11th Scientific Meeting of the SIS Group
"Statistics for the Evaluation and Quality in Services"

BOOK OF SHORT PAPERS

Editors
Andrea Bucci
Alfredo Cartone
Adelia Evangelista
Andrea Marletta

IES PESCARA 2023

STATISTICAL METHODS FOR EVALUATION AND QUALITY: TECHNIQUES, TECHNOLOGIES AND TRENDS (T^3)

EDIZIONI
Il Viandante

IES 2023 - Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)

BOOK OF SHORT PAPERS

Editors: Andrea Bucci, Alfredo Cartone, Adelia Evangelista and Andrea Marletta

Book of Short papers
11th International Conference **IES 2023**
Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)

University 'G. d'Annunzio' of Chieti-Pescara

Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services - SVQS

Maurizio Carpita	University of Brescia
Pietro Amenta	University of Sannio
Matilde Bini	European University of Rome
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Luigi D'Ambra	University of Naples 'Federico II'
Paolo Mariani	University of Milan 'Bicocca'
Marica Manisera	University of Brescia
Monica Palma	University of Salento
Pasquale Sarnacchiaro	University of Naples 'Federico II'

Program Committee of the conference IES 2023

Chair: Paolo Mariani	University of Milan 'Bicocca'
Matilde Bini	European University of Rome
Mario Bolzan	University of Padova
Giuseppe Bove	University of Rome 3
Eugenio Brentari	University of Brescia
Ida Camminatiello	University of Campania 'L. Vanvitelli'
Maurizio Carpita	University of Brescia
Carlo Cavicchia	Erasmus University Rotterdam
Roy Cerqueti	University of Rome 'La Sapienza'
Enrico Ciavolino	University of Salento
Corrado Crocetta	University of Bari 'Aldo Moro'
Antonello D'Ambra	University of Campania 'L. Vanvitelli'
Pierpaolo D'Urso	University of Rome 'La Sapienza'
Tonio Di Battista	University 'G. d'Annunzio' of Chieti-Pescara
Simone di Zio	University 'G. d'Annunzio' of Chieti-Pescara
Mauro Ferrante	University of Palermo
Francesca Fortuna	University of Rome 3
Michele Gallo	University of Naples 'Orientale'
Claudiu Herteliu	Bucharest University of Economic Studies
Salvatore Ingrassia	University of Catania
Ron Kenett	Samuel Neaman Institute & KPA Group
Michele La Rocca	University of Salerno
Rosaria Lombardo	University of Campania 'L. Vanvitelli'
Antonio Lucadamo	University of Sannio
Marica Manisera	University of Brescia
Andrea Marletta	University of Milan 'Bicocca'
Lucio Masserini	University of Pisa
Fabrizio Maturo	University Mercatorum
Matteo Mazziotta	ISTAT
Rodolfo Metulini	University of Bergamo
Stefania Mignani	University of Bologna
Enrico Ripamonti	University of Brescia

Silvia Salini	University of Milano
Pasquale Sarnacchiaro	University of Naples ‘Federico II’
Violetta Simonacci	University of Naples ‘Federico II’
Maurizio Vichi	University of Rome ‘La Sapienza’
Mariangela Zenga	University of Milan ‘Bicocca’

Organizing Committee

Chair: Tonio Di Battista	University ‘G. d’Annunzio’ of Chieti-Pescara
Andrea Bucci	University ‘G. d’Annunzio’ of Chieti-Pescara
Alfredo Cartone	University ‘G. d’Annunzio’ of Chieti-Pescara
Marco Di Marzio	University ‘G. d’Annunzio’ of Chieti-Pescara
Marta Di Nicola	University ‘G. d’Annunzio’ of Chieti-Pescara
Simone Di Zio	University ‘G. d’Annunzio’ of Chieti-Pescara
Stefania Fensore	University ‘G. d’Annunzio’ of Chieti-Pescara
Lara Fontanella	University ‘G. d’Annunzio’ of Chieti-Pescara
Stefano Antonio Gattone	University ‘G. d’Annunzio’ of Chieti-Pescara
Luigi Ippoliti	University ‘G. d’Annunzio’ of Chieti-Pescara
Eugenia Nissi	University ‘G. d’Annunzio’ of Chieti-Pescara
Paolo Postiglione	University ‘G. d’Annunzio’ of Chieti-Pescara
Annalina Sarra	University ‘G. d’Annunzio’ of Chieti-Pescara
Pasquale Valentini	University ‘G. d’Annunzio’ of Chieti-Pescara
Adelia Evangelista (Secretariat)	University ‘G. d’Annunzio’ of Chieti-Pescara

Editors

Andrea Bucci	University of Macerata
Alfredo Cartone	University 'G. d'Annunzio' of Chieti-Pescara
Adelia Evangelista	University 'G. d'Annunzio' of Chieti-Pescara
Andrea Marletta	University of Milan 'Bicocca'

Sede legale: Via Silvino Olivieri, 111 - 66100 Chieti (CH)

ISBN 979-12-803-3369-8
DOI 10.60984/978-88-94593-36-5-IES2023
https://doi.org/10.60984/978-88-94593-36-5-IES2023

Sponsored by

Preface

Statistical thinking, design and analysis play a crucial role in social life and are useful to society at large. Besides, promoting advanced methodological research is useful to facilitate the dissemination of ideas related to various fields of interest. For this purpose, experts in statistics, data analysis, data mining, statistical methods for decision making, machine learning and related methods come together to understand and analyse phenomena through data.

