COUNTING IN NON-ITALIAN RESIDENTS: THE USE OF THE "PERMITS TO STAY" ARCHIVE IN THE NEXT POPULATION CENSUS

Marco Fortini[1], Gerardo Gallo[1], Federico Benassi[1], Luca Mancini[1], Luigi Marcone[1]

Keywords: population census, foreign population, record linkage, deduplication.

1. INTRODUCTION

There is evidence that a significant proportion of foreigners living in Italy has been systematically missed by general population censuses. According to the 2001 PES, foreigners made up about 25% of the total undercoverage in metropolitan areas (Fortini and Gallo, 2001; Gallo, 2010). Borrelli *et al.* (2011) find similar results looking at the 2009 Census Pilot Survey. The 2001 Italian Population Census will be officially assisted for the first time by municipal population registers (LACs). This transition from a traditional enumeration-based to a register-assisted population census could potentially exacerbate the undercounting of the resident foreign population. In fact poorly maintained population registers can induce a significant discrepancy between the registered population and the actual number of non-Italians living in Italy with a regular permit to stay[2]. In order to obviate these deficiencies auxiliary administrative registers will be used to supplement the LACs and reduce the undercount of the resident foreign population. The main auxiliary archive considered for non-Italian residents is based on the "Permits To Stay" databases (PS) which are kept by the Italian Ministry of Interior. The aim of this paper is to describe the criteria used by the National Institute of Statistics (ISTAT) to construct the PS archive for census purposes. The rest of the paper is organized as follows: Section 2 will illustrate the composition of the PS archive and present the deduplication strategy including the decision rules applied to duplicates in order to decide which record(s) to retain in the final auxiliary PS archive. Section 3 will conclude and outline the next research steps.

2. DATA AND DEDUPLICATION STRATEGY

The administrative data used in this paper refer to the foreign population with a valid resident permit as of November 2010. The data come from two sources: The De-

[1] Division for General Censuses, National Institute of Statistics, Rome, Italy. Email: benassi@istat.it. The authors are grateful to Francesco Borrelli and Alessandra Ronconi for their excellent research assistance on deduplication. We are also indebted to Luca Valentino and Monica Scannapieco for their assistance on Relais 2.2. and to Edoardo Patruno for his help on the geo-codification of dwelling addresses.

[2] To a lesser extent this is also true for Italian citizens: although LACs should provide at any time a precise snapshot of the resident population living within the municipal borders, coverage errors – either permanent residents not listed (undercounts) or people who have left the municipality or passed away but have not yet been written off (over counts)- are not uncommon.

partment of Public Security (DPS) and the Department for Civil Rights (DCR) which are both part of the Italian Ministry of Interior. Individual information is stored in different databases depending on the processing stage of each individual application. At the time of data acquisition, DPS' databases contained information respectively on about 185,000 newly filed applications still at the pre-processing state[3]; 2,600,000 regular permit holders; 500,000 applications still under scrutiny; 9,500 rejected applications; 755,000 expired permits; 39,000 resident minors registered on their parents' permit; 57,000 work permits issued to first-time entrants; and 235,000 permits granted for family reunion purposes. DCR's archives covered information on about 250,000 previously "illegal" migrants applying for 'graduation' to regular permit holders; 131,000 applications for family reunion and 122,000 work permit extension/renewal applications. The overlap between DPS's and DCR's archives it is not only the result of the duplication of recordkeeping between two departments of the same Ministry as in the case of family reunion permits/applications, but it also occurs because obtaining a resident permit is a lengthy process taking the applicant through a series of successive statuses[4]. Therefore de-duplication is an important preliminary step toward the formation of an auxiliary archive to support a register-based census. Whereas the Italian tax code (*codice fiscale*) is the key of choice to uniquely identify each Italian citizen, its superiority is less clear cut when it comes to identify foreign residents. Usually not all foreigners have been assigned a *codice fiscale* while cases of individuals with two different codes are not uncommon. Given these limitations the *codice fiscale* cannot be used as the only key to identify duplicate records across the PS databases. This paper has therefore taken a more comprehensive approach to deduplication based on a probabilistic record linkage model[5]. Table 1 presents the deduplication results for a sample of countries of origin. Column (1) shows the total number of records to deduplicate while column (2) contains the number of individual records from the matched pairs. Column (3) shows the individual records matched only on *codice fiscale*. A comparison between (2) and (3) is summarized in (4) and stands as an indicator of the added value of probabilistic multiple-key deduplication compared to a single-key deterministic approach. For China a value of 10.2 indicates that about 10% additional matched pairs have been found using a probabilistic record linkage. This is a net gain as it takes already into accounts the error (5) represented by the percentage of undetected matches on *codice fiscale* due to misspecification of matching model[6]. Finally, the last column (6) shows the number of du-

[3] This is a first screening phase where the completeness of the information provided is verified.

[4] For instance one applicant can emerge from a clandestine status (I), have her application under examination for a while (C) and then be granted a first-timer short-term work resident permit (G) before becoming eligible to a more permanent permit to stay (B). In other words the same hypothetical individual will be present at least 4 times in the consolidated archive.

[5] Deduplication can be envisaged as a linkage operation between records belonging to the same archive where - depending on the strategy adopted - each record is compared with a subset of (or possibly all) other records. We use neighborhood matching methods with sorting on individual names and blocking on nationality to reduce the search space. The estimation is based on the EM algorithm proposed by Fellegi and Sunter. The matching variables include year, month and day of birth, full name and *codice fiscale*. We impose equality on all matching keys with the exception of name for which a Levenshtein distance function was used. No 1:1 reduction is chosen for linkage results (a cluster solution is adopted). Matching thresholds can vary from country to country. The software used for the deduplication is Relais (REcord Linkage At IStat) 2.2.

