United Nations ECE/CES/ GE.41/2016/3

Economic and Social Council

Distr.: General
7 July 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Eighteenth Meeting
Geneva, 28 - 30 September 2016
Item 5 of the provisional agenda
Methods for assessing quality and usability of registers and administrative sources

The integration of administrative data sources in Italy to increase Population Census data availability

Note by the National Institute of Statistics of Italy[1]

Summary

The next population census round in Italy will mark the definitive transition from the traditional "door-to-door" enumeration to a "register-based" system, which combines official population registers with other subject-specific administrative sources. By using an integrated system of registers it is possible to identify useful patterns in huge amounts of administrative data and create sub-population profiles. The Italian National Institute of Statistics (ISTAT) carried out an initial trial to identify the usually-resident population by using administrative data, which produced useful results. ISTAT analyzed the quality of the registers and identified patterns in the data. These patterns enabled ISTAT to classify individuals into specific groups, which also represent the "critical" subpopulation to be considered when defining the new census strategy. ISTAT then defined a preliminary workflow for deriving usually-resident population counts.

[1] Prepared by Gerardo Gallo, Angela Chieppa, Valeria Tomeo and Stefano Falorsi.

I. Introduction

1. The next Population census round in Italy will mark the definitive transition from the traditional "door-to-door" enumeration to a "register-based" system. Despite the strong efforts made during the 2011 Population census to increase the coverage by using different modes of data collection and administrative data as frames for enumeration, some groups of people remained very difficult to count by using traditional census. After the 2011 Post-Enumeration Survey, which involved more than 320,000 individuals, the Italian National Institute of Statistics (ISTAT) estimated that about 650,000 usual residents in Italy had not been counted by the census and that about 80% of them had foreign citizenship (ISTAT, 2014). These results, together with the huge cost of traditional enumeration and the significant statistical burden on respondents, led ISTAT to adopt a new census strategy based on the combined use of multiple data sources and surveys (Crescenzi et al., 2015).

2. Poulain et al. (2013) and Wallgren et al. (2011) pointed out that the use of the population register to count the usually-resident population does not automatically solve coverage problems: the accuracy of this register represents a critical issue even in the countries that have been using the register-based census for decades. Administrative data on a population involve problems of under-coverage, due mainly to the difficulty in recording international migration. In Italy, nationals living abroad and foreigners who have left the country permanently or long-term should be removed from the population register. However, emigrants see no reason to notify the authorities of their departure. In addition, local authorities have an incentive to maintain the stability of their population numbers by considering these people "temporary" emigrants, so they keep them in the register. A second critical group is the foreign population usually resident in Italy but without official residence. Although coverage of the usually resident foreign population in Italy is monitored, there are foreign citizens who do not want to be recorded in the population register, especially those nationals from the European Union member States. Moreover, those who have lost the requirements for staying in Italy and immigrants who entered without authorization cannot be recorded in the population register [2]. The latter, who do not appear in any administrative sources, represent a hard-to-count sub-population and they require a suitable survey to be counted[3].

3. A preliminary solution to face the coverage of population register could be the use of an integrated system of registers: official population registers would be linked together with other subject-specific administrative sources (related to labor, education, taxation, earnings, etc.). This system could be used to identify groups that correspond to the national or international definition of "usually resident population". Under-coverage in population registers could be corrected by using individuals' signals of presence on Italian territory coming from other registers; whereas the absence of signals for people in the population register could be evidence of over-coverage. Deriving census figures for the usually-resident population by using a system of registers is a challenge that involves the development of a new production process based on this new data framework.

4. With the prospect of a continuous census in Italy (censimento permenente), which should be strongly oriented towards the optimization of administrative archives, ISTAT arranged a Working Group with the aim of testing the use of administrative information for

[2] As of 8 August 2009 a new bill (Law no. 94 of 15 July 2009) makes it a crime to enter or stay in Italy illegally. Therefore, foreign nationals caught entering or staying in Italy without permission commit the offence of illegal immigration.

[3] Among the hard-to-count sub-populations, the homeless and people who do not have usual residence in the same place during the year should also be included.

census purposes and for the production of population counts. The results of a trial carried out by ISTAT's researchers will be presented in the following paragraphs. In order to exploit the wealth of information contained in the administrative sources, ISTAT used the scheme of Knowledge Discovery from Databases (KDD). This is a structured, iterative process, in which some of the variables to be analyzed are created during construction. This process allows the identification of useful patterns and specific subpopulations in a large amount of administrative data. The exploitation and analysis of ISTAT's repository and administrative registers has been carried out as an exploratory step to define the new census strategy.