In line with this objective, the Statistics Group for the Evaluation and Quality of Services (SVQS; www.svqs.it) of the Italian Statistical Society (SIS) has been organizing the Innovation and Society (IeS) conference biennially since 2009, focusing on new developments and ideas in statistics applied to the evaluation and quality of public and private services, attracting national and international statisticians and data scientists. The meeting contributes to spot light on the main statistical approaches and methodologies for the evaluation of public services currently in use in different contexts, as well as to facilitate discussion on the impact of innovative statistical evaluation systems for these services, involving various economic and social policy actors.

The conference "Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)" recorded valuable contributions that are reported in this volume. The papers underscore how the growing availability of data has tasked social and economic actors, organizations, and researchers with the management and analysis of large volumes of unstructured and heterogeneous data. In recent years, many tools for both qualitative and quantitative models have been developed to better describe and understand complex systems and their underlying behaviors, and the papers reported in this volume bear witness to this.

Techniques, technologies and trends: the study of data complexity presents the potential to provide analyses with increased frequency and timeliness, accuracy and objectivity, and to define sustainable models. Traditional quantitative methods for capturing socioeconomic data have often shown limitations in their ability to examine underlying systems, and with the three 'T' just mentioned, the outlines of future developments are starting to emerge.

The volume reports 127 contributions in the following areas:

- Advanced statistical methods for pattern recognition
- Advances in statistical learning from high-dimensional data
- Data analysis for web sources
- Distance and depth-based statistical learning methods for robust data analysis

- Economics and environment
- Education and labour
- Inequalities in the labour market
- Innovations and challenges in official statistics
- Labour market: trends, perspectives and new challenges
- Methodological and applicative contributions for evaluating sustainable development
- Methodological developments and applications for the assessment of student competencies
- Networks data analysis: new perspectives and applications
- New advanced statistical methods for data science
- Recent advances in statistical learning and data analysis
- Statistical analysis and modeling of environmental pollution data
- Statistical methods and complexity for evaluation in finance
- Statistical methods and composite indicators for healthcare
- Statistical methods and models for land monitoring with spatio-temporal data
- Statistical methods for environmental monitoring and sustainability
- Statistical methods for the analysis of university student choices and academic performance
- Statistical methods for the assessment of transport services and sustainable emissions
- Statistical methods for education and educational services
- Statistics in sports
- Tourism and territory.

The Conference event attracted many contributions as well as numerous Authors, not just from Italy but also from abroad. Over the three-day meeting, the Community has the opportunity to witness some of the state-of-the arts, new trajectories, and methodological challenges in 24 solicited sessions, 7 sessions of free contributes, two round tables - organized by Maurizio Vichi and Matilde Bini respectively - and three keynotes sessions with Ron S. Kennet of Samuel Neaman Institute of Israel, Luigi D'Ambra of Federico II University of Naples, and the former Minister Enrico Giovannini from University of Tor Vergata.

Organizers

Paolo Mariani
Chair of the Program Committee
Tonio Di Battista
Chair of Local Organizing Committee
Maurizio Carpita
Coordinator of the SVQS Group

Contents

Plenary Session **1**
Befitting Cross Validation with Three Case Studies 2
Kenett R.S.

Solicited Session SS1 - *Statistical analysis and modeling of environmental pollution data* **3**
Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution 4
Bottazzi Schenone M., Grimaccia E. and Vichi M.
Evaluating the nonlinear association between PM_{10} and emergency department visits 10
Bucci A., Sanmarchi F., Santi L., Giostra F., Tubertini E., Rosa S., Nante N. and Golinelli D.
Estimating spatially varying Gaussian Graphical Models to unveil relationships among pollutants in the Red River Delta in Vietnam 16
Pronello N., Cucco A., Ignaccolo R. and Ippoliti L.

Solicited Session SS2 - *Statistics in sports* **22**
Clustering Athlete Performances in Track and Field Sports 23
Argiento R., Colombi A., Modotti L. and Montagna S.
A Cross-Country Analysis of Engagement in Physical Activity and Sport Practice Learnt from Eurobarometer Survey Data 29
Simone R.
Strong eras in male professional tennis 35
Breznik K., Candila V., Milekhina A. and Restaino M.
NonParametric Combination method for data analytics in basketball matches 41
Barzizza E., Biasetton N., Ceccato R., Disegna M. and Vezzosi G.

Solicited Session SS3 - *Statistical methods for the analysis of university student choices and academic performance* **47**
The influence of labor market conditions on students' career disruption: first insights from Italy 48
Usala C., Sulis I. and Porcu M.
Socio-economic aspects that may affect South-North students' mobility in Italy 53
Genova V.G. and Boscaino G.
An analysis of student's performance in bachelor's degree 59

La Rocca M., Niglio M. and Restaino M.
An exploratory strategy for analyzing students' mobility data 65
Primerano I. and Giordano G.

Solicited Session SS4 - *Statistics in football* **71**
Community Detection in Sport Market Networks: The Case of Italian Professional Football 72
Rondinelli R. and Ievoli R.
An Original Application to Football of PLS-SEM for the xG Model 78
Cefis M. and Carpita M.
Performance Assessment of Football Players and Combined Permutation Tests with application to Home-Field Advantage 84
Bonnini S.
A First Proposal of the Triad Census for Weighted Networks: an Application to Football 90
Rondinelli R. and Palazzo L.

Solicited Session SS5 - *Advances in statistical learning from high-dimensional data* **96**
PCA approaches for vector functional time series 97
Aguilera A.M., Alonso F.J. and Acal C.
Conformal Prediction for Functional Kriging Models 102
Diana A., Romano E. and Adzic J.
Measuring Public-Private Connectedness in Financial Markets 108
Sánchez García J. and Cruz Rambaud S.
An original approach to anomalies in intertemporal choices through Functional Data Analysis: Theory and application for the study of Hikikomori syndrome 112
Martino R., Ventre V., Cruz Rambaud S. and Maturo F.

