United Nations ECE/CES/GE.41/2017/13/Rev.1

# Economic and Social Council

Distr.: General
27 July 2017

Original: English
English and Russian only

## Economic Commission for Europe

Conference of European Statisticians

### Group of Experts on Population and Housing Censuses

**Nineteenth Meeting**
Geneva, 4–6 October 2017
Item 2 of the provisional agenda
**Innovations in census methodology and technology, and results of testing**

# The use of continuity patterns in administrative data to define the usually resident population in Italy

### Note by the Italian National Institute of Statistics

*Summary*

The strong effort in micro-level integration among different statistical sources together with the availability of an increasing number of administrative archives is determining a big change in the processes that the National Institutes of Statistics adopt to produce population statistics. The Italian National Statistical Institute (Istat) is planning a new design for the next Census round, based on a convenient integration of administrative data and surveys.

A thematic database has been created to study how administrative sources could improve the quality and information of population registers: sources integrated are official municipal population registers together with administrative archives from labour market, education, data on income and taxation.

The aim of this work is to point out how this integration of data in proper registers could allow discovery of new relevant information about population: clusters of individuals determined by patterns emerging when analyzing 'records of presence' in different sources and according to a time-period could be of great interest for population studies.

# I. Introduction

1. For years the population counts have been based on demographic surveys and Population Census on one hand, and municipal population registers on the other. In the past, the integration among these sources was set up at aggregated level and results of the Census were used to increase accuracy of municipal population register.

2. Nowadays, the strong effort at micro-level integration among different statistical sources together with the availability of an increasing number of administrative data is determining a big change in the production of population counts.

3. The population Census still remains the largest and most important statistical data collection to provide population figures at the smallest geographical units: while most statistical advanced countries still use the traditional scheme, with complete enumeration of population and housing units (i.e, United States of America and Canada), an increasing number of countries base their Census production on statistical registers. Census register-based can use exclusively registers data, as in the case of Scandinavian countries (Netherlands, Sweden, Denmark, Finland and Norway), or can use a combination of both registers and sample surveys data within the frame of the so-called 'combined Census' (i.e. Spanish 2011 Population Census).

4. The Italian National Statistical Institute (Istat) is planning a combined Census scheme for the next Census round, by conveniently carrying on the integration of administrative data and surveys and then exploiting this new informative richness.

5. Adding vital events (births, deaths, internal and international migrations) to Population Census microdata, Istat has been computing a statistical population register, the so-called "ANagrafe VIrtuale Statistica" (ANVIS). This statistical register ensures higher level of quality than municipal administrative population one and represents a solid component for a frame of register-based production.

6. Since 2015 many tests were carried out at Istat on the use of administrative sources in order to comply with the definition of usual resident population included in the European Union Regulation (No. 1260/2013). According to this regulation, the "usual residence" is the place where the person lived for 12 months before the reference date. People who arrived in the place less than twelve months before the reference data with the intention of staying there for at least one year, are also considered as usual residents.

7. The aim of this work is to evaluate whether records from administrative sources may meet the requirements of the international definition and they allow computing the usually resident population in Italy. This goal requires selecting and evaluating administrative data with the aim to recognize patterns and associations.

8. Another relevant issue is to verify whether the integration of data in proper registers could allow knowledge discovery of clusters of individuals determined by patterns emerging when analyzing 'records of presence' in different sources.

# II. The integrated system of administrative records

9. To manage the increasing number of administrative data sets and to maximize the benefit, Istat built an integrated system of available administrative sources, named SIM (Integrated System of Microdata). When a new administrative archive is loaded in this system, the recognition process identify any individual or economic unit which are present in data and assigns it a permanent and unique identification number (ID): if the unit is already present in Istat databases, this ID is the same the unit was assigned in the past (di

Bella and Ambroselli, 2014). Then, starting from this base, it is possible to construct specific data structures for statistical processes and to create thematic data base.

10. Among all the archives loaded, SIM comprises data coming from ANVIS, permits to stay, data referring to employees and self-employed workers, compulsory education students, university students, retired people, non-pension benefits records, and individual data on income and taxation. These integrated data have been used to create a thematic database to analyze how administrative sources could improve the quality and information of population registers (Chieppa et al, 2016). The permanent identification number allows to link social and economic variables of individuals, households, and economic units to the place where people perform their activities or spend their time.

11. Linking official population registers to subject-specific administrative sources (Labor and Education registers, Tax Returns register, Earnings, Retired, and Non-Pension Benefits registers, Permits to Stay archive) could help identify groups corresponding to the national or international definition of "usually-resident population". Under-coverage errors could be corrected by using individuals' records of presence on the Italian territory; whereas the absence of records for people in the population register could testify for over-coverage.