5. The first goal of the test was to define a preliminary workflow on how to process available administrative data for deriving usually-resident population counts. This involved the selection and quality evaluation of administrative data. The second goal was to discover patterns and associations in administrative data and then to select the more relevant features and derived information from the available attributes in the registers, or to create new features as functions of the original ones. Moreover, the discovery of patterns would enable ISTAT to classify individuals into specific groups, which represent "critical subpopulations" related to under- and over-coverage in official population registers and which must be taken into account when defining the new census strategy.

II Trial to derive usually-resident population using administrative registers available to ISTAT

6. To manage the increasing number of administrative data sets acquired for statistical uses and to maximize the benefit deriving from the huge amount of information available, it is necessary to build an integrated system of the administrative sources. ISTAT has moved in this direction by centralizing some functions for acquisition, storage, integration and administrative data quality evaluation in the system called SIM (Integrated System of Microdata). This system manages social and economic administrative data for: individuals and household characteristics; demographic aspects, employment status, level of education; places, in terms of residence, labor or education; the typology of the unit where people perform their activity or spend their time: houses, schools, places of work; the typology of the relationships among individuals, units and places. The integration step in the SIM system is the process of linkage and physical integration of microdata coming from different sources: depending on the linking variables available, a suitable integration strategy and a set of algorithms are applied. Integration means: (i) identifying each object (individual, economic unit) in the administrative data sources with a unique and stable (over time) ID number; (ii) defining, for each object, the logical and physical relationships among sets of data coming from different sources. SIM has been created with the aim of supporting the statistical production processes: the use of the SIM IDs enables the building of data structures that could constitute the starting point for the statistical processes based on administrative data. It is possible to create a thematic database using a syntax to define the rules of extraction, by using the reference time or statistical domains or subsets of variables or objects etc.. To evaluate how administrative data could improve population counts, a thematic database of 'administrative signals' was created in order to support an initial experiment. Data used in the trial came from specific administrative sources already stored and integrated in the SIM system: Municipal Resident Population, Permits to Stay, Employees and Self-employed, Compulsory Education, University Students, Retired People, Non-Pension Benefits, Income and Taxation.

7. The base Population Register is still in a developmental phase at ISTAT: a specific "test version" has been created by combining all Municipal Population Registers at the reference date chosen for analysis which is 31 December 2012. Before analyzing signals

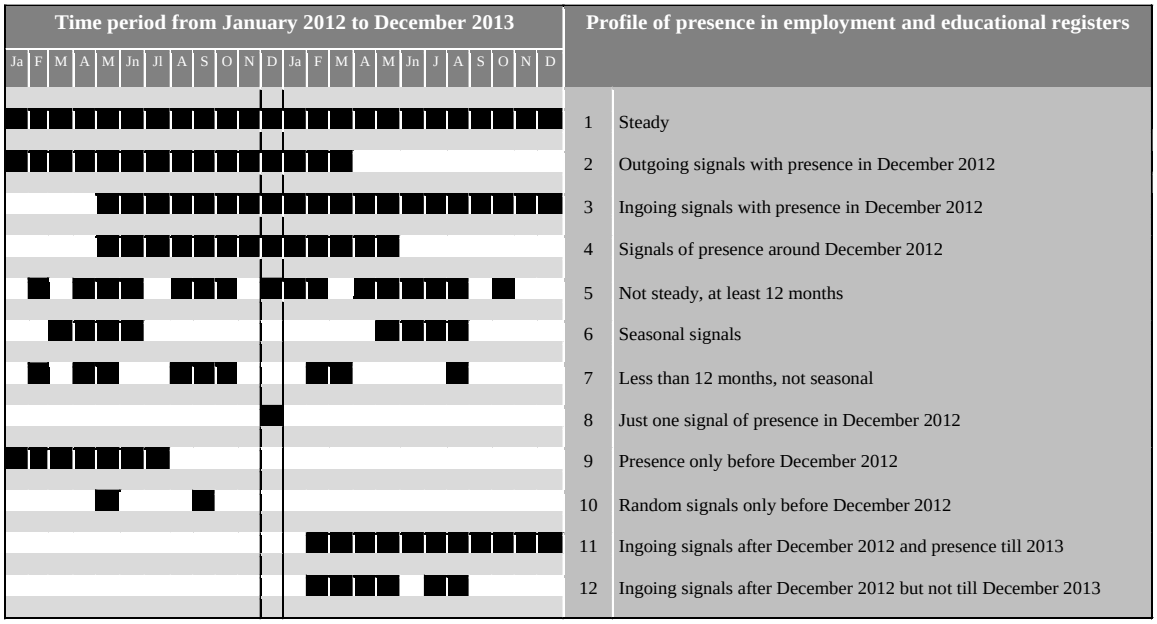
from administrative archives, the reference period must be chosen: an individual's presence on a territory is inevitably linked to a time interval. Choosing the reference period is a crucial factor when identifying and analyzing the signals of presence: the longer the period is, the easier it is to assess the weight of the signal, in terms of “continuity” and “stability” over time, number and repetition of migration processes, job mobility, course of study etc.. For the purposes of the trial, ISTAT used the definition of “usual residence” as set out in the European Union Regulation 1260/2013. On this basis, ISTAT chose a reference period from January, year T to December, year T+1, with 31 December, year T being the reference date. The period considered for the evaluation of signals, therefore, has a length of 24 months. Each signal can be attributed to a specific individual and to a certain geographical area; the signal 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. In cases where the geographical information is not available, a dummy code is used to represent the lack of information.