Solicited Session SS6 - *Labour market: trends, perspectives and new challenges* **118**
Enriching Job Vacancy Official Information with Online Job Advertisements: Chances and Limits 119
Lucarelli A. and Righi A.
Innovation in Management: towards the Open Manager 126
Bruttini P., Gallo M., Mariani P. and Menini T.

Solicited Session SS7 - *Data analysis for web sources* **132**
Enhancing SMEs default prediction with web-scraped data 133
Crosato L., Domenech J. and Liberati C.
Web data as enabler for informed decisions in Labour Market 137
Maggioni G.
The metaverse & luxury fashion brands: strategic communication exercise 141
Forciniti A. and Zavarrone E.
Increasing the Geographical Granularity of Economic Indicators with Google Trends 147

Domenech J. and Marletta A.

Solicited Session SS8 - *Methodological and applicative contributions for evaluating sustainable development* **153**

Evaluating sustainable development in EU countries through synthetic indicators . . . 154
Alaimo L.S. and Cucci M.

Naples and tourism sustainability: A survey of citizens' perceptions 160
Aria M., Pagliara F., D'Aniello L. and Della Corte V.

Modelling inequalities for sustainable development in Italy countries 166
Musella M., Borrata G., Camminatiello I. and Lombardo R.

Food Security and Sustainability: A Science Mapping Analysis 172
Piscitelli A.

Solicited Session SS9 - *Inequalities in the labour market* **178**

Skill similarities across Italian regions: an analysis based on the online job advertisements . . . 179
Kahlawi A., Buzzigoli L., Grassini L., Martelli C. and Giambona F.

Italian Labour Market reform and gender inequalities 185
Marini C. and Nicolardi V.

Intergenerational transmission of disadvantages in the Italian labour market: evidence from AD-SILC data . . . 191
Busetta A., Fabrizi E., Ragozini G. and Sulis I.

Solicited Session SS10 - *Statistical methods and complexity for evaluation in finance* **197**

Financial networks resilience and shocks propagation 198
Cerqueti R., Cinelli M., Ferraro G. and Iovanella A.

How the choice of one parameter impacts the numerical stability of the efficient frontier . . . 204
Fassino C. and Uberti P.

Dynamic shrinkage for minimum variance combination of forecasts 210
Mattera R.

Exploring the perception of the gender issue of Italian female entrepreneurs 217
Castellano R., Riccioni J. and Rinaldi A.

Solicited Session SS11 - *Networks data analysis: new perspectives and applications* **223**

Describing Italian mobility trajectories in higher education 224
Genova V.G., Giordano G., Ragozini G. and Vitale M.P.

Collaboration networks: methodological issues and updated empirical evidence on Italian statisticians . . . 230
De Stefano D., Fabbrucci Barbagli A.G., Santelli F. and Zaccarin S.

Mapping Ashtma Complexity with Graph Theory: an Integrative Approach 236
Cucco A., Simpson A., Murray C., Fontanella S. and Custovic A.

Investigating the patterns of Italian internal mobility: a network analysis at provincial level . . . 242

Sarra A., D'Ingiullo D., Evangelista A., Nissi E., Quaglione D. and Di Battista T.

Solicited Session SS12 - *Innovations and challenges in official statistics* **248**

Formal and informal networks of care for the elderly: regional profiles compared . . . 249
Sicuro L., Tucci D. and Coniglio R.

Gender Gap: a multidimensional approach . . . 255
Acampora C., Fusco D., Liguori M.A. and Pagliuca M.M.

Using Whatsapp in Official Statistics: a New tool for managing the Agriculture Census . . . 259
Fabi C.

Solicited Session SS13 - *Statistical methods and composite indicators for healthcare* **265**

Longitudinal composite indicators to measure the quality of health services 266
Crocetta C., Antonucci L., Cataldo R. and Mazza R.

Past and Future of Doctor-Patient Communication . . . 272
Tedesco N., Zavarrone E. and Forciniti A.

Network Analysis approach to customer satisfaction and service quality detection: an application to health-care services . . . 277
Crocetta C., Grassia M.G., Marino M., Mazza R., Simonacci V. and Stavolo A.

A project evaluation study on multiset Likert scale data . . . 283
Simonacci V., Marino M., Grassia M.G. and Gallo M.

Solicited Session SS14 - *Distance and depth-based statistical learning methods for robust data analysis* **289**

Robust distance-based predictive models . . . 290
Boj E., Grané A. and Parron D.

Data depth for mixed-type data through multidimensional scaling. An application to biological age imputation . . . 294
Cascos I., Grané A. and Qian J.

A compared protocol to improve clustering procedures . . . 298
Grané A., Riani M. and Salini S.

Robust diagnostics for Linear Mixed Models with the Forward Search . . . 304
Corbellini A., Grossi L. and Laurini F.

Solicited Session SS15 - *Advanced statistical methods for pattern recognition* **310**

Unsupervised classification of NPLs recovery curves . . . 311
Carleo A. and Rocci R.

Living alone in Italian municipalities . . . 317
Vellucci P., Benassi F., Naccarato A. and Gallo G.

Supervised learning from high-dimensional data through dynamic updating of functional classification rules . . . 323

Maturo F., Fortuna F. and Di Battista T.
Assessing the effectiveness of coordination among public authorities in cohesion expenditure . . . 329
Coco G., Monturano G. and Resce G.