[6] More precisely this error is due to inconsistencies in the name field.

plicate records organized in clusters rather than pairs. This rearrangement of duplicates draws upon graph theory: each record is a knot that can be associated directly or indirectly with one or more other knots by means of arches. As a result figures are lower than in (2) because some records typically appear in more than one matched pair.

Country	(1)	(2)	(3)	(4)	(5)	(6)
Albania	441,311	85,292	82,312	3.6	5.1	81,543
Bangladesh	111,948	39,572	37,770	4.8	5.9	37,712
China	255,772	60,182	54,600	10.2	5.0	57,413
Ecuador	87,298	23,118	20,672	11.8	5.2	22,004
India	162,480	75,178	63,936	17.6	8.1	68,510
Macedonia	70,314	15,890	13,724	15.8	4.7	15,147
Tunisia	107,824	31,502	25,972	21.3	4.9	24,818
Ukraine	269,966	118,954	109,892	8.2	5.5	113,663

Tab 1. Deduplication results for some countries

3.1. *Decision rules*

A set of decision rules is applied to each cluster of duplicate records in order to identify the record(s) to include in the auxiliary archive. The results are summarized in Table 3.

Country	(1)	(2)	(3)	(4)	(5)	(6)
Albania	81,543	41,263	201	399,847	2.3	3.5
Bangladesh	38,669	20,096	62	91,790	4.3	3.1
China	57,413	29,102	75	226,595	5.8	4.7
Ecuador	22,004	11,147	38	76,113	6.2	3.3
India	68,510	35,140	75	127,265	9.3	7.7
Macedonia	15,147	7,640	37	62,637	7.9	7.1
Tunisia	24,818	12,493	106	95,225	1.6	6.7
Ukraine	113,663	57,767	106	212,093	3.8	5.2

Tab 2. Decision rules

The chosen records are the "survivors" to the following exclusion rules applied in hierarchical order to each cluster: (a) all duplicate records are deleted with the exception of the one(s) with most recent filing date; (b) after sorting the residual duplicate records by the degree of reliability of the database of origin all non-first rank records are deleted provided their addresses coincide with the address of the top-rank record. Columns (2) and (3) show the number of records excluded after applying rule (a) and rule (b), respectively. Column (4) reports the number of surviving records for each country to be included in the auxiliary PS archive. Whenever either the *codice fiscale* of the surviving records was incomplete or their address could not be geo-referenced and geo-coded to a valid EA address, suitable donors from the same cluster were identified according to the criterion of the second most recent filing date. A valid *codice fiscale* and address were then retrieved and donated to the surviving record as supplementary fields. The last two columns of Table 2 show respectively the percentage of records with missing or

incomplete *codice fiscale* (5) and those with an address which could not be geo-referenced/coded (6).

4. CONCLUDING REMARKS AND FUTURE DEVELOPMENTS

Foreign residents traditionally account for a significant portion of the undercounts in Italian population censuses. At the eve of the 15th population census to take place next October ISTAT has been taking important steps to count in those foreigners living in the country on a permanent basis. A solution to the "foreigners' issue" is especially crucial in the forthcoming census round where the transition from a traditional enumeration-based to a register-assisted census could significantly increase the number of non-Italian residents who are entitled to receive a census questionnaire but will never receive one. ISTAT is planning to use auxiliary archives to guide a selective field search of households and individuals expected to be living within municipal borders but not yet enlisted in municipal population registers (LACs). The "Permits To Stay" archive is the primary instrument which will be used to fill LACs' information gaps on foreign residents. The paper has illustrated how the PS archive was constructed. After a careful process of data cleaning, deduplication, geo-coding and geo-referencing applied to about 4 million individual records collected by the Ministry of Interior into 11 different and partially overlapping databases the PS archive will contain approximately 3,300 thousands records. This means that about 700,000 duplicate records have been identified and dropped from the archive according to an agreed protocol of exclusion rules. The surviving records will be linked with the LACs at the municipal level. For each municipality, all those individual records found in PS at a valid address but not in LAC will be pooled together to form a supplementary municipal population register. These registers will then guide enumerators to reach individuals and households who will have otherwise been missed in the mail-out stage of the census. This information will also help us to better find the elusive component of foreign population.

References

Berger J. (1990) Robust Bayesian analysis: sensitivity to prior, *Journal of Statistical Planning and Inference*, 25, 303-328.

Borrelli, F., Fortini, M., Mancini, L., Marcone L. and A. Ronconi (2011) *Assessing the Effectiveness of Administrative Registers in Reducing Under-Coverage Errors in a Population Census: Evidence from the 2009 Italian Census Pilot Survey*, paper submitted to the 2011 SIS Conference.

Cooper M.C., Milligan G.W. (1988) The effect of measurement error on determining the number of clusters in cluster analysis, in: *Data, Expert Knowledge and Decision*, Gaul, W. & Shader, M. (Eds.), Springer, 319-328.

Duda R.O., Hart P.E. (1973) *Pattern Classification and Scene Analysis*, Wiley, New York.

Fortini, M. e G. Gallo (2009) *Misure di sottocopertura anagrafica in base alla revisione post-censuaria del 2001*, paper presented at the 2009 SIS Conference.

Gallo G. (2010) *The use of permits to stay to check the local population registers undercount*, Note presented to the Export Group Meeting on Register-Based Censuses, The Hague, 10-11.

Gu, L., Baxter, R., Vickers, D. e C. Rainsford (2003) *Record Linkage: Current Practice and Future Directions*, CMIS Technical report n. 03/83, Canberra, Australia.

Scannapieco, M., Tuoto, T., Valentino, L., Cibella, N. and M. Fortini (2010) *Relais User's Guide*, Version 2.1, DCMT and DCCG, Istat, Rome.