12. To exploit the administrative sources, the Italian statistical office uses the scheme of Knowledge Discovery from Databases which is based on a learning-from-examples technique. It includes data storage and access, scaling algorithms to massive data sets. This process leads to identify patterns and groups.

# III. The identification of continuity patterns from administrative records

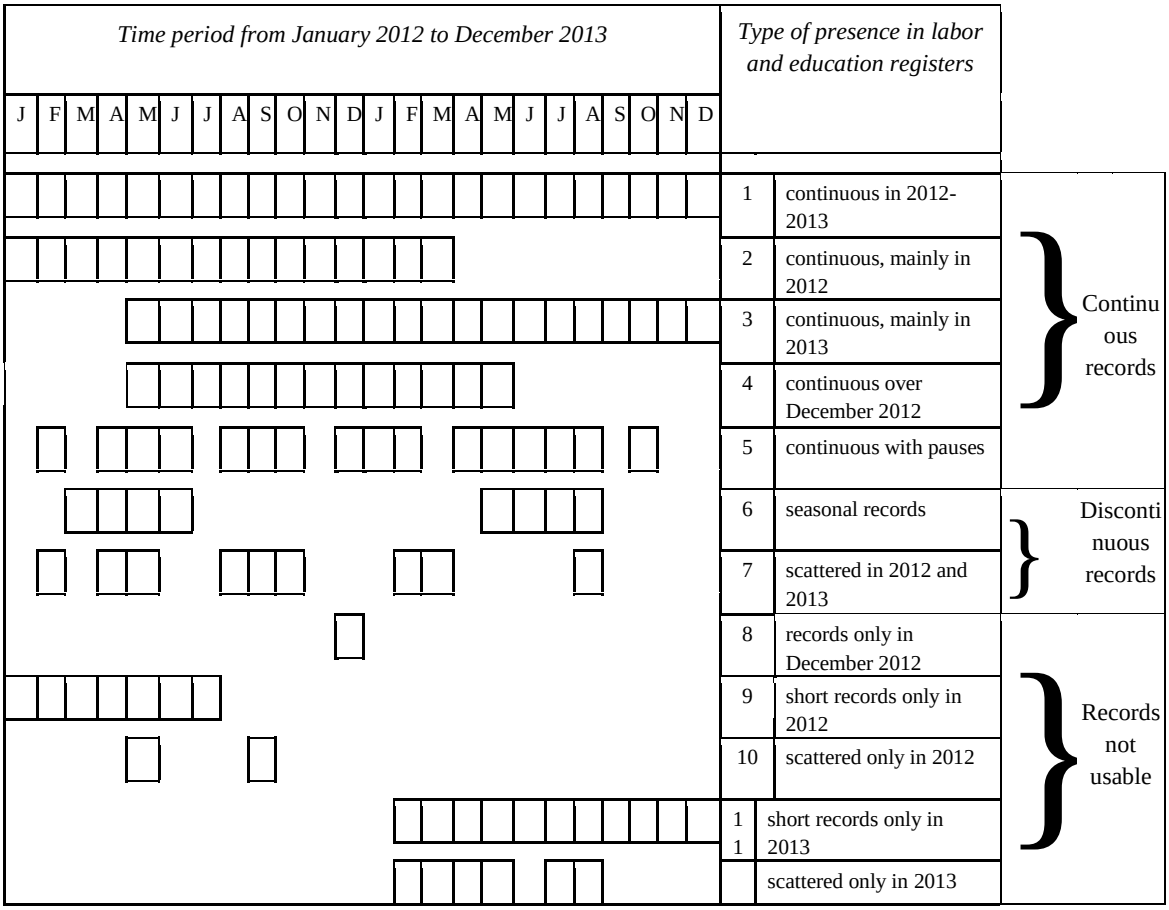
13. With the goal of discover records of presence that are consistent with the international definition of usual residence in Italy, administrative data were selected and evaluated with the aim to recognize patterns and association of records. Before analyzing records from administrative data, the reference period must be chosen: an individual's presence on a territory is inevitably linked to a time interval. Choosing the reference period is a crucial point when identifying and analyzing the records of presence in Italy: the longer the period is, the easier it is to assess the weight of the record, in terms of continuity or discontinuity over time, number and repetition of migration processes, job mobility, course of study, etc.

14. In order to meet the requirements of usual residence definition (even considering the intention to stay for at least 12 months), Istat considered a time-period of 24 months. Then considering a reference date, for example 31 December of 2012, records analysis is performed observing the previous 12 months and 12 months after the reference date.

15. Each signal can be attributed to a specific individual and to a certain geographical area. During the reference period, if a record is detected in a labour register and another in an education register, but both are located in the municipality, we will have a single record from that specific municipality. This record is tagged with an attribute that allows the tracking of the individual in both the original sources, and also an attribute relating to the duration of presence expressed in terms of job or study activity.

