, one of the first issues one has to wonder is related to methods of testing the statistical significance of the AUC difference of correlated ROC curves, as those corresponding to nested models on the same data. The so-called DeLong test (3) for comparing the AUC difference between correlated ROC curves may raise some critical issues, as it can be conservative if based on in-sample analysis and it is based on multivariate normal data (a setting that is very common in the biomedical framework). For these reasons, the Authors in (4) advice against its direct application and suggest evaluations based on the corresponding confidence intevals instead. Issues due to the asymptotic distribution of the AUC under the null are discussed also in (9), and motivate the introduction of a test procedure based on AUC for logistic regression but only for two binary predictors and solely for testing the first-stage null $H_0 : AUC = \frac{1}{2}$.

[1]Receiving operating characteristic

Hereafter, in order to obtain an automatic procedure and embed a test on AUC differences within a variable selection algorithm, we resort to the (stratified) bootstrap procedure implemented in the `R` package `pROC` (12) to test that the prediction improvement between subsequent steps is statistically significant.

## 1.1 AUC-based forward variable selection for logistic regression

Several measures can be chosen to assess the predictive performance of a binary classification, as the classical Brier Score (2) and the (empirical) AUC index of the receiving operating characteristic curve constructed on out-of-sample observations. Here we will rely on the AUC since it is the most popular choice and also since it is a normalized indicator.

Since the selection depends on prediction performances, $K$-fold cross validation will be exploited and at each step (with $K = 10$), model estimation will be carried out in-sample on a training set, whereas its discrimination ability will be assessed on out-of-sample observations. The algorithm can be summarized as follows:

- Let $\mathcal{M}^{(h-1)}$ be the model estimated and validated at step $h-1$, with $h-1$ covariates $\mathbf{Z}^{(h-1)}$.
- The goal of the $h$-th step of the forward procedure is to identify the covariate $X$ in the available predictor space $\mathbf{X}(\mathcal{M})$ that optimizes both fitting and prediction with respect to the previous step. With some more details:
  1. for each $k = 1, \ldots, K$, the dataset is divided into $K$ folds, say $\mathcal{D}_1, \ldots, \mathcal{D}_K$;
  2. for each $k = 1, \ldots, K$, consider $\mathcal{F}_k = \bigcup_{i \neq k} \mathcal{D}_i$ as training set and $\mathcal{D}_k$ as test set;
  3. for each $k$, estimate the Model $Y \sim (X, \mathbf{Z}^{(h-1)})$ for each available covariate $X$, say $\mathcal{M}_X^{(k,h)}$: let $BIC_X^{(k,h)}$ be its BIC on $\mathcal{F}_k$, $AUC_X^{(k,h)}$ be the corresponding AUC on the test set $\mathcal{D}_k$ and $\beta_X^{(k,h)}$ its estimated regression effect;
  4. for each covariate $X$ entailing significant effects over all the folds, consider the average BIC, AUC and regression coefficients over the folds:

$$\widehat{BIC}_X^{(h)} = \frac{1}{K}\sum_{k=1}^{K} BIC_X^{(k,h)}, \qquad \widehat{AUC}_X^{(h)} = \frac{1}{K}\sum_{k=1}^{K} AUC_X^{(k,h)}, \qquad \widehat{\beta}_X^{(h)} = \frac{1}{K}\sum_{k=1}^{K} \beta_X^{(k,h)} \tag{1}$$

- the $h$-th variable selected to enter the model is then given by $X^\star$ if:

$$\widehat{AUC}_{X^\star}^{(h)} = \max_{X \in \mathbf{X}(\mathcal{M})} \{\widehat{AUC}_X^{(h)}\}$$

  and if the difference with respect to $\widehat{AUC}_{X^\star}^{(h-1)}$ is statistically significant, at given $\alpha$ level, so that $\mathbf{Z}^{(h)} = (\mathbf{Z}^{(h-1)}, X^\star)$ is the new predictor vector. If $h = 1$, the procedure will resort to testing the null $H_0 : AUC = \frac{1}{2}$ against the alternative $H_1 : AUC > \frac{1}{2}$ for each significant covariate, choosing in the end the one entailing the largest significant difference with respect to the null.
- Then step $h+1$ is initiated and the procedure is re-iterated until either there is no covariate entailing significant effects or if there is no significant covariate that implies a significant increment in the AUC with respect to the previous step.

# 2. Some examples

In recent decades, the growing diffusion of financial technology (FinTech) and the progressive ageing of the population have profoundly changed the behavior of individuals. In particular, the longer lifespan requires individuals to make greater efforts in planning their retirement deciding how and how

DRAFT

much to save in order to ensure an adequate level of well-being and adequate health care in old age. Similarly FinTech is revolutionizing the financial services sector with a strong impact both in terms of financial planning and financial well-being (10). In this context we chose to test our proposal to the identification of the most predictive factors of some indicators of the propensity to pension planning and to use FinTech services. We refer to the 2020 edition of the Survey on Household Income and Wealth (SHIW) run by the bank of Italy (6), one of the most comprehensive studies carried out by the Bank of Italy to collect information on economic and financial status of Italian households. We consider the usage of home banking as an indicator of the propensity to use the FinTech services, and health insurance and pension plan subscription as indicators of pension planning. Indeed, these indicators are deemed to reveal important aspects of the respondent economic and financial behaviour, thus their prediction can be of interest to stakeholders.

The set example used hereafter consists of $n = 4044$ responses from household's head, after omission of missing values. The predictor space consists of all socio-economic and demographic variables collected within the survey: gender[2], having or nor a university education, work condition[3], household composition and its economic and living conditions, and the age generation of the household head. For the sake of simplicity, the procedure will consider only binary indicators as predictors.

Table 1 reports the most influential predictive factors of the indicator of usage of home banking service, listed in the ordering determined by the procedure. Specifically, for each row, we report the corresponding average AUC and BIC referring to the model including the covariates up to that row, where, at each step, we have averaged the AUC measures referring to the out-of sample ROC curves constructed on the left-out fold, for varying folds. We also report the average $p$-value for the test of significant differences in AUC between subsequent steps and the average regression coefficient since our ultimate goal is to identify both explanatory and predictive properties of available predictors. Tables 2-3 report the analogous results for the analysis on the indicators of health insurance and pension plan subscription, respectively.

It is worth to notice that the ordering in average AUC across folds does not match the corresponding one relative to BIC chosen as fitting indicator. A graphical display of the results with the plot of (empirical) ROC curves computed stepwisely is provided in Figure 1.

DRAFT

Table 1: Results from AUC-based forward variable selection applied to the logistic regression of the binary indicator of usage of home banking services in 2020

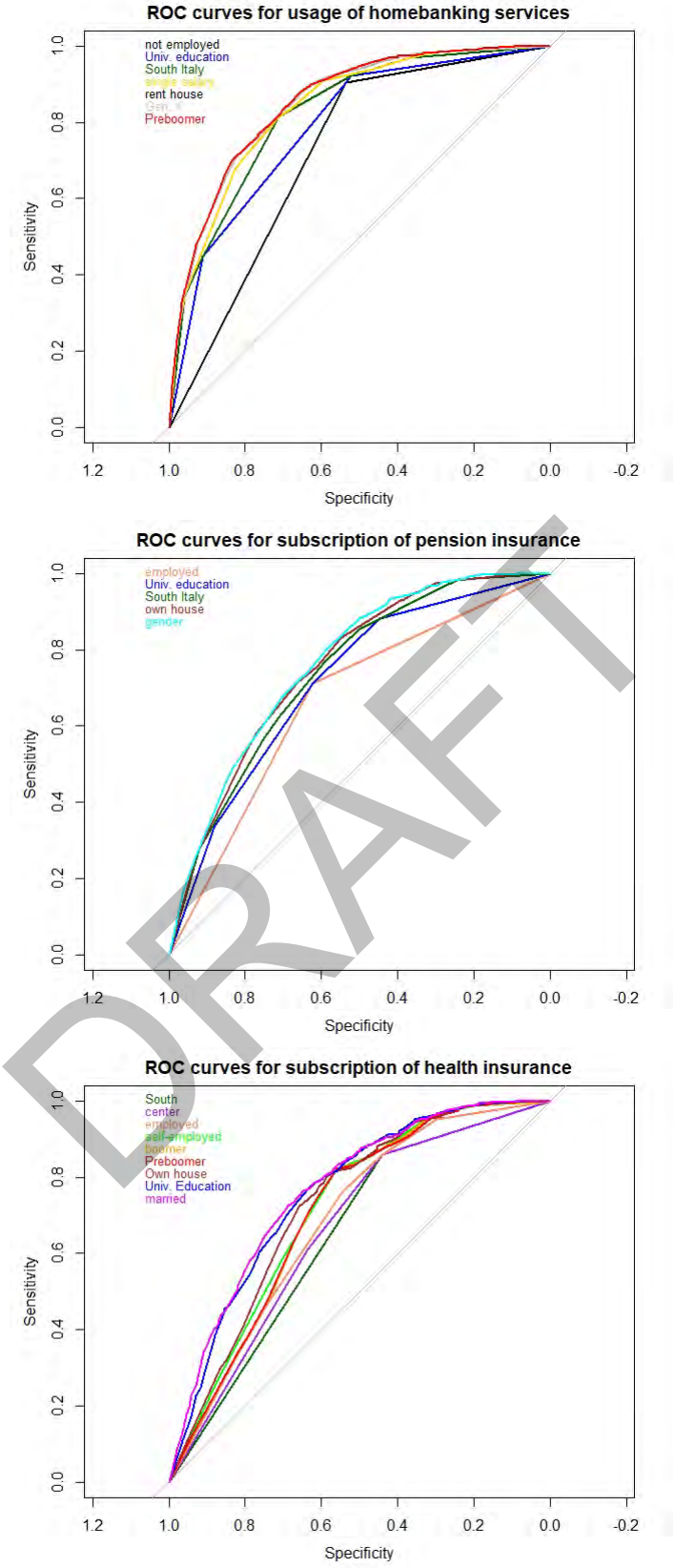
| | | Mean-AUC | Mean-Coef | Mean-BIC | Mean p-value |
|---|---|---|---|---|---|
| 1 | not employed or retired | 0.7210 | 0.0062 | 4373.2254 | 0.0000 |
| 2 | univ. education | 0.7875 | 1.7978 | 4128.4322 | 0.0000 |
| 3 | South Italy | 0.8240 | -1.4685 | 3810.6599 | 0.0000 |
| 4 | single salary | 0.8391 | -0.7843 | 3719.8381 | 0.0000 |
| 5 | tenant of rent house | 0.8446 | -0.8166 | 3672.5934 | 0.0003 |
| 6 | generation X | 0.8497 | 0.4316 | 3655.5302 | 0.0004 |
| 7 | pre-boomer | 0.8521 | -1.5445 | 3593.8227 | 0.0409 |

## 3. Concluding remarks

Further research will concern comparison with predictions and variable importance measures obtained with ensemble methods, like random forests (7), as well as with penalization and shrinkage tech-

[2]gender= 1 for women, gender = 0 for men

[3]The binary covariate *employee* does not consider self-employed people, for which a separate binary indicator is built.




