Distr.: General
27 March 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 3: Availability, completeness and quality of data from registers and other sources

The use of population registers in the 15th italian census: challenges and preliminary evidence

Note by ISTAT, Italy[1]

Summary

One of the major innovations introduced by the Italian National Institute of Statistics (ISTAT) in the last Population Census round held in October 2011 was the use of administrative registers to assist census operations. The purpose of the paper is twofold: a) to document the effort undertaken by ISTAT to make these registers amenable for census use and illustrate the main challenges encountered in terms of data cleaning, integration and standardization, and b) to present some preliminary results based on the census metadata available to date on the coverage of the municipal population registers as well as on the degree of success of the auxiliary registers to deal with deficiencies if the municipal population registers.

I. Introduction

1. One of the major innovations introduced by the Italian National Institute of Statistics (ISTAT) in the last Population Census round held in October 2011 was the use of

[1] Section II was prepared by Anna Pezone and Federica Pellizzaro (SCD/F), section III.1 by Daniela Casale (SCD/C), section III.2 by Luca Mancini (MTC/A) and section IV by Luigi Marcone (MTC/A).

.

administrative registers to assist census operations. Millions of households were no longer visited by enumerators but received their questionnaires by post at the address printed in the municipal population registers (Liste anagrafiche comunali - henceforth LAC) of the municipality where the household head resided.

2. Coverage errors due to births and deaths as well as changes of address occurred between the mailing-out date and the census date were tackled by instructing enumerators to deal with the variations during field operations.

3. By law LAC should be kept constantly up-to-date to provide at any time a precise snapshot of the resident population living within the municipal borders. However cases of households or individuals not registered in the municipality and/or at the address where they usually live are not uncommon.

4. In order to help locating these individuals and count them in auxiliary administrative archives – including among others the National Tax Register, the Permits to Stay Archive and the Addresses Register - were used.

5. The purpose of the paper is twofold: a) to document the effort undertaken by ISTAT to make these registers amenable for census use and illustrate the main challenges encountered in terms of data cleaning, integration and standardization, and b) to present some preliminary results based on the census metadata available to date on the coverage of the LAC as well as on the degree of success of the auxiliary registers to deal with LAC' deficiencies.

II. The *Liste Anagrafiche Comunali*

6. The LAC is a municipal register containing individual and household-level information of each person registered as resident in a given municipality at a given time. In the months preceding Census Day the LACs have served a number of purposes, including: a) identifying the relevant census units (household or cohabitation); b) printing, mailing-out and password creation for web-based access to census questionnaires; c) pre-loading of the data into the Survey Management System (Sistema di gestione della rilevazione – henceforth SGR).

7. ISTAT defined a common data transmission protocol for all municipalities with the following technical specifications: a) the name, structure and format of the files, including instructions for decoding variables such as "relationship", "marital status" and "gender"; b) features, format and length of the variables.

8. The electronic files were sent through a secure protocol (https) in order to ensure privacy. ISTAT enforced these technical specifications by implementing a web-based application (Starlac) which was used to collect and check the data transmitted by the municipalities. Starlac allows checking two kinds of errors in the input phase: a) Random checks for non-compliance to the required technical standards; b) Systematic checks for non-compliance to the legal provisions on population registers recordkeeping (Regolamento Anagrafico)[2].

[2] The *Regolamento Anagrafico* (D.P.R. n.223 del 30 maggio 1989) sets the recordkeeping standards for the LACs. Owing to the different pace at which municipalities have been migrating from paper to electronic registers, the harmonization process is still under way.

9. Only those LACs fully compliant to technical specifications and with type b)'s errors falling within defined tolerance ranges were further checked for variable coherence and correctness and processed according to the following steps:

a) Identification of census units (households or cohabitations);

b) Imputation of households' heads (in case of none or multiple heads of household);

c) Standardization and geo-coding of the addresses to EAs codes;

d) Sampling of households from selected EAs to receive the long form questionnaire;

e) Choice of questionnaire's model (_6p or _3p) based on household's size;

f) Assignment of a unique questionnaire's id code;

g) Password creation to access the online questionnaire;

h) Printing and mail-out of questionnaires;

i) Data loading on the SGR.

10. The final database was made up of about 61 million individuals, 25 million households and about 300,000 cohabitations. Over 93% of all municipalities successfully transmitted the data at their first attempt. Only one out of 8094 municipalities did not send the LAC, while 77 municipalities had to re-transmit the data due to errors. During the operations of data collection and elaboration, municipalities were supported by a task force providing troubleshooting over the phone.

11. To carry out all the preparatory operations detailed above, municipalities were required to transmit their LACs to ISTAT between January and mid-February 2011 detailing information on all registered residents as of December 31st 2010. In order to update SGR with all demographic variations such as births, deaths, or changes of address occurred between the end of 2010 and Census Day (October 9th 2011) ISTAT has devised three different strategies: 1) municipalities under 5,000 inhabitants were allowed to manually loading all demographic changes occurred up to October 8th directly on the SGR; 2) municipalities with a population size between 5,000 and 20,000 were also given the option of manual loading alongside the alternative of a second release of the full archive; 3) municipalities over 20,000 has to submit a full updated copy of their register. About 5,600 municipalities have manually loaded the variations directly into the SGR. About 2,400 municipalities (50 million individuals and 20 million households) send a second LAC referred to the 8th of October 2011. For these municipalities, the update of the SGR was carried out by ISTAT. About 100 municipalities had to resend the data because of the following errors: a) Data files referred again to 31/12/2010; 2) Data files containing only demographic changes; c) Errors in the coding or formatting of some of the main linkage keys (e.g. family code or address). The data were screened for errors as after the January release.