Solicited Session SS16 - *Recent advances in statistical learning and data analysis* **335**
A Predictive Functional Principal Component Analysis of Well-Being Data 336
Marcis L., Pagliarella M.C. and Salvatore R.
Detecting the partition in the extended hierarchy of a dendrogram: an application on biomedical data . . . 343
Policastro V., Palazzo L. and Vistocco D.
Concordance measure for rankings . . . 350
Bissiri P.G. and Nai Ruscone, M.
Quadratic discriminant scoring for selecting clustering solutions . . . 355
Coraggio L. and Coretto P.

Solicited Session SS17 - *Statistical methods for education and educational services* **361**
Association between INVALSI scores and students' mobility in Italy: a preliminary analysis . . . 362
Bacci S., Bertaccini B., Lombardi G. and Tocchioni V.
Modelling Responses and Response Times: an application to Mathematics INVALSI data . . . 368
Bungaro L., Desiderati R. and Mignani S.
Latent potential outcomes: An analysis of the effects of programs aimed at improving students' non-cognitive skills . . . 374
Pennoni F., Bartolucci F. and Vittadini G.
Cognitive Skills and Non Cognitive Skills to Analyze School and Students Performances . . . 380
Vittadini G.

Solicited Session SS18 - *Statistical methods for the assessment of transport services and sustainable emissions* **386**
Sustainability assessment of urban transport by an LCA comparison on different technologies vehicles . . . 387
Della Ragione L. and Meccariello G.
Passenger comfort prediction via time-series classification . . . 393
Vanacore A., Pellegrino M.S. and Ciardiello A.
A statistical model to analyse driving behavior: a case study . . . 399
Rodia G., Sarnacchiaro P. and Acciarino V.
Aggregating judgments in non negotiable group decisions in transport system 403
Amenta P. and Lucadamo A.

Solicited Session SS19 - *New advanced statistical methods for data science* **408**

A unified framework for two-dimensional clustering on preference-approvals: an analysis of Eurobarometer data 409
Albano A., Sciandra M. and Plaia A.
Pandemic Data Quality Modelling: A Bayesian Approach 415
Ferrari L., Manzi G., Micheletti A., Nicolussi F. and Salini S.
Explainable AI for Peer-to-Peer Credit Risk Management 421
Babaei G., Pagnottoni P. and Do T. T.
Tackling misclassification in surveys about undeclared work via the EM algorithm 427
Arezzo M.F., Guagnano G. and Vitale D.

Solicited Session SS20 - *Tourism, territory and data analysis* **432**
Tourism, sustainability, and territorial impact: an input-output analysis . . 433
Garau G. and El Meligi A.K.
The Impact of Big Data in Tourism 439
Ciuffreda R., Choedon C. and Simonetti B.
Neural network-based prediction of domestic tourists' length of stay in Italy 443
Antolini F. and Cesarini S.
The management of cultural heritage in contexts of undertourism: a model for assessing the economic sustainability of public-private partnerships 450
Calabrò F.

Solicited Session SS21 - *Statistical methods and models for land monitoring with spatio-temporal data* **457**
Geo-referenced data and complex networks for measuring road accident risk 458
Cantaluppi G., Clemente C., Della Corte F. and Zappa D.
A comparison of geospatial models for car crash risk 464
Cantaluppi G., Giardino G. and Zappa D.
Geostatistical modelling of livestock-related $PM_{2.5}$ pollution and scenario analysis for policymakers - Work in progress 470
Fassò A., Rodeschini J., Fusta Moro A. and Finazzi F.
Functional clustering methods for space-time big data from mobile phone networks 476
Perazzini S., Metulini R. and Carpita M.

Solicited Session SS22 - *Methodological developments and applications for the assessment of student competencies* **483**
Modeling the main drivers of mathematical literacy of school-leaving students. Some evidence from the Invalsi tests 484
Davino C., Palumbo F., Romano R. and Vistocco D.
Educational Data Mining: clustering students' performance over time . . . 490
Taraborrelli G. and Farnè M.
The nexus of cultural capital with participation in early childhood education 496
Ripamonti E.
High- and Low-Performing students and future career: a gender and social issue 502
Falzetti P. and Ricci R.

Solicited Session SS23 - *Statistical methods for environmental monitoring and sustainability* **508**
Clustering spatial data through optimal transports 509
Balzanella A. and Verde R.
New interpretative insights for environmental air quality by means of FDA 514
Terzi S., Naccarato A. and Fortuna F.
A Bayesian State-Space Model to Mitigate Unmeasured Confounding . . . 520
Zaccardi C., Valentini P. and Ippoliti L.
Mining social media data for damage assessment in environmental disasters 526
del Gobbo E., Cafarelli B., Ippoliti L. and Fontanella L.

Solicited Session SS24 - *Satisfaction and behavior in tourism* **531**
The evaluation of the hotel stay through a new development of correspondence analysis coping with ordinal variables 532
D'Ambra A. and Amenta P.
Assessing the role of knowledge and authenticity in the formation of attendee loyalty at cultural festivals 536
Rivetti F., Lucadamo A. and Rossi C.
Residents' Opinions and Perceptions of Tourism Development in the Historic City of Matera 542
Sarnacchiaro P., Di Gennaro R. and Di Taranto E.
Exploring tourism at religious sites: The case of Assisi 547
Rivetti F., Dini M. and Splendiani S.