DRAFT

Figure 1: Empirical ROC curves at each step of the proposed forward variable selection applied to the logistic regression of the binary indicator of usage of homebanking services, pension plan and health insurance subscription (from top to bottom)

Table 2: Results from AUC-based forward variable selection applied to the logistic regression of the binary indicator of health insurance subscription in 2020

| | | Mean-AUC | Mean-Coef | Mean-BIC | Mean p-value |
|---|---|---|---|---|---|
| 1 | south Italy | 0.6492 | 0.0667 | 3026.4791 | 0.0000 |
| 2 | central Italy | 0.7218 | 1.2416 | 2897.5322 | 0.0000 |
| 3 | employed | 0.7284 | 1.2369 | 2899.9611 | 0.0000 |
| 4 | self employed | 0.7445 | 1.2070 | 2880.0764 | 0.0000 |
| 5 | boomer | 0.7544 | 0.9909 | 2811.2971 | 0.0000 |
| 6 | pre boomer | 0.7507 | 0.9730 | 2813.7462 | 0.0000 |
| 7 | own house | 0.7499 | 0.9640 | 2816.7228 | 0.0000 |
| 8 | univ. education | 0.7592 | 0.9014 | 2793.9727 | 0.0000 |
| 9 | married | 0.7670 | 0.3419 | 2793.5198 | 0.0002 |

Table 3: Results from AUC-based forward variable selection applied to the logistic regression of the binary indicator of pension plan subscription in 2020

| | | Mean-AUC | Mean-Coef | Mean-BIC | Mean p-value |
|---|---|---|---|---|---|
| 1 | employee | 0.6675 | 0.0424 | 3036.5627 | 0.0000 |
| 2 | univ. education | 0.7156 | 1.0117 | 2948.6740 | 0.0000 |
| 3 | south Italy | 0.7379 | -1.2710 | 2825.1729 | 0.0157 |
| 4 | own house | 0.7557 | 1.0180 | 2777.0147 | 0.0000 |
| 5 | gender | 0.7653 | -0.7096 | 2736.8597 | 0.0021 |

niques for the ranking of fitting performance of covariates. More generally, the proposal could be extended to a best-subset variable selection where - at each step - the procedure includes a set of covariates that maximizes a given prediction accuracy measure (8), or to a general unknown link function if the logit transform is not adequate.

DRAFT

## References


[1] Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology. **12**(4):387–415 (1975)

[2] Brier, G.W. Verification of forecasts expressed in terms of probability. Month. Weather Rev. **78**(1):1–3. (1950)

[3] DeLong, E.R., DeLong, D.M., Clarke-PearsonD.L.: Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. **44**(3), 837–845 (1988)

[4] Demlera, O.V., Pencina, M.J., D'Agostino R.B. Sr.: Misuse of DeLong test to compare AIC for nested models . Statistics in Medicine. **31**(23), 2577–2587 (2012)

[5] Fawcett T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, **27**(8): 861–874.

[6] Faiella I. and Gambacorta R. (2007). The weighting process in the SHIW, Bank of Italy. Economic Working Paper No. 636.

[7] Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Pattern Recognition Letters. **31**(5), 2225–2236 (2010)

[8] Hastie, T., Tibshirani, R. and Tibshirani R.: Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. Statistical Science. **35**(4), 579–592 (2020)

[9] Lieli R.P., Hsu Y.-C.: Using the area under an estimated ROC curve to test the adequacy of binary predictors. Journal of Nonparametric Statistics. **31**(1):100-130 (2019)

[10] Panos, G.A., Wilson, J.O.S.: Financial literacy and responsible finance in the FinTech era: capabilities and challenges. The European Journal of Finance. **26**:4–5, 297-301 (2020)

[11] Pepe, M.S., Cay, T., Longton, G. :Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve. Biometrics, **62**: 221–229 (2006)
[12] Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J.C. and Müller M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, **12**, p. 77 (2011)
[13] Shmueli, G. To explain or to predict? Statistical Science. **25**:289–310 (2010)
[14] Zhou, X.H., Chen, B., Xie, Y.M., Tian, F., Liu, H. and Liang, X.: Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoporosis disease. Statistics in Medicine. **31**:628–635 (2012)

DRAFT

# Comparison of binary regressions with asymmetric link function for imbalanced data

Michele La Rocca[a], Marcella Niglio[a], and Marialuisa Restaino[a]

[a]Di.S.E.S., University of Salerno, Via Giovanni Paolo II, 132 - Fisciano (SA) - Italy ,
[larocca, mniglio, mlrestaino]@unisa.it

**Abstract**

Logit and Probit link functions are largely used in the regression domain to model binary response data. In the presence of an imbalanced dependent variable their use can lead to obtaining biased estimates of the regression model parameters and then the selection of asymmetric links may be more appropriate in this case. We here compare two asymmetric link functions for regression models, the Generalized Extreme Values and the Generalized Logistic links, presenting some results related to their maximum likelihood estimators. Using different rates of imbalance in the response data, the comparison of the empirical distribution and the accuracy of the estimates of the models' parameters based on the two link functions are evaluated through a simulation study.

***Keywords:*** Asymmetric link, GEV, Generalized Logistic, likelihood

DRAFT

## 1. Introduction

Estimating binary response variable is an important statistical task in many areas, including social sciences, biology, and economics. In binary regression, the two commonly used symmetric link functions are the Logit and Probit [10]. Several studies have investigated their limitations. In particular, statistical models based on the symmetric link functions lie on the assumption of equal class distribution for data [8, 14] and they might produce biased estimates [10]. However, these functions might be improper in the presence of imbalanced data, which occurs when one of the classes is much smaller than the other. If the degree of imbalance is extreme, the events become rare [8, 14]. The main reason for this inadequacy is that the probability of a binary response, as a function of covariates, approaches zero and one at different rates [3, 7], and consequently the probability of imbalanced events should be underestimated [8, 14].

To deal with these problems, some proposals have been suggested such as i) the bias correction method [8], ii) the penalized methods [6], and iii) the use of asymmetric link functions [9, 11, 12, 14].

Given this framework, our attention is on the last point, i.e. on the use of asymmetric link functions, and we investigate the effect of the finite sample size on the choice of the link functions. In more detail, the aim is to evaluate and compare the performance of two classes of asymmetric link functions, Generalized Extreme Values (GEV) distribution [14], and Generalized Logistic distribution [5], in the presence of different rates of imbalance in the binary response variable, according to the finite sample sizes.

In Sect. 2, the regression models for unbalanced binary data are introduced. In Sect. 3, through a simulation study, the empirical distribution of the estimates and their accuracy are shown for the two classes of asymmetric link functions is done.

## 2. Regression models for imbalanced binary data

Consider a response variable $Y$ whose distribution belongs to the exponential family and a vector of $p$ covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Further, $E[Y_i] = \mu_i$ is the expectation of $Y_i$, for $i = 1, 2, \ldots, n$, with $n$ the sample size and where $g(\cdot)$ is a monotone and differentiable function such that:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the $[(p+1) \times 1]$ vector of parameters and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of explanatory variables of unit $i$. The function $g(\cdot)$, called *link function*, relates $\mathbf{x}_i'\boldsymbol{\beta}$ to $\mu_i$ and has to be chosen to properly deal with the set of values assumed by $\mu_i$, for $i = 1, 2, \ldots, n$. The equation (1) defines the generalized linear model (GLM) characterized by a link function that is an increasing or decreasing function of $\mu_i$ (among the others see [10]).

When $Y$ is a binary response variable which assumes values $y = \{0, 1\}$, the probability associated to $Y_i = y_i$ is $\pi_i$ if $y_i = 1$ and $1 - \pi_i$ if $y_i = 0$. Therefore, $Y_i$ can be modeled by using a Bernoulli random variable with probability density function $P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$, for $y_i = \{0, 1\}$ and $i = 1, 2, \ldots, n$.

Furthermore,

$$E[Y_i] = \pi_i = P(Y_i = 1) = F(\mathbf{x}_i'\boldsymbol{\beta}), \tag{2}$$

where $F(\cdot)$ is the cumulative distribution function and using the GLM notation (1):

$$\pi_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) = F(\mathbf{x}_i'\boldsymbol{\beta}).$$

The selection of the link function $F(\cdot)$ is strictly related to the nature of the dependent variable $Y$ and its misspecification may affect the estimated regression coefficients (for small and large-sample effects in the binary response case see [4]).

As underlined in Section 1, the symmetric links function (logit and probit) are mostly used even if they might produce biased estimates in the presence of imbalanced data. Therefore, asymmetric link functions should be preferable.

In this manuscript, we will focus on two classes of asymmetric link functions:

1. *Generalized Extreme Value* (GEV) *distribution* [14]:

$$\pi_i = \exp\{-[1 + \xi\mathbf{x}_i'\boldsymbol{\beta}]^{-\frac{1}{\xi}}\}, \tag{3}$$

   where $(1 + \xi\mathbf{x}_i'\boldsymbol{\beta}) > 0$ and $\xi \in \mathbb{R}$ is the shape parameter.
2. *Generalized Logistic distribution* [5]:

$$\pi_i = \frac{1}{(1 + \exp\{-(\mathbf{x}_i'\boldsymbol{\beta})\})^{\alpha}}, \tag{4}$$

   where $\alpha > 0$ is the parameter that controls the asymmetry. If $\alpha = 1$, the symmetric logistic distribution is obtained whereas as $\alpha \to \infty$ the (4) becomes a Gumbel distribution [13].