III. Integrating the LAC: The auxiliary population registers

A. The *Liste Integrative Anagrafiche Comunali*

12. This section describes the role of the Supplementary Municipal Population Registers (Liste Integrative Anagrafi Comunali or LIACs) within the computerized procedures that were put in place to support the 15th Italian Population Census.

13. Firstly, a relational database was designed to help the process. For each entity (individuals, households and cohabitations) two main tables were created: the first one with the raw data of the LACs, the second one with the derived data and the indicators. The maximum result in terms of data homogeneity was obtained by applying techniques of cleaning and checking. Standardization methods (e.g.: well-defined formats for all the fields) and common rules (e.g.: exceptions' handling for missing values) operated during the collection phases. In particular, the formal controls were very careful on those attributes to be used as linkage keys between the two releases of the LACs such as the individual id, tax code (codice fiscale), household/cohabitation code, toponym code.

14. Secondly, once the formal controls were passed, the data were subjected to consistency checks at several levels: single record, type of residence, municipality. An indicator was calculated for each error at every level, according to the following taxonomy: a) individual-record indicators concerning errors related to the id codes; b) individual-record indicators concerning errors related to other individual attributes; c) household-level indicators; and d) municipality-level at indicators.

15. A taxonomy of changes between the two LAC's releases with their frequency of occurrence are summarized in Table 1:

Table 1: The nature and size of the LACs' update

Type of variation	Households	Individuals
New households	662,931	1,099,563
Households with at least one component added	356,034	851,870
Households with at least one component deleted	947,764	1,183,497
Households which changed their residence address	232,416	380,372
Deleted households	523,911	523,911
Households with a new head	139,840	263,771

16. The linkage process was organized by steps in accordance with a logical sequence. Every step is preparatory to the following one, and produces a detailed log on the results of its processing. The structure of the log is such that it enables the execution of both simple automatic analyses and complex statistical investigations, in order to decide on the prosecution of the process every time at the end of a step. In addition, every step included comparisons with information coming from certified external sources, in order to carry out quantitative consistency checks based on thresholds set up depending on the population size of the municipalities.

17. In a first stage the changes were identified at the individual level: the primary key for the record linkage was the individual code followed by the tax code and, in sequence, up to the concatenation of several attributes. A more sophisticated procedure was adopted in order to determine the residence changes of households and cohabitations, using both raw and normalized addresses as well as the toponym code.

18. The effectiveness of each linkage key was rated and recorded municipality by municipality, according to the results of the matching procedures3. The resulting LIACs

[3] These measures of effectiveness are expected to be useful for future surveys, particularly for the idea of an Italian "rolling census" currently under study by ISTAT.

were stored into two tables, one for the households and one for the cohabitations, which include all the typologies of change both at the individual level and at the households/cohabitations level. Such tables include the linkage keys to the two reference LACs as well as all the register attributes. In conclusion, the major difficulties arose from the controls on the residence addresses. However this aspect should improve after the alignment, to be operated by the Register Offices under the set of rules provided by ISTAT.

B. The LIFA and the RNC

19. Although by law LACs should be kept constantly up-to-date in order to provide at any time a precise snapshot of the resident population living within municipal borders, mismatches between resident and registered population are not uncommon. It can happen that households or individuals live in a municipality without appearing in the local population register. On the other hand, there are cases of households or individuals who are regularly registered in a LAC but live in a different municipality. The reasons behind these anomalies can be diverse. For non-Italians, for instance, the red tape and hassle usually involved in obtaining the registration typically discourages the applicant. For Italians, on the other hand, electing a place where they do not actually live as their official residence is often an expedient to elude local taxation.

20. These flaws have prompted ISTAT to consider auxiliary registers from other sources to complement the LACs and LIACs during the census. As for the LIACs, the information was used to help enumerators locate households or individual respondents at risk of being excluded from the census. Two alternative archives were created for each municipality: a) a list of individuals (LIFA) who were either unregistered or, if registered, were using a second address in a different municipality; b) a list of valid addresses (RNC) not associated with any registered household.

21. The LIFAs are the result of a series of record linkage operations between the following archives: a) the LACs as of 31/12/2010; b) the Tax Register (AT) as of 30 April 2011 which was mainly used to obtain second addresses (*domicilio fiscale*) for Italian as well as other European citizens residing in Italy; c) the Permit to Stay (PS) register as of 30 November 2010 for individuals with a non-European passport and a regular permit to live in Italy[4].