Session of free contributes SFC1 - *Education and labour* **552**
Local concordance among the items of questionnaires on student's opinion (OPIS) 553
Terzi S. and Petrarca F.
High School Proficiency of Future University Students: An Analysis based on INVALSI Data 560
Santelli F., Di Credico G. and Di Caterina C.
Employment vulnerability of immigrants in the labour market – Does origin matter? 566
Bittaye M.
The effect of pricing policies on students' use of university canteens 572
Masserini L., Bini M. and Lorenzoni V.
Gig workers' identikit 576
Zavarrone E. and Forciniti A.

Session of free contributes SFC2 - *Tourism and territory* **582**
Multi-source approach for statistics in tourism sector 583
Bianchino A., Fusco D., Giordano P., Liguori M.A. and Summa D.
Statistical analysis of tourism sustainability in Campania: post Covid-19 review 589
Giacalone M., Basile V. and Bellucci M.
Investigating recent changes in dietary behavior 596
D'Uggento A., d'Ovidio F.D., Toma E. and Onorati M.G.

Depopulation in the Abruzzo municipalities 603
Carulli A.L., Di Spalatro D. and Valentini A.
The Productions System of Inland Areas 607
Madia Carucci A.M. and Regano A.

Session of free contributes SFC3 - *Social issues* **613**
Modelling the gender gap in youth mortality with an Age-Period-Cohort analysis 614
Lanfiuti Baldi G. and Nigri A.
Random forest for classifying odor emission sources 620
Distefano V., Palma M., De Iaco S. and Mazuruse G.
An Experimental Annotation Task Investigating Annotator Agreement Within a Misogynistic Dictionary and Corpus 626
Tontodimamma A., Ignazzi E., Anzani S., Fontanella L. and Di Zio S.
Statistical analysis of COVID19 impact on Italian mortality 632
Franchetti G. and Politano M.
Measuring multidimensional deprivation using objective and subjective data: an application of the Voronoi ranking method 638
Ciommi M., Mariani F., Polinesi G. and Recchioni M.C.

Session of free contributes SFC4 - *Trends* **644**
The role of big data analytics in circular supply chains: A bibliometric analysis 645
Khan F. and Rapposelli A.
Estimation of the ranking of incentive policies for the adoption of 4.0 technologies 652
Bonnini S. and Borghesi M.
Risk Management and Future Scenarios. A proposal based on a mixed-method approach 658
Di Zio S., Bolzan M., Marozzi M. and Scioni M.
Explainable artificial intelligence (XAI) through artificial intelligence from a human in the loop (HITL) perspective: an interview with ChatGPT 664
Santarcangelo V., Lamacchia A., Vitullo S., Di Lecce M. and Giacalone M.
Relevance in official statistics: information needs, satisfaction with data quality, some results and future perspectives 670
Loporcaro M.F.

Session of free contributes SFC5 - *Economic issues* **676**
Evolutionary trends of start-ups in Italy: a case study 677
Duttilo P., Caruso G., Iannone B. and Gattone S.A.
Permanent establishments and efficiency analysis with global enterprises . 682
Frenda A. and Sepe E.
Techniques and constructs in some recent market and organizational research 685
Sciascia I.A.
The value of buildings in the Italian general government balance sheet . . . 689
Santoro P. and Regano A.

Linear and nonlinear factors affecting default risk in the peer-to-peer lending market . . . 695
Giordano F., Milito S. and Parrella M.L.

Session of free contributes SFC6 - *Methodological issues* **701**
Macroeconomic Time Series Classification by Nonparametric Trend Estimation . . . 702
Feo G., Giordano F., Niglio M. and Parrella M.L.
A Normalization Method for Space-time Analysis of Evaluation and Quality Indicators . . . 709
Mazziotta M. and Pareto A.
Unveiling Latent Structures: exploring Multidimensional IRT Models using Dirichlet Process Mixtures . . . 713
Valentini P., Fontanella S. and Fontanella L.
On a technique to detect accounting data manipulation . . . 718
Passamonti C.

Session of free contributes SFC7 - *Economics and environment* **724**
Determinants of Water Conservation Behaviour and Spatial Heterogeneity in their Coefficients . . . 725
Mammadli R. and Gigliarano C.
Modeling the economic burden of grass pollen allergoid immunotherapy . . 731
Bilancia M. and Di Bona D.
Swine fewer in Liguria: who does pay for economic losses? A causal analysis 735
Baggetta C., Cavalletti B. and Corsi M.
Measuring Multidimensional Poverty of the Italian Regions in the era of COVID-19 . . . 741
Chelli F.M.C., Ciommi M., Gigliarano C. and Polinesi G.
Drivers of inflation: relationships changing over time . . . 747
Sokolenko O., Palumbo A., Fortuna F., Naccarato A. and Marie J.
Environmental accounting and sustainable cities: an explorative bibliometric–based literature analysis . . . 753
Kaur A., Javed A., D'Andreamatteo A. and Rapposelli A.

Solicited Session SS16 - *Recent advances in statistical learning and data analysis*

Session of the SIS-CLADAG organized by Domenico Vistocco and Pietro Coretto
Chair: Marcella Niglio
Discussant: Cristina Davino

1. *A Predictive Functional Principal Component Analysis of Well-Being Data* (Marcis L., Pagliarella M.C. and Salvatore R.)

2. *Detecting the partition in the extended hierarchy of a dendrogram: an application on biomedical data* (Policastro V., Palazzo L. and Vistocco D.)

3. *Concordance measure for rankings* (Bissiri P.G. and Nai Ruscone, M.)

4. *Quadratic discriminant scoring for selecting clustering solutions* (Coraggio L. and Coretto P.)