The corresponding log-likelihood function associated with the binary regression model is given by:

$$\ell(\boldsymbol{\beta}, \gamma; \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\}, \tag{5}$$

where $\pi_i$ is given by (3) and (4), according to which distribution is chosen, $\gamma$ is the shape parameter, and therefore $\gamma = \xi$ in the GEV case and $\gamma = \alpha$ with the Generalized Logistic distribution.

In Fig. 1 Generalized Logistic and GEV cumulative distributions are plotted for different shape parameters and for the sake of comparison Logistic distribution is also shown. It can be clearly appreciated the asymmetry of the Generalized Logistic and of the GEV distribution and their different behaviors as the probability approaches 0 or 1.

DRAFT

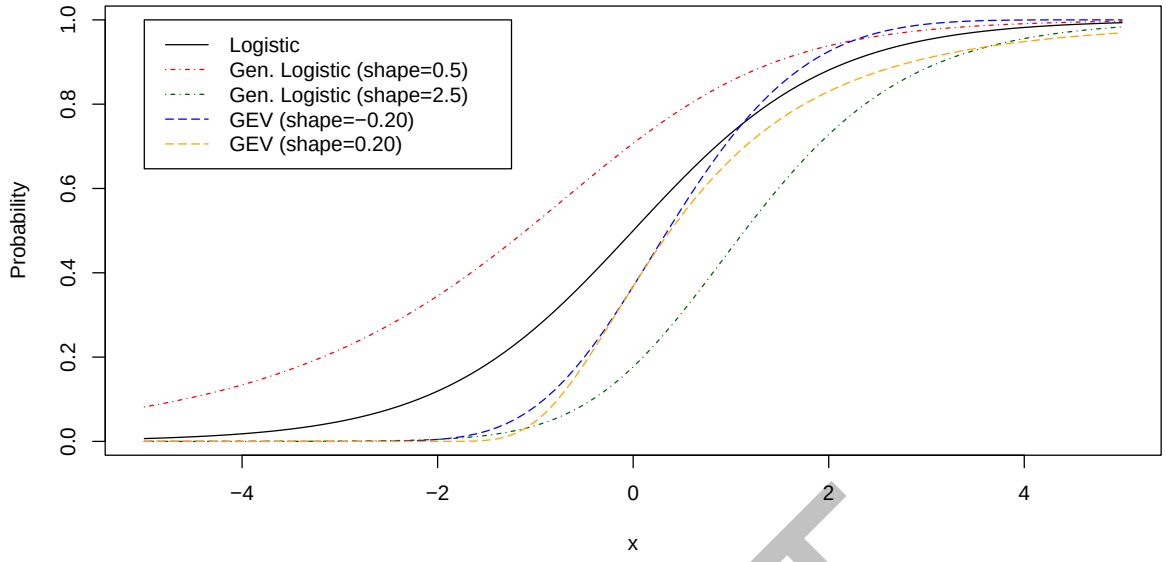



Figure 1: GEV and Generalized Logistic cumulative distributions for different values of the shape parameters.

DRAFT

The estimation of the parameters of the GEV regression model based on the maximum likelihood method has been largely investigated in [1] whereas different approaches have been considered for the estimation of the parameters when the link function is given by (4).

In both cases, the score functions do not have closed form and then the maximum-likelihood estimators can be obtained by maximizing the log-likelihood by numerical algorithms. In this case, the knowledge of the score vector can be of help to reduce the computational time and simplify the search of the maximum (mainly in this case where the function to be maximized is not linear).

The score vector of the GEV regression model has been presented in [1] but, in our knowledge, any corresponding results are given for the Generalized Logistic regression. For this reason, we present here the score function of the Generalized Logistic regression model, given by:

$$\frac{\partial \ell(\boldsymbol{\beta}, \alpha; \mathbf{X}, \mathbf{y})}{\partial \beta_j} = \sum_{i=1}^{n} \alpha \frac{\pi_i^{1/\alpha}}{(1-\pi_i)} x_{ij} \exp[-\mathbf{x}_i' \boldsymbol{\beta}](y_i - \pi_i), \quad j = 0, 1, \ldots, p \tag{6}$$

$$\frac{\partial \ell(\boldsymbol{\beta}, \alpha; \mathbf{X}, \mathbf{y})}{\partial \alpha} = \sum_{i=1}^{n} \alpha^{-1} \left[ (y_i - \pi_i) \frac{\log(\pi_i)}{1-\pi_i} \right]. \tag{7}$$

where $\pi_i$ is given in (4), for $i = 1, 2, \ldots, n$.

What associates the two distributions (3) and (4) is not only the asymmetric behavior but also the difficulties that can arise to estimate the shape parameter ($\xi$ and $\alpha$ respectively). These difficulties are mainly due to the complexity of the function to be maximized and the behavior of the log-likelihood that can be empirically appreciated in Fig. 2.

This problem is overtaken in the GEV case, defining for $\xi$ a grid of values over which the $\boldsymbol{\beta}$ vector is estimated, as in [2] among the others. Bayesian approaches are instead used for the Generalized Logistic case.

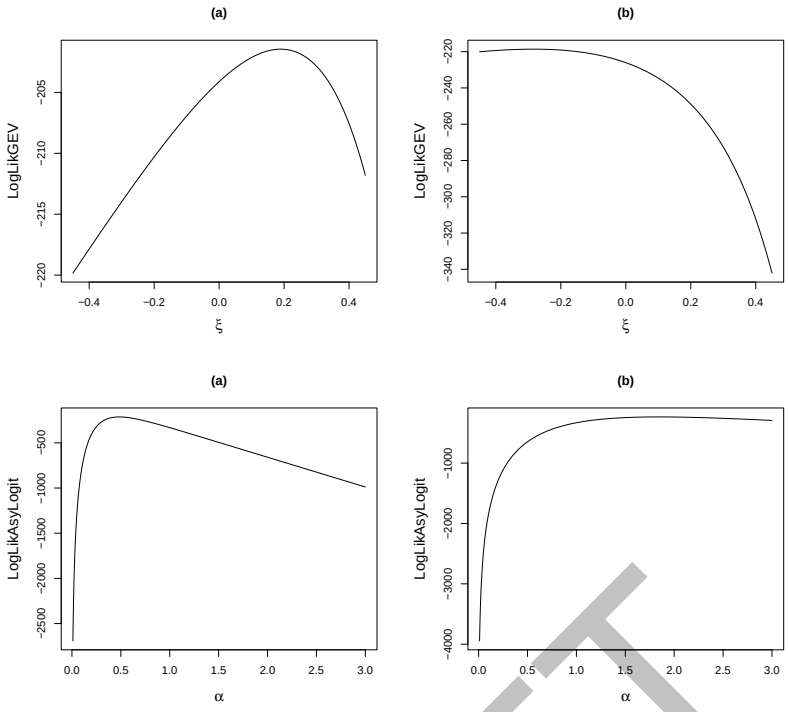



Figure 2: Top: Profile log-likelihood of the GEV regression when $\xi = 0.20$ frame (a) and $\xi = -0.20$ frame (b); Low: Profile log-likelihood of the Generalized Logistic regression when $\alpha = 0.5$ frame (a) and $\alpha = 2$ frame (b).

## 3. Simulation study

The choice of the asymmetric link function in the presence of imbalanced binary response data cannot be an easy task when the selection is made in large classes of functions. We have presented in Sect. 2. two asymmetric links and we have discussed some problems that can arise when the log-likelihood has to be maximized. Another problem that needs to be faced is how to choose between them. We here consider this point evaluating how the imbalance of the dependent variable affects the estimate of the vector $\boldsymbol{\beta}$, and consequently its variability, with finite samples.

To evaluate how the imbalance of the Bernoulli dependent variable ($Y$) affects the maximum likelihood estimates when $n$ is finite, we have implemented a simulation study to compare the estimate of the vector $\boldsymbol{\beta}$ in the presence of GEV and Generalized Logistic regression models, when $Y$ is characterized by different proportions of imbalance.

For this aim, we have generated the data from both regression models using four sample sizes $n = \{100, 250, 500, 1000\}$, three different proportions of imbalance $Pr(Y = 1) = \{0.05, 0.10, 0.20\}$, with $Pr(Y = 1)$ the proportion of one's of the dependent variable $Y$. The covariates $X_j$, for $j = 1, 2, 3, 4$ have been generated from independent standard Normal random variables, the true value of $\beta_j = 0$, for $j = 1, 2, 3, 4$, whereas the true value of $\beta_0$ is selected to guarantee the proportion of imbalance fixed in each case examined in the simulation design.

For all 24 scenarios of the simulation design we have generated 1000 datasets and for each of them, we have estimated the parameters of the vector $\boldsymbol{\beta}$, fixing the shape parameter ($\xi = -0.20$ in the GEV model and $\alpha = 2$ in the Generalized Logistic model).

In Fig. 3 the empirical distributions of $(\hat{\beta}_j - \beta_j)$, for $j = 0, 1$ are presented for all values of $Pr(Y = 1)$ and $n$ (the remaining $\beta_j$ parameters, $j = 2, 3, 4$, are not included in the figure because, as expected, their behavior is similar to $\beta_1$).

The corresponding empirical distributions of the Generalized Logistic distribution are instead given in Fig. 4 where the main difference with the GEV case, can be appreciated only for small $n$ and small

DRAFT

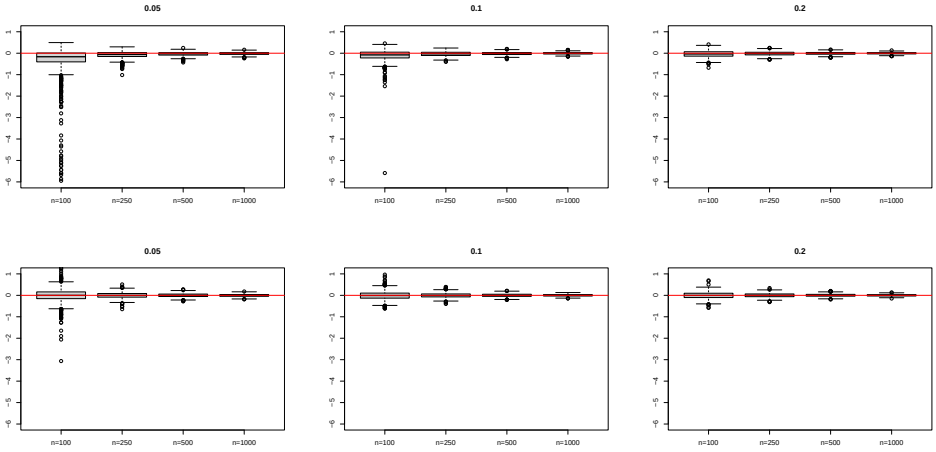



Figure 3: Empirical distribution of $(\hat{\beta}_0 - \beta_0)$ (first row) and $(\hat{\beta}_1 - \beta_1)$ (second row) for the GEV regression model, with $Pr(Y = 1) = 0.05$ (left), $Pr(Y = 1) = 0.10$ (center) and $Pr(Y = 1) = 0.20$ (right).