16. Records related to labor and education are to be considered the most reliable, and they include duration and type of activity. They allow to know where an individual was, by month, and what he or she was doing. Clustering these sequences of records, table 1 shows patterns with continuous records and those with discontinuous presence in Italy. The pattern of "continuous presence in 2012 and 2013" (1) corresponds to continuity during the period.

Table 1.
**Scheme of continuity patterns in work and study activities**




*Source:* Istat

17. Continuous presence, mainly in 2012 or in 2013 (patterns 2, and 3), characterizes people present during at least twelve months, and who have changed, started, or ceased an activity, or have moved. Continuous presence over December 2012 regroups people with at least twelve months of continuous presence from 2012 to 2013 (pattern 4). Pattern 5 regroups discontinuous records lasting more than twelve months; pattern 6 includes "seasonal records" appearing only once a year; pattern 7 combines scattered records with less than twelve months. The profiles 8 to 12 are not exploitable, because they escape the definition of usual residence. Their records show only one month of presence (pattern 8), disappear before December 2012 (pattern 9), appear after that date (pattern 11), or recur randomly in 2012 or in 2013 with less than twelve months of presence (pattern 10, pattern 11, and pattern 12).

18. So, records can be used to derive new relevant variables for related individuals and their type of living conditions in Italy. More specifically with regard to population counts, this new information could identify cases of permanent presence that correspond to the usual residence definition and concept of the international regulations, that not always correspond to what is record in the population registers.

19. Demographic variables, especially gender, age and country of citizenship as well as the location of the records on the territory have proved to be very significant variables in defining specific sub-population profiles.

# IV. Population classified according to register consistency

20. We classify the cases recorded in the experimental database on the basis of record consistency (Table 2). We first link the Population register, the activity records from Labor and education registers, and the permits to stay register together, second we isolate retired people or people with other benefits from the Labor and education registers; third, the Tax Returns register inform on people able to justify their presence in Italy.

21. In the first step, the linkage allows us to delineate four groups: Group A, comprising 36.6 million individuals recorded in the Population register and in Labor and education registers. Group B comprises 24.5 million people recorded in the Population register but not in the Labor and education registers. Group C comprises 1.1 million people not recorded in the Population register but recorded in Labor and education registers. Group D comprises 351 thousand people registered only in the Permits to stay archive.

Table 2.
**Proceeding scheme and population counts (in thousands) according to eligible or possible residency in Italy at 1st January 2013**

| Step I | | | | | Step II | | | Step III | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Population register | Labor-Education register | Permits to stay archive | Group | Count | Retired-non pension-benefits registers | Group | Count | Tax return register | Group | Count |
| 61,068 | 37,704 | 3,378 | | | 20,764 | | | 26,649 | | |
| records | | | | | records | | | records | | |
| yes | yes | - | A | 36,618 | | | | | | |
| yes | no | - | B | 24,450 | yes | E | 14,485 | | | |
| | | | | | no | F | 9,965 | yes | G | 6,939 |
| | | | | | | | | no | H | 3,026 |
| no | yes | - | C | 1,086 | | | | | | |
| no | no | yes | D | 351 | | | | | | |

Eligible as usual residents

Possible residents

*Source:* data from *Istat*

22. In the second step, we link Group B to the register of retired-people and maternity allowance and unemployment benefit. From this linkage, we distinguish whether the individuals are recorded in either one of these two registers (Group E, comprising 14.5 million people) or not (Group F comprising 10 million people).

23. In the third step, we link Group F to the Tax Returns register, in order to identify people financially supported by individuals who have taxable income (Group G, comprising 6.9 million people) or people recorded in the Population register without records in other available sources (Group H, comprising 3 million people).

24. Members of Groups A, E, and G are residents, while those of Group C, D, and H are "might-be" residents in Italy. So, from this point on, we focus on Group C (1.1 million of individuals) which brings together the possible under-coverage of population registers. Table 3 distinguishes people from Group C by type of record and duration: 409,157 individuals (Groups C1, C2, and C3) have continuous records, most of them are foreigners

(90%), and have a professional activity (78%), while the others are in school (58,000) or students (33,000).

Table 3.
**Groups C: people not recorded in population registers but in other administrative sources at 1 January 2013**

| Groups | Sources and records | Absolute values |
|---|---|---|
| C1 | Workers with continuous records | 318,159 |
| C2 | University students with continuous records | 32,671 |
| C3 | Primary and secondary school students with continuous records | 58,327 |
| C4 | Discontinuous records of presence | 266,763 |
| C5 | Non usable records of presence or no monthly information | 410,242 |
| Total | | 1,086,162 |

*Source:* data from *Istat*

25. Discontinuous records (Group C4) amount to 266,763, mostly are foreigners (90%). Among discontinuous records 410 thousand individuals are not exploitable because of the lack of information on them (Group C5).