Quadratic discriminant scoring for selecting clustering solutions

Funzione discriminante quadratica per la selezione di soluzioni di clustering

Luca Coraggio and Pietro Coretto

Abstract Selecting an optimal clustering solutions is a difficult problem. There exist many data-driven validation strategies in the literature to perform this task. In this paper, we focus on a recent proposal, based on quadratic discriminant scores and bootstrap resampling, namely the BQH and BQS from Coraggio and Coretto [4]. These strategies proved to be extremely successful with elliptic-symmetric clusters and, in general, when clusters can be separated by quadratic boundaries. In this work, we review the BQH and BQS strategies, and try to shed more light on their functioning, by comparing them with alternative likelihood-based validation indexes, and with different resampling schemes.

Abstract *La selezione di soluzioni di clustering ottimali è un problema complesso. In letteratura, esistono molte strategie di validazione per svolgere questo compito. In questo lavoro, ci concentriamo su una proposta recente, basata su funzioni di discriminante quadratica e tecniche di ricampionamento bootstrap, ovvero i metodi BQH e BQS proposti in Coraggio and Coretto [4]. Queste strategie si sono dimostrate estremamente efficaci con cluster ellittico-simmetrici e, più in generale, quando i cluster possono essere separati da confini quadratici. In questo lavoro, riprendiamo le strategie BQH e BQS, e indaghiamo il loro funzionamento in uno studio comparativo, utilizzando nuovi indici di validazione, basati sulla funzione di verosimiglianza, assieme a schemi di ricampionamento alternativi.*

Key words: Cluster validation, Mixture models, Model-based clustering, Resampling methods

Luca Coraggio
Department of Economics and Statistics, University of Naples Federico II, e-mail: luca.coraggio@unina.it

Pietro Coretto
Department of Economics and Statistics, University of Salerno e-mail: pcoretto@unisa.it

1 Introduction

Typically, in cluster analysis, the researcher produces many solutions, running several clustering algorithms with various settings. The problem is that, while a single final solution may be required, it may well be the case that multiple of them provide a good description of the data, according to different clusters' concepts [17]. Recently, in [4], we proposed a novel validation index aimed at selecting clustering solutions in cases where clusters can be expected to have elliptic-symmetric shapes, or to be separable by quadratic boundaries. The proposed strategy proved to be effective in conjunction with a bootstrap resampling strategy that hedges against overoptimism in the selection process, and helps avoid choosing overly-complex solutions. In this paper, we review this methodology and investigate the different components that make it successful. We do this by comparing it with an alternative likelihood-based validation index, and combining it with different resampling schemes. The outline of the paper is as follows. Section 2 reviews the BQH and BQS indexes, and introduces the alternative strategies; Section 3 presents the empirical analysis; Section 4 concludes.

2 Quadratic Discriminant Scoring

Let $\mathbb{X}_n$ be an observed sample of size n, with feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$. Let $\mathscr{G}^{(m)} = \left\{G_k^{(m)},\ k = 1,\dots,K_m\right\}$ be a clustering solution, allocating the n objects into K_m groups; $\mathscr{G}^{(m)}$ is obtained from a clustering method $m \in \mathscr{M}$.We assume that $\mathscr{G}^{(m)}$ can be meaningfully described by a collection of K_m triplets $\boldsymbol{\theta}^{(m)} = \left\{\boldsymbol{\theta}_k^{(m)},\ k = 1,\dots,K_m\right\}$, where each $\boldsymbol{\theta}_k^{(m)}$ collects unique elements of the following objects: *(i)* π_k: the expected fraction of points belonging to the *k-th* group; *(ii)* $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the *k-th* cluster's center; *(iii)* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is a positive definite scatter matrix that either coincides with or is proportional to the *k-th* cluster's covariance matrix. $\boldsymbol{\theta}^{(m)}$, together with K_m, provides a description of the *m-th* clustering configuration $m \in \mathscr{M}$.[1]

For a point $\boldsymbol{x}$ and a triplet $\boldsymbol{\theta}_k$, we define the quadratic scoring of point $\boldsymbol{x}$ for the *k-th* cluster as

$$\mathrm{qs}(x, \boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) \tag{1}$$

The quadratic score is inspired to Quadratic Discriminant Analysis, where (1) represents the optimal classification boundaries under the Gaussian assumption (e.g., see [10]). It can be seen as a measure of how well point $\boldsymbol{x}$ is accommodated into

[1] Depending on m, the triplets $\boldsymbol{\theta}_k^{(m)}$ may be taken to coincide with elements defined by the approach itself (e.g., in model-based clustering, each triplet coincides with the parameters of one mixture component; e.g., see [13]), or can be estimated with sample quantities, computing within-cluster empirical counterparts (using estimated point-to-cluster assignments).

cluster k (higher values correspond to a better fit). Using (1) we define the quadratic partition, $\mathscr{Q}$, as

$$\mathscr{Q} = \{Q_k,\ k \in \{1,\ldots,K\}\}, \quad Q_k = \left\{ \boldsymbol{x} \in \mathbb{R}^p : k = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \mathrm{qs}(\boldsymbol{x}, \boldsymbol{\theta}_k) \right\}. \tag{2}$$

Finally, for a given $\boldsymbol{\theta}$ we define the hard and smooth quadratic scoring criteria respectively as

$$H(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} s_H(\boldsymbol{x}, \boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} \mathbb{I}\left\{ \boldsymbol{x} \in Q_k(\boldsymbol{\theta}^{(m)}) \right\} \mathrm{qs}\left(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)} \right); \tag{3}$$

$$T(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} s_T(\boldsymbol{x}, \boldsymbol{\theta}^{(m)}) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \sum_{k=1}^{K^{(m)}} \tau_k \left(\boldsymbol{x}_i, \boldsymbol{\theta}^{(m)} \right) \mathrm{qs}\left(\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(m)} \right), \tag{4}$$

where τ_k defines a smooth weight of point-to-cluster membership as measured by the quadratic score:

$$\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\mathrm{qs}(\boldsymbol{x}, \theta_k)}{\sum_{k=1}^{K} \mathrm{qs}(\boldsymbol{x}, \theta_k)}.$$