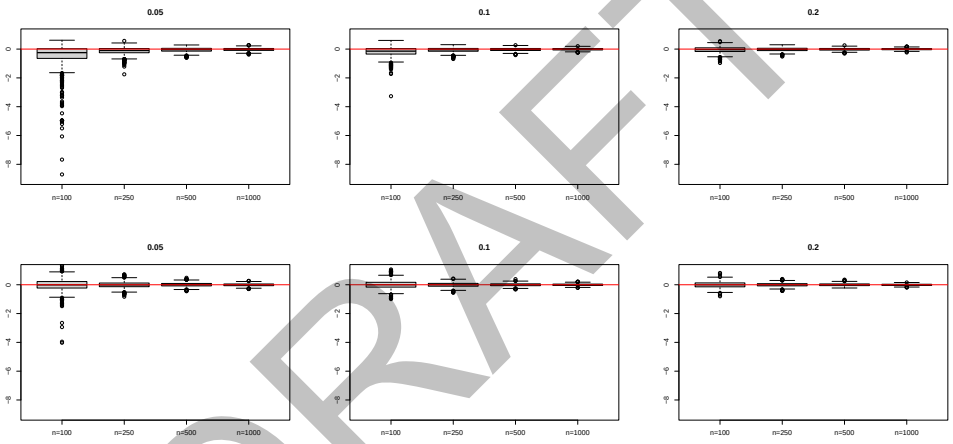



Figure 4: Empirical distribution of $(\hat{\beta}_0 - \beta_0)$ (first row) and $(\hat{\beta}_1 - \beta_1)$ (second row) for the Generalized Logistic regression model, with $Pr(Y = 1) = 0.05$ (left), $Pr(Y = 1) = 0.10$ (center) and $Pr(Y = 1) = 0.20$ (right).

DRAFT

$Pr(Y = 1)$.

To further evaluate the estimates obtained for $\boldsymbol{\beta}$ by using both models, in Table 1 the MSE ratios of the $\hat{\beta}_j$, $j = 0, 1$, estimated from the GEV and the Generalized Logistic regression are presented. In all cases, the MSE of the parameters estimated from the GEV regression is lower than the Generalized Logistic case and this difference is more marked when $n$ is small and the imbalance of data is quite high $(Pr(Y = 1) = 0.05)$.

As expected, the choice of the link function in binary regression with imbalanced data becomes crucial in finite sample sizes, especially when $n$ is small.

This is the first step of the comparison that needs to be further explored considering a larger class of asymmetric link functions that is left for future research.

Table 1: MSE ratios of the $\hat{\beta}_j$, $j = 0, 1$ estimated from the GEV and the Generalized Logistic regression.

| | $n$=100 | | | $n$=250 | | | $n$=500 | | | $n$=1000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Pr(Y=1)$ | 0.05 | 0.10 | 0.20 | 0.05 | 0.10 | 0.20 | 0.05 | 0.10 | 0.20 | 0.05 | 0.10 | 0.20 |
| $\hat{\beta}_0$ | 0.016 | 0.618 | 0.536 | 0.391 | 0.481 | 0.560 | 0.396 | 0.478 | 0.537 | 0.353 | 0.455 | 0.522 |
| $\hat{\beta}_1$ | 0.025 | 0.520 | 0.560 | 0.411 | 0.540 | 0.595 | 0.450 | 0.505 | 0.559 | 0.475 | 0.508 | 0.545 |

## References


[1] Calabrese, R., Osmetti, S.: Modelling SME loan defaults as rare events: An application to credit defaults. J. Appl. Stat. **40**, 1172–1188 (2013)

[2] Calabrese, R., Giudici, P.: Estimating bank default with generalised extreme value regression models. J. Oper. Res. Soc. **66**, 1783–1792 (2015)

[3] Chen, M.H., Dey, D.K., Shao, Q.M.: A new skewed link model for dichotomous quantal response data. J. Am. Stat. Assoc. **94**, 1172–1186 (1999)

[4] Czado, C., Santner, T.J.: The effect of link misspecification on binary regression inference. J. Stat. Plan. Infer., **33**, 213–231 (1994)

[5] Devidas, M., George, EO., Zelterman, D.: Generalized logistic models for low-dose response data. Stat. Med., **12**, 881–92 (1993)

[6] Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika, **80**, 27–38 (1993)

[7] Kim S., Chen M.H., Dey, D.K.: Flexible generalized t-link models for binary response data. Biometrika, **95**, 93–106 (2007)

[8] King, G., Zeng, L.: Logistic regression in rare events data. Polit. Anal., **9**, 137–163 (2001)

[9] La Rocca, M., Niglio, M., Restaino, M.: Bootstrapping binary GEV regressions for imbalanced datasets. Comp. Stat. (2023) doi: https://doi.org/10.1007/s00180-023-01330-y

[10] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapmann Hall, London (1989)

[11] Mirzadeh, S., Iranmanesh, A.: A new class of skew-logistic distribution. Math. Sci., **13** 375–385 (2019)

[12] Nagler, J.: Scobit: An Alternative Estimator to Logit and Probit. Am. J. Polit. Anal., **38**, 230–255 (1994)

[13] Shao, Q.: Maximum likelihood estimation for generalized logistic distributions. Commun. Stat. Theory Methods, **31**, 1687–1700 (2002)

[14] Wang, X., Dey, D.K. : Generalised extreme value regression for binary response data: An application to B2B electronic payments system adoption. Ann. Appl. Stat., **4**, 2000–2023 (2010)


DRAFT

# New advances in Regression Forests

Mila Andreani[a], Lea Petrella[b], and Nicola Salvati[c]

[a]Scuola Normale Superiore, Pisa, Italy; mila.andreani@sns.it

[b]MEMOTEF Depart., Sapienza University of Rome, Rome, Italy; lea.petrella@uniroma1.it

[c]Department of Economics and Management, University of Pisa, Pisa, Italy; nicola.salvati@unipi.it

**Abstract**

In this paper we propose a new Mixed-Effects Quantile Regression Forest by generalizing the Quantile Regression Forest approach to longitudinal data. The inferential procedure is based on the Nonparametric Maximum Likelihood exploiting the Asymmetric Laplace distribution tool. The performance of the ME-QRF is tested in a simulation study and compared with the results of standard quantile regression models. Finally, the ME-QRF is applied to a data set for analysing the effect of the tratment on lead-exposed children.

***Keywords:*** Quantile Regression, Random Forests, mixed-effects, longitudinal data

DRAFT

## 1. Introduction

Mixed-effects quantile regression models are used in longitudinal studies to obtain a more complete picture of the response variable distribution with respect to standard linear regression while accounting for serial correlation among observations of the same statistical unit [5; 4; 2; 7]. This paper proposes a novel machine learning algorithm, denoted Mixed- Effects Quantile Regression Forest (ME-QRF) to estimate quantiles of longitudinal data generalizing the Quantile Regression Forest (QRF) algorithm of [9]. The inferential approach is based on the Asymmetric Laplace distribution tool by applying the Non Parametric Maximum Likelihood approach (NPML) of [6] already introduced in a quantile regression framework by [1; 10] to the Quantile Regression Forest contest. In particular, we develop an EM algorithm to estimate quantiles by decoupling the fixed-effects estimation part from the random-effects one without making any parametric assumption. The ME-QRF performance is tested by means of a simulation study and by comparing its performance with standard quantile regression models. The ME-QRF is also applied empirically using a dataset from the study of [13] conducted to assess whether the succimer treatment of children with Blood Lead Levels ($BLL$) $< 45\mu$g/dL is beneficial and safe.

## 2. Methodology

Let $y_{it}, i = 1, \ldots, N, t = 1, \ldots, T_i$ be the response variable for the $i$-th statistical unit observed at time $t$, and $\mathbf{x}_{it} \in \mathbb{R}^p$ be the vector of explanatory variables where $x_{it,1} \equiv 1$.

By indicating with $\tau \in (0, 1)$ the quantile probability level, the standard quantile regression linear mixed-model (LQMM) is:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}_\tau + b_{i,\tau} + \varepsilon_{it} \quad \text{where} \quad Q_\tau(\varepsilon_{ij}|\mathbf{x}_{it}, \boldsymbol{\beta}_\tau, b_{i,\tau}) = 0, \quad \forall\tau \in (0, 1) \tag{1}$$

where $\boldsymbol{\beta}_\tau$ is a vector of $\tau$- dependent regression coefficients common to all statistical units, the term $\mathbf{x}'_{it}\boldsymbol{\beta}_\tau$ is the "fixed-effects" part of the model, whereas the term $b_{i,\tau}$ is the "random-effects" part, represented by a time-constant parameter that varies across statistical units according to a distribution $f_b(\cdot)$ with support $\mathcal{B}$.