# V. Longitudinal approach and preliminary results

26. A more detailed analysis of groups at risk of under-coverage in populations registers has highlighted some important topics when we use a longitudinal approach. Analyzing all individuals who are not registered in the population registers at 2012 and looking at them for the next three years, you may notice the following three groups (Table 4)[1]:

(a) People with records of presence in 2012[2] recorded in the population registers by the next two years. This group includes more than 165,000 individuals (about 34% of possible under-coverage) who shown continuous records of presence in Italy, and 76,000 other people with discontinuous records;

(b) People with records of continuously presence from 2012 to 2014 but never recorded in the population registers. This group brings together more than 180,000 people (more than 37%) with continuous records, and about 90,000 with discontinuous records. Both types of people represent the "core" of under-coverage of population registers;

(c) People with records of presence only in 2012 and 2013 (no in 2014) without any registration in the population registers count more than 140,000 individuals (almost 30%) with continuous records, and other 317,000 individuals (about 66%) with discontinuous records. Both types of people might have definitely left the country or might be in a precarious condition for losing work.

[1] Individuals who are not exploitable have been excluded from longitudinal analysis (group C5 of table 3).

[2] It should be considered that for each year (2012, 2013, 2014) we observe the records of administrative data from the previous 12 months to the 12 months after the 1st January reference data.

Table 4.
**Longitudinal counts of people with records of presence in Italy in the time period 2012-2014**

| **Records in the Labor-education registers** | | | | **Records in the Population registers** | | | **Continuous records** | **Discontinuous records** | **Profile of groups** |
|---|---|---|---|---|---|---|---|---|---|
| 1 Jan 2012 | 1 Jan 2013 | 1 Jan 2014 | | 1 Jan 2012 | 1 Jan 2013 | 1 Jan 2014 | Abs. Val. (000) | Abs. Val. (000) | |
| yes | yes | yes | | no | yes | yes | 165,4 | 76,0 | People with records of presence in 2012 recorded in the population registers by the next two years |
| | | | | | | | *33.8%* | *15.7%* | |
| yes | yes | yes | | no | no | no | 182,5 | 89,7 | People with records of presence from 2012 to 2014 NOT recorded in the population registers |
| | | | | | | | *37.2%* | *18.6%* | |
| yes | yes | no | | no | no | no | 142,0 | 317,6 | People with records of presence only in 2012 and 2013 without any registration in the population registers |
| | | | | | | | *29.0%* | *65.7%* | |
| | | | | | | | **489,9** | **483,3** | **Total amounts** |

*Source:* data from Istat

27. When considering under-coverage by using longitudinal approach, the analysis shows relevant results. First of all, the criteria of observing the records of administrative source in a time-period of 24 months in order to meet the usual residence definition is performing. Moreover, clustering the individuals according to continuity or discontinuity patterns represents a useful classification tool to identify a stable presence on the territory, especially in the case of foreign nationals.

28. Geographical location and specific citizenship will be essential for identifying the cross-border workers, whose absence from the population register is admissible.

29. However, longitudinal analysis revealed that also some individuals with not continuous records may be associated with stable presence on the territory (with a range between 15 and 18%), and for this reason an improved characterization of records is needed.

30. Records in administrative registers improve the knowledge of the population, and shown that once the population registers data have been linked to other administrative sources, Istat may evaluate the quality and accuracy of the sources. Some foreigners or some mobile people elude the population registers, but are recorded in labor registers.

# VI. References


- di Bella G. and Ambroselli, S. (2014). Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat. Paper presented to the European Conference on Quality in Official Statistics Q2014. Vienna: 2-5, June:1-14.

- Chieppa, A., Gallo, G., Tomeo, V., Borrelli, F., Di Domenico, S. (2017). Knowledge Discovery Process to Derive Usually-Resident Population from Administrative Registers. Mimeo.
- Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology. 40(2): 137-161.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining, U.
- Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds). AAAI Press, Menlo Park: 134.
- ISTAT (2014). La misurazione della qualità del 15° Censimento generale della popolazione e delle abitazioni: i risultati dell'indagine di copertura (PES). Seminario del 27 giugno, Roma: http://www.istat.it/it/archivio/126014.
- Unece (2013). Population Definitions at the 2010 Censuses Round in the Countries of the Unece Region. Paper presented to the Fifteenth Meeting of Group of Experts on Population and Housing Censuses. Geneva: 30 September – 3 October 2013. https://www.researchgate.net/publication/260096889_Population_definitions_at_the_2010_censuses_round_in_the_countries_of_the_UNECE_region: 1-16.