2.1 Likelihood-based validation

The quadratic score (1) is strongly connected to likelihood theory, and it is easy to show that it is proportional to the Gaussian density function [4]. Thus, a natural alternative to the scoring criteria (3) and (4) appears to be the likelihood function of a Gaussian mixture model. A similar proposal was also made by Smyth [16], used in combination with cross-validation. Thus, the data log-likelihood can be used to score a clustering solution (solutions achieving higher likelihood are preferred), and is defined as:

$$l(\boldsymbol{\theta}^{(m)};\ \mathbb{X}_n) = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbb{X}_n} \log \left(\sum_{k=1}^{K^{(m)}} \pi_k^{(m)} \phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)}) \right), \tag{5}$$

where $\phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})$ is the density function of a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. Note that, differently from (3) and (4), the usage of (5) is only justified for model-based clustering, where a probabilistic model is assumed, and it is needed to motivate the construction of the likelihood function.

Algorithm 1 Bootstrap quadratic scoring

input: observed sample $\mathbb{X}_n$ (with ecdf $\mathbb{F}_n$), $\alpha \in (0,1)$; clustering method $m \in \mathscr{M}$; integers $B > 0$
output: bootstrap quadratic scoring for method m: $\widetilde{L}_n^{(m)}$.
(to ease notation, dependence on m is dropped and reintroduced in step 3)
for $b \in \{1, \ldots, B\}$ do
(step 1.1) $\mathbb{X}_n^{(b)} \leftarrow$ non-parametric bootstrap resample from $\mathbb{X}_n$ (sample of size n from $\mathbb{F}_n$)
(step 1.2) $\hat{\boldsymbol{\theta}}_n^{(b)} \leftarrow$ triplets of parameters from clustering solution m fitted on $\mathbb{X}_n^{(b)}$
(step 1.3) $S_n^{(b)} \leftarrow$ score solution on points in $\mathbb{X}_n$ not in $\mathbb{X}_n^{(b)}$

$$S_n^{(b)} = H(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n) \quad \text{or} \quad S_n^{(b)} = T(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n) \quad \text{or} \quad S_n^{(b)} = l(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n)$$

end for
(step 2) $\widetilde{W}_n \leftarrow \frac{1}{B}\sum_{b=1}^{B} S_n^{(b)}$ $\quad R_n^{(b)} \leftarrow \sqrt{n}\left(S_n^{(b)} - \widetilde{W}_n\right)$
(step 3) Compute $(\alpha/2)$-level and $(1-\alpha/2)$-level empirical quantiles:

$$\widetilde{L}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \leq t\right\} \geq \frac{\alpha}{2} \right\}; \qquad \widetilde{U}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \leq t\right\} \geq 1 - \frac{\alpha}{2} \right\}$$

2.2 Resampling schemes

Choosing the solution that maximizes (3) or (4) may give poor results: since the sample data $\mathbb{X}_n$ is used both for estimation and scoring, overly-complex solutions may be selected due to overoptimism in the evaluation process. It is well known that increasingly complex solutions can fit the data better, and many selection methods takes this into account in several ways. For example, information criteria typically include a penalization term for the model's complexity (e.g., BIC [15] and ICL [3]). To cope with this issue, in Coraggio and Coretto [4], we propose to estimate clustering solutions on non-parametric bootstrap resamples [7] from $\mathbb{X}_n$, and to evaluate (3) and (4) on $\mathbb{X}_n$. The procedure is reviewed in Algorithm 1, and here we extend it to include a likelihood-based scoring, using (5) as the scoring function.

3 Empirical analysis

The experimental analysis is a scaled-down version of that in Coraggio and Coretto [4], using the Iris, Olive and Banknote data sets [1, 9, 8]. Since we are going to compare the quadratic scores with likelihood-based ones, $\mathscr{M}$ includes only model-based clustering methods: *(i)* 140 model-based clustering methods, using Gaussian mixture models with parsimonious representation of the covariance matrices [2], implemented with the Mclust software [14] (setting $K = 1, \ldots, 10$, and 14 covariance models); *(ii)* 180 model-based clustering methods, using Gaussian mixture models with eigen-ratio contraints (ERC; [12]), implemented with Otrimle software

Table 1: Selected clustering solution by selection criteria (left-most column). Each sub-table shows results from a real data set: the first column shows the selected solution, and the second column reports its ARI, computed against true classes.

	(b) Iris		(c) Olive		(d) Banknote	
Criterion	Selected m	ARI	Selected m	ARI[1]	Selected m	ARI
QH	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
QS	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
LK	Otrimle, K=10, γ=10^4	0.44	Mclust, K=10, VVV	0.54	Otrimle, K=10, γ=10^4	0.26
CVQH	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=6, VVV	0.79	Mclust, K=4, VVE	0.68
CVQS	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=6, VVV	0.79	Otrimle, K=3, γ=10	**0.86**
CVLK	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=9, VEE	0.65	Mclust, K=4, VVE	0.68
BQH	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Otrimle, K=3, γ=10	**0.86**
BQS	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Otrimle, K=3, γ=10	**0.86**
BLK	Otrimle, K=3, γ=10^2	**0.90**	Mclust, K=8, VVV	**0.86**	Mclust, K=6, EEE	0.47

(Olive) 1. refers to the finer 9-regions classification.