8. Signals could be directly associated with an individual unit or could refer to another person who has a relationship with the individual unit; in this last case they are “indirect signals” and less strong than the direct ones. Furthermore, there is a hierarchy of the sources, in that the signals coming from the “activity registers”, related to labor and education, should be considered robust signals because going to school or holding a job represent guaranteed presence on a territory.

9. Moreover, the activity registers included more detailed information, such as duration and type of activity, which provided useful features to analyze. By using a standardized time reference for the duration of an activity, it was possible to derive a monthly presence scheme. This feature consists of a sequence with a length of 24 digits, each of them representing whether a person was present for a specific activity and in which geographical area of the country. Clustering on these sequences, evident in the majority of the activity signals, showed the existence of patterns of continuity (Figure 1)

Figure 1

Monthly presence scheme of continuity’s patterns in job and study activities



10. “Steady”cluster (pattern 1) includes signals with a continuous presence and maximum steadiness during the entire reference period. The “ingoing” and“outgoing”

signals (patterns 2 and 3) with a presence of more than twelve months, including December 2012, represent people that have changed, begun or lost an activity or that move on the territory. “Around December 2012” represents the signals characterized by continuous presence, beginning in 2012 and ending in 2013, with at least twelve months (pattern 4). Also the discontinuous signals with more than twelve months are a relevant pattern (pattern 5). The “seasonal” signals include cases when the pattern scheme is repeated once a year (pattern 6). Then there are some less specific patterns, including clusters of signals with only one month of presence (pattern 8), discontinuous signals with less than twelve months of duration (pattern 7) and, finally, patterns of continuous signals but positioned entirely in 2013. For the purposes of the trial, Population Register and activity signals were loaded in a database which also contained individual characteristics, such as demographic attributes, which were very useful when analyzing the final results.

III Population groups classified according to register consistency

11. The evaluation of consistency between the registers examined determines a classification of all the cases recorded in the experimental database, into groups that are useful when producing population counts. The process to derive these groups was organized in three steps (Table 1). The first step involved linkage between the experimental Population Register (PR), the activity signals from Labor and Education registers (LE) and the Permits to Stay Register (PS); in the second step, which could explain the absence of previous activity signals, retired people or people with other benefit signals emerged from all the individuals without signals in the LE; finally, the Tax Returns Register (TR) provided indirect signals on people who could in some way justify their presence in Italy.

Table 1

Proceeding scheme and counts of population groups *(in thousands)* according to their eligibility or uncertainty to be usual resident in Italy

Step I					**Step II**			**Step III**		
PR	**LE**	**PS**	Group	Count	**RR-NPR**	Group	Count	**TR**	**Group**	Count
61,068	37,704	3,378			20,764			26,649		
Signals					Signals			Signals		
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						

Source: our own elaboration on ISTAT data

Table legend:

Eligible as usually residents

Uncertain residents

Administrative sources legend:

PR	Population Register
LE	Labor and Education Registers
PS	Permits to stay Register
RR-NPR	Retired and Non-Pension Benefits Registers
TR	Tax Returns Register

11. In the first step,12. In the first step, the linkage between the registers examined determines the four groups of people. Group A, which counts 36.6 million individuals, includes people recorded in the PR and in LE registers. Group B, which counts about 24.5 million people, includes people recorded in the PR but not in the LE sources. Those people not recorded in the PR but with activity signals from the LE sources are in Group C, which counts almost 1.1 million individuals. Finally, Group D, which counts 351 thousand people, represents the individuals registered only in the PS archive and not in the PR or LE sources. In the second step, Group B was linked to the Retired-people Register and to the Non-pension-Benefits Register (maternity allowance, unemployment benefit, etc.): the results of this linkage created two sets, according to whether the individuals have signals in these two registers (Group E) or not (Group F). So, Group E, with about 14.5 million people, included people in the PR with signals in the Retired and Non-Pension Benefits Registers (RR-NPR); Group F, which counted about 10 million people, is composed of the people recorded in the PR without signals of retirement or benefits.