22. The PS archive was obtained after a process of data cleaning, de-duplication, geo-coding and geo-referencing applied to about 4 million individual records collected by the Department of Public Security (DPS) and the Department for Civil Rights (DCR) of the Ministry of Interior. The raw data are stored into 11 different and partially overlapping databases depending on the processing stage of each individual application[5]. At end of the process about 700,000 duplicate records were identified and dropped according to an

[4] Whenever an individual ended-up in the LIFA both from PS and AT, it was the more up-to-date record from the latter source which was retained after a final de-duplication step.

[5] At the time of acquisition, DPS' databases contained information respectively about 185,000 newly filed applications still at the pre-processing state; 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 is in charge of three additional archives covering information 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.

agreed protocol of exclusion rules leaving a final PS archive of approximately 3.3 million records[6].

23. The AT register gathers basic personal information on all individuals endowed with a tax code including their reported *domicilio fiscale*[7] and contains approximately 80 million records. After excluding all those records having transitory nature (e.g. foreign artists who come to perform and then leave the country) or still referring to deceased individuals, the surviving records underwent a validation process whereby they were retained in the archive only if the information provided was confirmed by other administrative registers such as student archives, business registers, maternity ward registries, pensioners' archives and others.

24. The RL between AT and LAC (RL1) was by and large a deterministic merge using the tax code as the matching key at the national level[8]. The RL between PS and LAC (RL2) required a more sophisticated strategy both deterministic and probabilistic in view of the lesser quality of the available matching keys for non-European residents. Given the number of records involved, the linkage was usually performed at the province level (regione). However, for densely populated localities it was necessary to bring down the analysis to the district (provincia) or even to the municipal level. The LIFA was obtained by appending the unmatched residuals from RL1 and RL2 and removing the overlap with a final de-duplication step[9]. The final archive contained just over 2.1 million records evenly divided between PS and AT. Of these about one third were individuals registered in a LAC but using a second address, while the remaining two thirds were resident individuals not registered in any LAC.

25. The RNC is geo-referenced database of valid addresses geo-coded to census EAs for all municipalities with the population size of 20,000 or more. Although the RNC has been primarily engineered for use in the Census of Buildings it has also been used alongside the LIACs and the LIFAs to help locate unregistered households and reduce LAC's undercoverage errors. For this purpose about 2.6 million addresses were loaded in the SGR and made available to enumerators.

IV. Preliminary results

26. At the time of writing this paper census operations are about to conclude but still underway in many large municipalities, particularly those with a population size exceeding 150,000. The metadata on the returned questionnaires are final only for municipalities under 20,000 where census operations closed on 31 December 2011.

27. Table 2 shows some preliminary evidence from the returned questionnaires on the coverage and performance of the registers used in the census, broken down by region and municipal population size. In municipalities with a population size below 20,000 about

[6] For details please refer to Fortini *et al.* (2011) *Counting in non-Italian residents: The use of the "Permits to Stay" archive in the next population census.* In Survey research methods and applications: proceedings of the second ITACOSM conference, June 27-29 2011, Pisa, Italy: Plus-Pisa University Press

[7] Article 43 of the Italian Civil Code defines the *domicilio fiscale* as the locality where the individual establishes her main business or income-generating activity.

[8] This RL operation as well as the preparation of the AT archive described in the previous paragraph were implemented by colleagues at the Directorate for Administrative and Statistical Registers (DCAR).

[9] The overlap between PS and AT was minimal accounting for less than 4,000 records (about 0.2% of the consolidated LIFA).

94% of all registered households/cohabitations have filled and returned the questionnaire. Percentages are clearly lower for more populated municipalities because enumerators are still scouting the EAs collecting questionnaires. In the below-20,000 group of municipalities about 6% of all mailed-out questionnaires were not returned because the household was no longer living at that address (LAC's overcoverage). On the other hand about 4% of all returned questionnaires were dispatched by enumerators to unregistered households (undercoverage). The results reveal that the retrieval of the undercoverage was carried out overwhelmingly on the basis of the LIACs. Somewhat disappointingly both the LIFA and the RNC appear to have played a very marginal role in supplementing the LACs. Whether this was due to the rapid obsolescence of these registers, to the lack of integration with the LIACs between the auxiliary archives or to their low degree of utilization by municipalities during the census remains to be established.

Table 2: Questionnaires' returning rates and LACs' coverage (%)

Population size	Archives				LAC's over coverage	LAC's under coverage
	LAC	LIAC	LIFA+RNC	All		
Total	89.5	62.7	2.8	69.0		
<=20,000	93.9	80.4	2.8	72.1	6.1	4.1
>20,000	85.8	49.5	2.9	66.3		

V. Conclusions

28. The 15th Italian population census has been assisted for the first time by administrative registers. The pivoting role was entrusted to the municipal population registers which were used for instance for the mail-out of about 25 millions questionnaires. These register were complemented by other archives in order to reduce potential deficiencies and consequently the risk of leaving a significant number of unregistered households uncounted. With the census still underway the preliminary evidence points to an overall success of LACs in terms of guaranteeing a high level of coverage. The overall contribution of the auxiliary archives deployed to supplement the LACs has so far fallen short of expectations and perhaps a careful cost-benefit analysis is in order at the conclusion of the census round to discuss about their use in the future.