Here we avoid making a parametric assumption concerning the fixed-effects part as in (1) and formulate the quantile mixed model as follows:

$$y_{it} = g_\tau(\mathbf{x}_{it}) + b_{i,\tau} + \varepsilon_{it} \quad \text{where} \quad Q_\tau(\varepsilon_{it}|\mathbf{x}_{it}, b_{i,\tau}) = 0, \quad \forall\tau \in (0,1) \tag{2}$$

where $g_\tau : \mathbb{R}^p \to \mathbb{R}$ is a non-parametric unknown function. The terms $g_\tau(\mathbf{x}_{it})$ and $b_{i,\tau}$ are estimated via Maximum Likelihood by means of an EM algorithm based on QRF. We exploit the Asymmetric Laplace ($AL$) distribution as suitable tool [14], where $y_{it} \sim AL(\mu_{it}, \sigma_\tau, \tau)$:

$$f(y_{it}|\mu_{it,\tau}, \sigma_\tau, \tau) = \frac{\tau(1-\tau)}{\sigma_\tau} \exp\left\{ -\rho_\tau\left(\frac{y_{it} - \mu_{it,\tau}}{\sigma_\tau}\right)\right\}, \tag{3}$$

where $\sigma_\tau > 0$ is the scale parameter, the function $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u<0\}})$ is the quantile loss function of [5] and the location parameter $\mu_{it,\tau} = g_\tau(\mathbf{x}_{it}) + b_{i,\tau}$ represents the quantile at level $\tau$.

The observed data likelihood is:

$$L(\boldsymbol{\Phi}_\tau) = \prod_{i=1}^{N}\left\{ \int_{\mathcal{B}} \prod_{t=1}^{T_i} f(y_{it}|\mu_{it,\tau}, \sigma_\tau, \tau) f_b(b_{i,\tau})\, \mathrm{d}b_{i,\tau}\right\} \tag{4}$$

where $\boldsymbol{\Phi}_\tau = \{\sigma, b_1, \ldots, b_N\}$. The main issue concerning the likelihood in (4) is that it involves a multidimensional integral that does not have a closed form solution and that it requires to specify the functional form of $f_b(\cdot)$. Thus, an EM algorithm is developed to estimate $g_\tau(\mathbf{x}_{it})$ with a QRF and to estimate $b_{i,\tau}$ by maximising (4) without making any parametric assumptions about the form of $f_b(\cdot)$.

In line with previous contributions [8; 11; 1], we approximate $f_b(\cdot)$ with a discrete distribution by exploiting the NPML approach of [6]. In particular, we consider a discrete distribution on $K < N$ locations $b_{k,\tau}$ such that $b_{i,\tau} \sim \sum_{k=1}^{K} \pi_{k,\tau}\delta_{b_{k,\tau}}$, where the probability $\pi_{k,\tau}$ is defined as $\pi_{k,\tau} = \mathbb{P}(b_{i,\tau} = b_{k,\tau})$ with $i = 1, \ldots, N$ and $k = 1, \ldots, K$ where $\delta_{b_{k,\tau}}$ is a one-point distribution putting a unit mass at $b_{k,\tau}$. The likelihood (4)is reformulated as:

$$L(\Phi_\tau) = \prod_{i=1}^{N}\left\{ \sum_{k=1}^{K} \prod_{t=1}^{T_i} f(y_{itk}|\mu_{itk,\tau}, \sigma_\tau, \tau)\pi_{k,\tau}\right\}, \tag{5}$$

where $\boldsymbol{\Phi}_\tau = \{\sigma, b_1, \ldots, b_K, \pi_1, \ldots, \pi_K\}$ is the parameter vector.

The next section described the EM algorithm based on (5) and QRF used to obtained $\hat{\Phi}_\tau$.

## 2.1 The EM algorithm

Given that each observation $i$ in (5) can be considered as drawn from one of the $K$ locations of the discrete distribution used to approximate $f_b(\cdot)$, we denote with $w_{ik}$ the indicator variable equal to 1 if the $i$-th unit belongs to the $k$-th component of the finite mixture, and 0 otherwise. The component membership $w_{ik}$ is considered as missing data and, from (5), the complete data log-likelihood is:

$$\ell_c(\boldsymbol{\Phi}_\tau) = \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik,\tau}\Big\{ \sum_{t=1}^{T_i} \log(f(y_{it}|\mu_{itk,\tau}, \sigma_\tau)) + \log(\pi_{k,\tau})\Big\} \tag{6}$$

Estimates $\widehat{g}_\tau(\mathbf{x}_{it})$ and $\widehat{b}_{i,\tau}$ in (2) are obtained from (6) in a EM algorithm by decoupling the fixed-effects estimation, obtained with a QRF, from the random-effects one as follows.

**Initialization** By indicating with $r$ the generic iteration of the algorithm, in the first step $r = 0$, $\widehat{b}^{(0)}_{i,\tau}, \hat{\sigma}^{(0)}_\tau, \hat{\pi}^{(0)}_{k,\tau}, \widehat{g}_\tau(\mathbf{x}_{it})^{(0)}$ are initialised. In particular, the initial value $\widehat{g}_\tau(\mathbf{x}_{it})^{(0)}$ is computed as the $\tau$-th quantile estimated with a QRF fitted with the training set $\mathcal{T}^{(0)} = \{(y_{it}, \mathbf{x}_{it})\}_{\substack{i=1,\ldots,N \\ t=1,\ldots,T_i}}$.

DRAFT

**E-step** The E-step consists in updating $\hat{w}_{ik,\tau}^{(r+1)}$ and $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$. In particular, $\hat{w}_{ik,\tau}^{(r+1)}$ is updated as:

$$\hat{w}_{ik,\tau}^{(r+1)} = \mathbb{E}[w_{ik,\tau}|y_{it}, \mathbf{x}_{it}, \hat{\Phi}_\tau^{(r)}] = \frac{\prod_{t=1}^{T_i} f_{itk,\tau}^{(r)} \hat{\pi}_{k,\tau}^{(r)}}{\sum_{l=1}^{K} \prod_{i=1}^{T_i} f_{itl,\tau}^{(r)} \hat{\pi}_{l,\tau}^{(r)}}, \tag{7}$$

where $f_{itk,\tau}^{(r)}$ is the response variable distribution when considering the $k$ -th component of the finite mixture.

The estimate $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$ is updated by decoupling the random-effects from the fixed-effects. To this end, $\widehat{g}_\tau(\mathbf{x}_{it})^{(r+1)}$ is estimated with the QRF fitted using the training set $\mathcal{T}^{(r+1)} = \left\{ \left( y_{it}^{*(r+1)}, \mathbf{x}_{it} \right) \right\}_{\substack{i=1,\dots,N \\ t=1,\dots,T_i}}$, in which $y_{it}^{*(r+1)} = y_{it} - \widehat{b}_{i,\tau}^{(r)}$.

**M-step** In the M-step, numerical optimisation techniques are applied to maximise $\mathbb{E}[\ell_c(\mathbf{\Phi}_\tau)|y_{it}, \mathbf{x}_{it}, \hat{\mathbf{\Phi}}_\tau^{(r)}]$ with respect to $\hat{\sigma}_\tau$ and $\widehat{b}_{k,\tau}$ .

The E- and M-steps are alternated iteratively until convergence.

## 3. Simulation study

This section reports the results of a simulation study carried out to assess the performance of the ME-QRF in a non-linear setting. To this end, the ME-QRF is used to predict quantiles at levels $\tau \in \{0.1, 0.5, 0.9\}$ of an outcome variable simulated under the following non-linear data generating process (DGP) [3]:

$$y_{it} = g(\mathbf{x}_{it}) + b_i + \varepsilon_{it} \quad \text{where} \quad g(\mathbf{x}_{it}) = 2x_{it,1} + x_{it,1}^2 + 4 \cdot \mathbf{1}_{\{x_{it,3}>0\}} + 2x_{it,3} \log|x_{it,1}|$$

The covariates are generated as $x_{it,1},\ x_{it,2},\ x_{it,3} \sim \mathcal{N}(0,1)$. The random-effects parameters and the error terms are generated independently according to two DGPs:

$$\textbf{(NN)}\ b_i \sim N(0,1),\ \ \varepsilon_{it} \sim N(0,1) \qquad \textbf{(TT)}\ b_i \sim t(3),\ \ \varepsilon_{it} \sim t(3)$$

As in [3], for each scenario we consider a training set of 500 observation for $N = 100$ statistical units and $T_i = 5$ measurements each, and an unbalanced test set with $T_i \in \{9, 27, 45, 63, 81\}$ for a total of 4500 observations. Each scenario has been replicated $S = 100$ times.

The average performance of the ME-QRF across the 100 replications is assessed in terms of Average Mean Absolute Error (MAE) and average Mean Squared Error (MSE) with respect to the theoretical quantile of the DGP, computed as in [12]:

$$MAE_\tau = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}|Q_{it,\tau}^s - \hat{Q}_{it,\tau}^s| \qquad MSE_\tau = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}(Q_{it,\tau}^s - \hat{Q}_{it,\tau}^s)^2$$

where $Q_{it,\tau}^s = Q_\tau^s(y_{it}|\mathbf{x}_{it})$ and $\hat{Q}_{it,\tau}^s = \hat{Q}_\tau^s(y_{it}|\mathbf{x}_{it})$ are respectively the theoretical and estimated conditional quantiles of variable $y_{it}$ at level $\tau$ of the $s$-th simulated dataset.

The ME-QRF is compared with three benchmark models: LQMM, Quantile Random Forest (QRF) and the Quantile Mixed Model (QMM) of [10]. The latter model exploits the same methodological approach of the ME-QRF in a linear setting. Results are reported in Table 1.

DRAFT

| | | $\tau = 0.1$ | | | | $\tau = 0.5$ | | | | $\tau = 0.9$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **ME-QRF** | **LQMM** | **RF** | **QMM** | **ME-QRF** | **LQMM** | **RF** | **QMM** | **ME-QRF** | **LQMM** | **RF** | **QMM** |
| **NN** | **MAE** | **1.83** | 1.86 | 2.64 | 4.67 | **1.65** | 1.72 | 2.11 | 2.80 | 1.67 | **1.57** | 2.20 | 5.16 |
| | **MSE** | **5.87** | 5.89 | 10.98 | 40.62 | **4.62** | 5.27 | 7.15 | 14.61 | 4.86 | **4.34** | 7.93 | 61.12 |
| **TT** | **MAE** | **1.80** | 1.94 | 2.02 | 4.11 | **1.43** | 1.57 | 1.44 | 2.01 | **1.87** | 2.06 | 2.06 | 5.13 |
| | **MSE** | 6.66 | **6.42** | 7.88 | 33.16 | **4.32** | 4.48 | 4.51 | 7.82 | 7.10 | **6.81** | 8.02 | 50.79 |

**Table 1:** Loss values for each scenario computed on the test set of the four fitted models. Values in bold indicate the smallest loss.