[5, 6] (setting $K \in \{1,\dots,10\}$, ERC $\gamma \in \{1,5,10,10^2,10^3,10^4\}$, and 3 initialization methods). The criteria compared to select optimal solutions are respectively based on equations (3), (4), and (5), and are divided as follows. QH, QS, and LK: clustering solutions are estimated and scored using the full data, $\mathbb{X}_n$; CVQH, CVQS, CVLK: clustering solutions are estimated on a "train set" and scored on a non-overlapping "test set", using a 10-fold cross-validation scheme, as in [16]. BQH, BQS, BLK: clustering solution are estimated and scored according to Algorithm 1, selecting the method m maximizing $\widetilde{L}_n^{(m)}$. For each criterion, the selected solutions are evaluated against the true class labels, reporting the achieved Adjusted Rand Index (ARI, [11]).

Results are presented in Table 1. The comparison gives a better understanding on the mechanism that lies behind the effectiveness of the BQH and BQS criteria. First, notice that all criteria where solutions are estimated and scored on the full data (QH, QS, LK) always select poor, overly-complex solutions: this is because of the problem discussed in Subsection 2.2. On the contrary, both resampling mechanisms choose lower complexity solutions, acting similar to complexity penalization. However, the cross-validation scheme produces poor (Olive) or inconsistent (Banknote) results, with respect to the bootstrap scheme. This is likely due to the fact that 10-fold cross-validation splits leave too few shared information between train and test sets, and is less adequate for clustering where, differently from prediction settings, the goal is to select a clustering method that has a good in-sample performance. This is also true for the likelihood-based criteria, CVLK and BLK, which perform worse with respect to the quadratic score criteria. Indeed, it can be shown that compared to the likelihood function, the quadratic scores add extra penalization for overlapping clusters. Overall, both the quadratic scores, (3) and (4), and the resampling scheme in Algorithm 1 seem equally important to consistently achieve good results.

4 Conclusion

In this paper, we reviewed the BQH and BQS procedures from [4]. We ran an empirical analysis comparing it with a likelihood-based validation index, and testing different resampling schemes. Our experiments show that neither the bootstrap resampling scheme in Algorithm 1 nor the quadratic discriminant score in (1), alone, are sufficient to obtain good results. Rather, both of them contribute to the procedure in different ways. The comparison with likelihood-based alternatives highlights that the extra penalty that the quadratic scores impose on overlapping clusters is needed to select better solutions in cases where clusters are not well separated. Similarly, the comparison with the 10-fold cross-validation shows that the bootstrap resampling scheme is better suited for cluster analysis, where the final goal is to describe the data at hand, rather than predicting cluster membership for unseen points.

References

1. Anderson, E.: The species problem in Iris. Ann. Missouri Bot. Gard. 23 (3) 471–483 (1936)
2. Banfield, J. D., Raftery, A. E.: Model-Based Gaussian and Non-Gaussian Clustering. Biometrics, 49 (3) (1993)
3. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 22 (7), 719–725 (2000)
4. Coraggio, L., Coretto, P.: Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. J. Multivariate Anal. , 196, 105181, (2023)
5. Coretto, P., Hennig, C.: Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. J. Mach. Learn. Res. 18 (142), 1–39 (2017)
6. Coretto, P., Hennig, C.: OTRIMLE: Robust model-based clustering. R package version 2.0 (2021)
7. Efron, B.: Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7 (1) (1979)
8. B. Flury and H. Riedwyl: Multivariate statistics: a practical approach., 1st ed. Chapman and Hall/CRC, London—New York (1988)
9. Forina, M., Armanino, C., Lanteri, S., Tiscornia, E.: Classification of olive oils from their fatty acid composition. Food Res. Data Anal. 189–214 (1983)
10. Hastie, T., Tibshirani, R. J., Friedman, J.: The Elements of Statistical Learning. 2nd ed., Springer New York (2009)
11. Hubert, L., Arabie, P.: Comparing partitions., J. Classification 2 (1), 193–218 (1985)
12. Ingrassia, S.: A likelihood-based constrained algorithm for multivariate normal mixture models. Stat Methods Appt 13 (2) (2004)
13. McLachlan, G., Peel, D.: Finite mixture models. In: Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, Inc., New York (2000)
14. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R J. 8 (1), 205–233 (2016)
15. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. 6 (2) 461–464 (1978)
16. Smyth, P.: Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. 10 (1) 63–72 (2000)
17. von Luxburg, U., Williamson, R. C., Guyon I.: Clustering: Science or Art?. In Proceedings of ICML Workshop on Unsupervised and Transfer Learning, Bellevue, Washington, USA, 27, pp. 65–79 (2012)

This book collects the papers presented at the 11th Scientific Meeting of the SIS Group "Statistics for the Evaluation and Quality in Services" *Statistical Methods for Evaluation and Quality: Techniques, Technologies and Trends (T^3)*, which took place at the University of Chieti-Pescara, on 30th August-1st September, 2023.
The papers, which had been selected through a refereeing process, contain topics on statistical approaches and methodologies for the evaluation of public services in different contexts, and cover the areas of digital transition, e-commerce and digital marketing, enterprises, environment and territory, healthcare and wellness, finance, bank and FinTech, justice system, labour market, official statistics, public administration, food and wine, school, education and training, social, sports, sustainability, tourism, transport, university and research, well-being and welfare.

ISBN: 979-12-803-3369-8
DOI: 10.60984/978-88-94593-36-5-IES2023

ISBN 9791280333698