The results highlight that the ME-QRF outperforms the benchmark models at almost all quantile levels in each scenario, especially when the data violate the Gaussianity assumptions. The only exception is represented by the quantile at $\tau = 0.9$ for the NN scenario. In this case, the LQMM outperforms the ME-QRF since the Gaussianity assumptions of the LQMM model hold. The ME-QRF and the LQMM perform similarly in terms of MSE and represent the two models with the lowest MAE and MSE values.

## 4. Empirical Application

In this section, the ME-QRF is applied to a dataset from a placebo-controlled, double-blind, randomised trial to study whether the succimer treatment of children with Blood Lead Levels (BLL) $<$ 45$\mu$g/dL is beneficial and safe. Following [1], three quantile levels are considered $\tau \in \{0.25, 0.5, 0.75\}$.

The dataset includes $T_i = 4$ weekly measurements of BLL for $N = 100$ children with BLL of 20–44 $\mu$g/dL. The covariates are the dummy variable Treatment ($R_i$) taking value 1 for children that have been treated and 0 otherwise, and Time $W_{it} \in \{0, 1, 4, 6\}$ representing week 0 -baseline-, week 1, week 4 and week 6. Given the results of the simulation study, the ME-QRF has been compared in terms of quantile loss with the LQMM with the following formulation:

$$BLL_{it} = \beta_{i1} R_i + \beta_{i2} W_{it}^2 + \beta_{i3}(R_i * W_{it}^2) + \beta_{i4}(R_i * W_{it}) + b_i \tag{8}$$

The quantile loss of [5] is computed as follows and the related results are presented in Table 2:

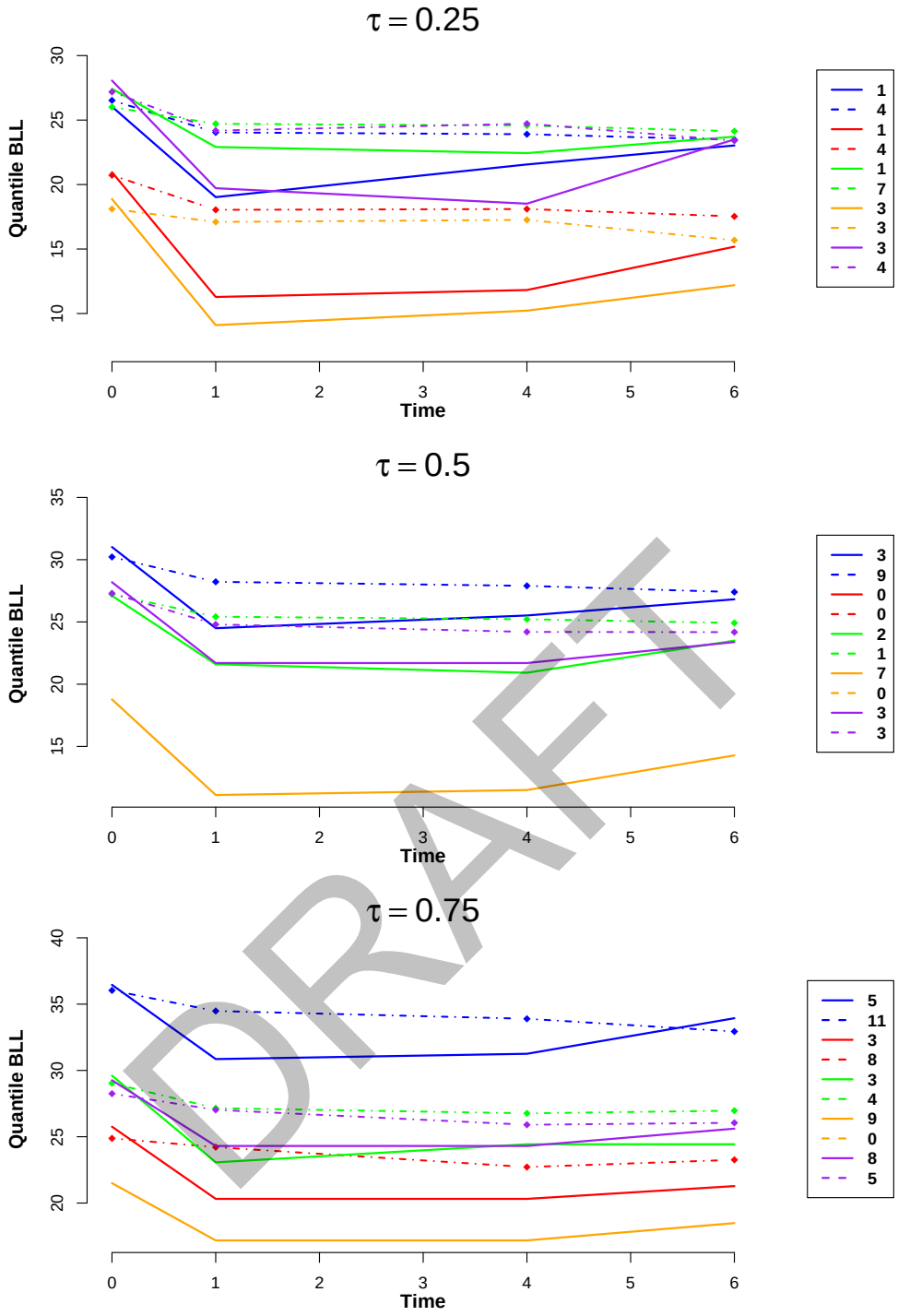
$$QLOSS_\tau = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} u_{it}(\tau - \mathbf{1}_{\{u_{it}<0\}}) \quad \text{where} \quad u_{it} = y_{it} - \hat{Q}_{it,\tau}. \tag{9}$$

| $\tau$ | **0.25** | **0.5** | **0.75** |
|---|---|---|---|
| QLOSS$^{\text{ME-QRF}}$ | **1.22** | **1.55** | **1.31** |
| QLOSS$^{\text{LQMM}}$ | 1.61 | 1.82 | 1.59 |

**Table 2:** Results in terms of QLOSS for the treatment of lead-exposed children dataset.

Results show that our model outperforms the LQMM at each quantile level. Figure 1 depicts the treatment and control group quantile trajectories estimated with the ME-QRF for each mixture component at level $\tau \in (0.25, 0.5, 0.75)$. Each trajectory is colour-coded according to the mixture component and the treatment and control group of each component are identified with the solid and dashed lines, respectively. The legend reports the number of statistical units in the control and treatment group of each mixture component. The trajectories estimated with our model are coherent with the findings of [1].

DRAFT




DRAFT

**Figure 1:** Estimated trajectories for ME-QRF for the first five mixture components. Each component is indicated with one colour, and the treatment (solid curves) and control (dashed lines) groups are represented separately. The legend reports the number of statistical units belonging to each mixture component.

## 5. Conclusions

This paper introduces the Mixed-Effects Quantile Regression Forest (ME-QRF) model, which combines QRF and mixed-models to estimate quantiles of longitudinal data wotjoud any parameteric as-

sumption on the fixed-effects and the random-effects distribution. Simulation results highlight that the ME-QRF outperforms benchmark models in non-linear settings, especially when Gaussianity assumptions are violated. The ME-QRF is applied empirically using data from the study of [13] in order to assess the effectiveness of the succimer treatment on lead-esposed children. Results show that the ME-QRF outperforms the LQMM model in terms of quantile loss and that findings are coherent with the ones presented in the previous literature.

## References


[1] Alfò, M., Salvati, N., and Ranallli, M. G. (2017). Finite mixtures of quantile and m-quantile regression models. *Statistics and Computing*, 27(2):547–570.

[2] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

[3] Hajjem, A., Bellavance, F., and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328.

[4] Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

[5] Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

[6] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.

[7] Liu, Y. and Bottai, M. (2009). Mixed-effects models for conditional quantiles with longitudinal data. *The International Journal of Biostatistics*, 5(1).

[8] Marino, M. F., Tzavidis, N., and Alfò, M. (2018). Mixed hidden markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical methods in medical research*, 27(7):2231–2246.

[9] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

[10] Merlo, L., Maruotti, A., and Petrella, L. (2021). Two-part quantile regression models for semi-continuous longitudinal data: A finite mixture approach. *Statistical Modelling*, page 1471082X21993603.

[11] Merlo, L., Petrella, L., and Tzavidis, N. (2022). Quantile mixed hidden markov models for multivariate longitudinal data: An application to children's strengths and difficulties questionnaire scores. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*.

[12] Min, I. and Kim, I. (2004). A monte carlo comparison of parametric and nonparametric quantile regressions. *Applied Economics Letters*, 11(2):71–74.

[13] Rogan, W., Bornschein, R., Chisolm, J., Damokosh, A., Dockery, D., Fay, M., Jones, R., Rhoads, G., Ragan, N., Salganik, M., et al. (2000). Safety and efficacy of succimer in toddlers with blood lead levels of 20-44 $\mu$g/dl. *Pediatric Research*, 48(5):593–599.

[14] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.


DRAFT

# Using expectile regression with latent variables for digital assets

Beatrice Foroni[a], Luca Merlo[b], and Lea Petrella[a]

[a]MEMOTEF Department, Sapienza University of Rome; beatrice.foroni@uniroma1.it, lea.petrella@uniroma1.it

[b]Department of Human Sciences, European University of Rome; luca.merlo@unier.it

**Abstract**

In this paper we introduce a linear expectile hidden Markov model with the goal of modeling the entire conditional distribution of asset returns and, at the same time, to grasp unobserved serial heterogeneity and rapid volatility jumps typical of financial time series. The temporal evolution of asset returns is captured by introducing time-dependent coefficients evolving according to a latent discrete homogeneous Markov chain. To implement the procedure, we consider the Asymmetric Normal distribution as a working likelihood for the estimation of model parameters and the estimation procedure is carried out using an efficient EM algorithm. The empirical application investigates the relationship between daily Bitcoin returns and major world market indices.

***Keywords:*** Bitcoin, financial time series, hidden Markov models, tail risk

DRAFT

## 1. Introduction

The recent exploit of the cryptocurrency market have attracted investors and risk managers as never before. The enormous price jumps and levels of high volatility of these digital assets, and Bitcoin in particular, caused by speculative behaviors have threatened the stability of financial markets (19), and research and practitioners have recently started to investigate the peculiar characteristics of cryptocurrencies, as well as their relationship with tradition financial markets. Many contributions to this strand of literature rely on well-known econometric techniques such as GARCH models (9), variance decomposition (6; 21) and Granger causality test (4). Rather than investigating measures of conditional central tendencies, fewer works have focused on the tails of returns distribution. For instance, (15) focused on the tail connectedness among major cryptocurrencies in extreme downward and upward market conditions using LASSO penalized quantile regressions, while (22) apply a risk spillover approach based on generalized quantiles, showing the existence of a downside risk spillover between Bitcoin and traditional assets. From a risk management perspective, it is of extreme importance to be able to investigate the dynamics of extreme occurrences. Since the seminal work of (10), quantile regression has represented one of the most used approaches for modeling the entire distribution of returns while accounting for the well-known stylized facts, i.e., high kurtosis, skewness and serial correlation, that typically characterize financial assets. Since then, several generalizations of the concept of quantiles have been presented, among which we find expectile regression (14), which, similar to quantile regression, allows to describe the entire conditional distribution of a response variable based on an asymmetric squared loss function. Despite having a more difficult interpretation, expectiles possess several advantages, both from an informative and a computational point of view. In particular, the asymmetric squared loss is continuously

differentiable, which makes the estimators and their covariance matrix easier to compute using fast and efficient algorithms. In the context of risk management, expectiles have gained an important role as potential competitors to the Value at Risk (VaR) and the Expected Shortfall measures, and indeed possess several interesting properties in terms of risk measures, being the only risk measure that is both coherent and elicitable (11; 23). However, homogeneous regression models are not able to capture the volatility clustering behavior that often financial time series exhibit. In this context, hidden Markov models (HMMs) have been intensively employed to characterize temporal evolution of returns distribution, modeling volatility regime shifting trough a latent Markov chain. Since (8), different works have combined the quantile framework with HMMs by introducing in the model parameters that vary according to the outcome of a latent Markov process (20; 12; 13). To the best of our knowledge, however, an expectile hidden Markov regression model has not been explored yet. In this paper we introduce an expectile regression model to analyze the entire conditional distribution of Bitcoin returns where the dynamics of returns over time is described by state-specific regression coefficients which follow a latent discrete homogeneous Markov chain. The proposed model contributes to the existing literature regarding the relations between cryptocurrencies and traditional asset class to control for potential inherent risks related to the participation in crypto exchanges. As usual for latent variable models, inference is carried out in a Maximum Likelihood (ML) approach using an Expectation-Maximization (EM) algorithm based on the asymmetric normal distribution of (17) as working likelihood. In the empirical analysis, we model daily Bitcoin log-returns as a function of major stock and global market indices, including Crude Oil, Standard & Poor's 500 (S&P500), Gold COMEX daily closing prices and the Volatility Index (VIX) from September 2014 until October 2022.

The rest of the paper is organized as follows. In Sect. 2. we specify the proposed model with the EM algorithm for estimating the model parameters and the computational aspects. Sect. 3. discusses the results obtained and concludes.

DRAFT

## 2. Model Specification and Inference

In this section we describe the proposed expectile hidden Markov regression model. Formally, let $\{S_t\}_{t=1}^T$ be a latent, homogeneous, first-order Markov chain defined on the discrete state space $\{1,\dots,K\}$. Let $\pi_k = Pr(S_1 = k)$ be the initial probability of state $k$, $k = 1,\dots,K$, and $\pi_{k|j} = Pr(S_{t+1} = k|S_t = j)$, with $\sum_{k=1}^K \pi_{k|j} = 1$ and $\pi_{k|j} \geq 0$, denote the transition probability between states $j$ and $k$, that is, the probability to visit state $k$ at time $t+1$ from state $j$ at time $t$, $j,k = 1,\dots,K$ and $t = 1,\dots,T$. More concisely, we collect the initial and transition probabilities in the $K$-dimensional vector $\boldsymbol{\pi}$ and in the $K \times K$ matrix $\boldsymbol{\Pi}$, respectively. To build the proposed model, let $Y_t$ denote a continuous observable response variable and $\boldsymbol{X}_t = (1, X_{t2},\dots,X_{tP})'$ be a vector of $P$ exogenous covariates, with the first element being the intercept, at time $t = 1,\dots,T$.

For a given expectile level $\tau \in (0,1)$, the proposed linear expectile hidden Markov model is defined as follows:

$$Y_t = \boldsymbol{X}_t'\boldsymbol{\beta}_k(\tau) + \epsilon_{tk}(\tau), \tag{1}$$

where $\mu_{tk} = \boldsymbol{X}_t'\boldsymbol{\beta}_k(\tau)$ defines the linear expectile model, $\boldsymbol{\beta}_k(\tau) = (\beta_{1k}(\tau),\dots,\beta_{Pk}(\tau))' \in \mathbb{R}^P$ is the state-specific coefficient vector that assumes one of the values $\{\boldsymbol{\beta}_1(\tau),\dots,\boldsymbol{\beta}_K(\tau)\}$ depending on the outcome of the unobservable Markov chain $S_t$ and $\epsilon_{tk}(\tau)$ is the error term whose conditional $\tau$-th expectile is assumed to be zero. When $\tau = \frac{1}{2}$, expectile regression reduces to the standard mean regression while, when $\tau \neq \frac{1}{2}$, the regression targets the entire conditional distribution of the response given the covariates. The estimation of the model parameters is carried on through a Maximum Likelihood approach. We employ the Asymmetric Normal (AN) distribution, originally introduced by (16), to describe the conditional distribution of the response given covariates and the state occupied by the latent process at time $t$, whose probability density function is given by

$$f_Y(y_t|\boldsymbol{X}_t = \boldsymbol{x}_t, S_t = k) = \frac{2\sqrt{\tau(1-\tau)}}{\sqrt{\pi\sigma_k^2}(\sqrt{\tau}+\sqrt{1-\tau})} \exp\left[-\omega_\tau\left(\frac{y_t - \mu_{tk}}{\sigma_k}\right)\right], \tag{2}$$

where $\omega_\tau(\cdot)$ is the expectile loss function defined as $\omega_\tau(u) = u^2|\tau - \mathbb{I}(u < 0)|$, which assigns weights $\tau$ and $1 - \tau$ to positive and negative deviations, respectively, and $\mathbb{I}(\cdot)$ denotes the indicator function. The location parameter $\mu_{tk}$ is defined by the linear model $\mu_{tk} = \boldsymbol{x}_t'\boldsymbol{\beta}_k(\tau)$ and corresponds to the $\tau$-th expectile, $\sigma_k > 0$ is a scale parameter and $\tau \in (0, 1)$ determines the asymmetry of the distribution. Particularly, when $\tau = \frac{1}{2}$ the density in eq. (2) reduces to the well-known normal distribution, and $\mu_{tk}$ and $\sigma_k$ coincide with its mean and standard deviation, respectively. The use of this distribution is deemed to be as a likelihood inferential tool for estimating the model parameters in a regression framework rather a parametric assumption. As common for latent variable models, and HMMs in particular, inference on model parameters is made through the development of an EM algorithm (1). To ease the notation, unless specified otherwise, hereinafter we omit the expectile level $\tau$, yet all model parameters are allowed to depend on it. The complete log-likelihood of the proposed model is defined as follows for a given number of hidden states $K$:

$$\begin{aligned}\ell_c(\boldsymbol{\theta}_\tau) = \sum_{k=1}^{K} \gamma_1(k) \log \pi_k + \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K} \xi_t(j,k) \log \pi_{k|j} \\ + \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t(k) \log f_Y(y_t|\boldsymbol{x}_t, S_t = k),\end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_\tau = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1, \ldots, \sigma_K, \boldsymbol{\pi}, \boldsymbol{\Pi})$ represents the vector of all model parameters, $\gamma_t(k)$ denotes a dummy variable equal to 1 if the latent process is in state $k$ at occasion $t$ and 0 otherwise, and $\xi_t(j,k)$ is a dummy variable equal to 1 if the process is in state $j$ in $t-1$ and in state $k$ at time $t$ and 0 otherwise.

To estimate $\boldsymbol{\theta}_\tau$, the algorithm iterates between the E- and M-steps until convergence, as briefly showed in what follows. In the E-step, at the generic $(h+1)$-th iteration, the unobservable indicator variables $\gamma_t(k)$ and $\xi_t(j,k)$ in eq. (3) are replaced by their conditional expectations given the observed data and the current parameter estimates $\boldsymbol{\theta}_\tau^{(h)}$. To compute such quantities one can use the Forward-Backward algorithm of (18). Then, we use these to calculate the conditional expectation of the complete log-likelihood function in eq. (3) given the observed data and the current estimates. In particular, in the M-step update of the regression coefficients is obtained by using Iteratively Reweighted Least Squares for cross-sectional data with appropriate weights.

The EM algorithm is initialized by assigning the initial states partition, $\{S_t^{(0)}\}_{t=1}^T$, to a Multinomial distribution with probabilities $1/K$. From the generated partition, the elements of $\boldsymbol{\Pi}^{(0)}$ are computed as proportions of transition, while we obtain $\boldsymbol{\beta}_k^{(0)}$ and $\sigma_k^{(0)}$ by fitting mean regressions on the observations within state $k$. A multiple random starts strategy is adapted to deal with the possibility of multiple roots. Once we computed the ML estimate of the model parameters, to estimate the standard errors we employ a parametric bootstrap scheme, refitting the model to $R$ bootstrap samples and approximating the standard error of each model parameter with the corresponding standard deviation of the bootstrap estimates.

## 3. Main Results and Conclusions

The empirical analysis is based on the log-returns of Bitcoin, Crude Oil, S&P500, Gold COMEX daily closing prices and the VIX from September 2014 to October 2022. We consider the following model with the idea of providing insights into the temporal evolution of Bitcoin returns and its relationship with traditional global financial assets

$$\mu_{tk}^{Bitcoin} = \beta_{1k}(\tau) + \beta_{2k}(\tau) r_t^{Crude\ Oil} + \beta_{3k}(\tau) r_t^{S\&P500} + \beta_{4k}(\tau) r_t^{Gold} + \beta_{5k}(\tau) r_t^{VIX}, \tag{4}$$

with $\mu_{tk}^{Bitcoin}$ corresponding to the $\tau$-th conditional expectile of Bitcoin return at time $t$ in state $k$, while $r_t^{Crude\ Oil}$ denotes the return of the same date for Crude Oil, and similarly for the other indices. We fit the proposed model for two values of $K$, representing high and low volatility market conditions, at three expectile levels $\tau = \{0.10, 0.50, 0.90\}$, which allow us to focus on both downside and upside

DRAFT

risks. For the selected models, we report the clustering results in Figure 1 at $\tau = 0.50$ expectile level. The plot shows the time series of Bitcoin daily returns colored according to the estimated posterior probability of class membership, $\max_k \gamma_t(k)$, with the vertical dashed lines representing globally relevant events such as the Chinese stock market crash in 2015, the cryptocurrencies crash at the beginning of 2018, the COVID-19 market crash in March 2020, Biden's election at the USA presidency in November 2020 and the Russian invasion of Ukraine at the beginning of 2022. Here we clearly see that the latent components can be associated to specific market regimes characterized by low and high volatility periods. Specifically, light-blue points (State 1) tend to identify low returns, while dark-blue ones (State 2) correspond to periods of extreme positive and negative returns. Table 1 shows the parameter estimates along with the standard errors (in brackets) computed by using the parametric bootstrap technique over $R = 1000$ resamples, as illustrated in Sec. 2. As it happens in the quantile regression framework, the state-specific intercepts are increasing somewhat with $\tau$, with State 1 having lower values than State 2 for all $\tau$'s. Moving forward with the analysis, at $\tau = 0.50$ we observe none or few interactions among Bitcoin and financial assets, during low and high volatility states, respectively. In particular, S&P500 and Gold significantly influence the mean of Bitcoin only in the second state, highlighting a weak hedge behavior of the crypto-asset during tranquil periods and confirming results founded in (2; 5). If we move to the tails of return distributions, in the not-at-risk state (State 1) at the extreme left-tail ($\tau = 0.10$) the S&P500, Gold and the VIX index positively influence Bitcoin returns, while only S&P500 and Gold significantly influence the right-tail ($\tau = 0.90$) expectiles of the cryptocurrency, exposing a connection during high volatility periods between traditional financial markets and Bitcoin both for negative and positive returns. In the at-risk state (State 2) we observe a positive influence of the S&P500 and Gold across the conditional distribution of returns. Also, one can see that Crude Oil is negatively associated with the crypto returns at the 10-th expectile. This finding is in line with (3) but it is contrary to the works of (7) and (6), which may be due to the events and crises occurred in the last years.

In conclusion, we developed a linear expectile hidden Markov model for the analysis of time series where temporal behaviors of the data are captured via time-dependent coefficients following an unobservable discrete homogeneous Markov chain. The proposed method enables us to model the entire conditional distribution of asset returns and, at the same time, to grasp unobserved serial heterogeneity and rapid volatility jumps that would otherwise go undetected. With this model we strengthen the existing literature in this field, contributing towards a deeper understanding of the interrelations between Bitcoin and traditional financial markets.

DRAFT

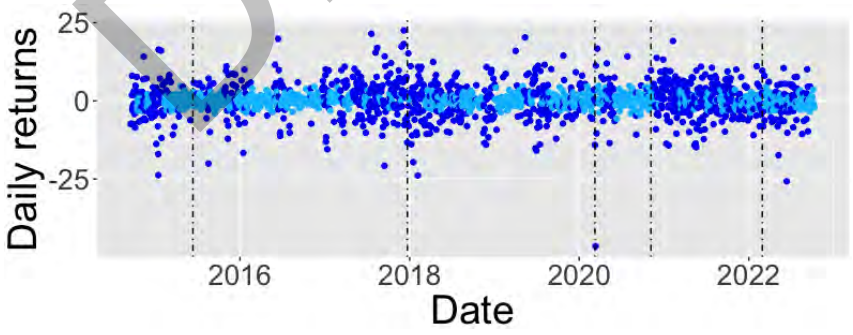



Figure 1: Bitcoin returns series classified according to the estimated posterior probability of class membership at $\tau = 0.50$. Vertical dashed lines indicate globally relevant events in the financial markets that occurred in 2015,06; 2017,12; 2020,03; 2020,11; and 2022,02.

| | Intercept | Crude Oil | S&P500 | Gold | VIX | $\sigma_k$ |
|---|---|---|---|---|---|---|
| State 1 | | | | | | |
| $\tau = 0.10$ | **-1.036 (0.280)** | 0.024 (0.021) | **0.595 (0.096)** | **0.189 (0.072)** | **0.029 (0.012)** | 1.433 (0.040) |
| $\tau = 0.50$ | 0.122 (0.158) | 0.031 (0.072) | 0.409 (0.383) | 0.263 (0.249) | 0.009 (0.036) | 1.695 (0.062) |
| $\tau = 0.90$ | **1.297 (0.061)** | -0.009 (0.020) | **0.589 (0.088)** | **0.134 (0.065)** | 0.014 (0.011) | 1.335 (0.041) |
| State 2 | | | | | | |
| $\tau = 0.10$ | **-6.52 (0.060)** | **-0.256 (0.096)** | **2.072 (0.476)** | **1.032 (0.320)** | -0.055 (0.058) | 4.964 (0.157) |
| $\tau = 0.50$ | 0.242 (0.092) | -0.056 (0.055) | **1.087 (0.357)** | **0.613 (0.214)** | -0.025 (0.026) | 6.164 (0.169) |
| $\tau = 0.90$ | **6.244 (0.229)** | 0.017 (0.079) | **0.948 (0.291)** | **0.835 (0.249)** | -0.002 (0.041) | 4.692 (0.128) |

Table 1: State-specific parameter estimates for three expectile levels, with bootstrapped standard errors (in brackets) obtained over 1000 replications. Point estimates are displayed in boldface when significant at the standard 5% level.

## References


[1] Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. of Math. Stat. **41**(1), 164–171 (1970)

[2] Bouri, E., Jalkh, N., Molnár, P., Roubaud, D.: Bitcoin for energy commodities before and after the december 2013 crash: diversifier, hedge or safe haven? Appl. Econ. **49**(50), 5063–5073 (2017)

[3] Bouri, E., Lucey, B., Roubaud, D.: Cryptocurrencies and the downside risk in equity investments. Fin. Res. Lett. **33**, 101,211 (2020)

[4] Bouri, E., Lucey, B., Roubaud, D.: The volatility surprise of leading cryptocurrencies: Transitory and permanent linkages. Fin. Res. Lett. **33**, 101,188 (2020)

[5] Bouri, E., Molnár, P., Azzi, G., Roubaud, D., Hagfors, L.I.: On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier? Fin. Res. Lett. **20**, 192–198 (2017)

[6] Corbet, S., Meegan, A., Larkin, C., Lucey, B., Yarovaya, L.: Exploring the dynamic relationships between cryptocurrencies and other financial assets. Econ. Lett. **165**, 28–34 (2018)

[7] Dyhrberg, A.H.: Hedging capabilities of Bitcoin. Is it the virtual gold? Fin. Res. Lett. **16**, 139–144 (2016)

[8] Farcomeni, A.: Quantile regression for longitudinal data based on latent markov subject-specific parameters. Stat. and Comput. **22**(1), 141–152 (2012)

[9] Guesmi, K., Saadi, S., Abid, I., Ftiti, Z.: Portfolio diversification with virtual currency: Evidence from Bitcoin. Int. Rev. of Fin. Anal. **63**, 431–437 (2019)

[10] Koenker, R., Bassett, G.: Regression quantiles. Econ.: J. of the Econ. Soc. **46**(1), 33–50 (1978)

[11] Lambert, N.S., Pennock, D.M., Shoham, Y.: Eliciting properties of probability distributions. In: Proc. of the 9th ACM Conf. on Electron. Commer., pp. 129–138. ACM (2008)

[12] Maruotti, A., Petrella, L., Sposito, L.: Hidden semi-Markov-switching quantile regression for time series. Comp. Stat. & Data Anal. **159**, 107,208 (2021)

[13] Merlo, L., Maruotti, A., Petrella, L., Punzo, A.: Quantile hidden semi-Markov models for multivariate time series. Stat. and Comp. **32**(4), 1–22 (2022)

[14] Newey, W.K., Powell, J.L.: Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society pp. 819–847 (1987)

[15] Shahzad, S.J.H., Bouri, E., Ahmad, T., Naeem, M.A.: Extreme tail network analysis of cryptocurrencies and trading strategies. Fin. Res. Lett. **44**, 102,106 (2022)

[16] Waldmann, E., Sobotka, F., Kneib, T.: Bayesian geoadditive expectile regression. arXiv preprint arXiv:1312.5054 (2013)

[17] Waldmann, E., Sobotka, F., Kneib, T.: Bayesian regularisation in geoadditive expectile regression. Stat. and Comp. **27**(6), 1539–1553 (2017)

[18] Welch, L.R.: Hidden Markov models and the Baum-Welch algorithm. IEEE Inf. Theory Soc. Newsl. **53**(4), 10–13 (2003)

[19] Yarovaya, L., Brzeszczyński, J., Lau, C.K.M.: Intra-and inter-regional return and volatility spillovers across emerging and developed markets: Evidence from stock indices and stock index


DRAFT

futures. Intern. Rev. of Fin. Anal. **43**, 96–114 (2016)
[20] Ye, W., Zhu, Y., Wu, Y., Miao, B.: Markov regime-switching quantile regression models and financial contagion detection. Insur.: Math. and Econ. **67**, 21–26 (2016)
[21] Yi, S., Xu, Z., Wang, G.J.: Volatility connectedness in the cryptocurrency market: Is Bitcoin a dominant cryptocurrency? Intern. Rev. of Fin. Anal. **60**, 98–114 (2018)
[22] Zhang, Y.J., Bouri, E., Gupta, R., Ma, S.J.: Risk spillover between bitcoin and conventional financial markets: An expectile-based approach. The N. Am. J. of Econ. and Fin. **55**, 101,296 (2021)
[23] Ziegel, J.F.: Coherence and elicitability. Math. Fin. **26**(4), 901–918 (2016)

DRAFT