13. In the third step of the process, Group F was finally linked to the Tax Returns Register (TR) to identify two other subpopulations based on the presence (Group G) or the absence (Group H) from the source of individuals declared as dependents for tax purposes: "spouses", "children" or "other family members". Therefore, Group G, composed of 6.9 million people, included those recorded in the PR with only indirect signals extracted from the TR. Group H, which counted 3 million people, included the cases recorded in PR but without signals coming from any other of the sources examined.

14. It is possible to organize the above mentioned groups according to evaluation of usual residence. So, the individuals belonging to Group A, Group E and Group G could be considered “eligible” as the usually-resident population. Instead, Group C, Group D and Group H could be considered as “uncertain” residents which would require a more thorough analysis. At the end of 2012, about 3 million people were recorded in Group H which represents the sub-population recorded in the population register (PR) without signals from other sources. Almost 75 per cent of these people had Italian citizenship. The gender composition appeared well-balanced for both Italians and foreigners and the median age slightly exceeded 40 years (Table 2). The distribution by age-group of this subpopulation perfectly reflected that of the total population resident in Italy in 2012. The percentage of Italian citizens at a more mature age (35-64 years) was about 14 percent higher than that of foreigners. In addition, there was a very substantial percentage of children under six years (just under 15 per cent for both Italians and foreigners). This percentage is attributable to births that occurred subsequent to the 2012 tax returns. About 400,000 children were not registered as dependents on the 2012 tax returns (they would be on the 2013 tax returns) and so this determines a number of children on the population/births register but with no signals in the administrative sources. Therefore, this sub-population should be reduced by 400 thousand units, because these children certainly belong to a family nucleus and, as such, do not represent "uncertain" usually-resident cases.

15. The geographical distribution by citizenship (Italians compared to foreigners) showed significant differences: the Central and Southern municipalities recorded the highest percentages of uncertain residents, both Italians and foreigners. The distribution of Italians was very concentrated in many municipalities of Lazio, Campania, Sicily and Calabria. On the contrary, foreigners were more widely distributed. In addition to the Southern areas, foreign people resided in municipalities situated in the west of the country and along the Tyrrhenian coast, especially in the Tuscan areas (Figure 2). For decades, municipalities in Tuscany have been the favorite destination of many retired people from Northern Europe, especially British citizens (Warners et al., 1999).

16. Medium or high percentages of people with Italian citizenship were only found in the Northern municipalities along the borders of Switzerland, France and Austria, due to the cross border-workers who usually resided in Italy.

17. Looking at the age structure for each country of citizenship group, the Chinese community showed the youngest median age (only 13 years) compared to Italians and other foreign communities. Among the latter, the Albanian community showed the highest median age.

18. Additionally, regarding foreigners, if the sex ratio (men per 100 women) reflected the gender composition of the foreign community in Italy, the distribution by country of citizenship of this subpopulation did not reproduce exactly the geography of the foreign presence in Italy. For example, immigrants from Moldova, India and Peru did not rank in the top ten nationalities resident in Italy (Table 2).

Table 2
Demographic structures of sub-population H: country of citizenship, sex ratio, age groups and median age

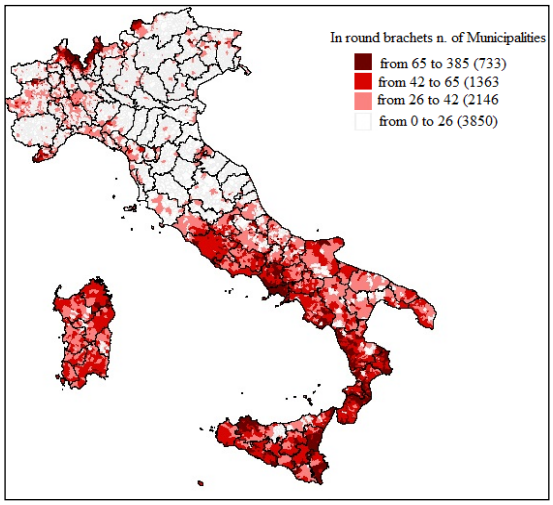
Country of citizenship	Absolute Values (*Thousands*)	Percentage	Sex ratio	Age groups (*Percentages*)						Median age
				Less than 6	6-14 years old	15-34 years old	35-64 years old	65 and over	Total	
Total	**3,023**	**100,0**	**101,3**	14.3	1.5	22.7	55.3	6.3	100.0	**40.6**
Italians	2,223	73,5	101,4	13.4	0.8	21.5	58.5	5.8	100.0	42.1
Foreigners *(of which):*	800	100,0	101,3	17.0	3.5	26.6	44.7	8.1	100.0	35.7
Romania	134	16,7	90,3	20.0	2.4	34.5	40.8	2.4	100.0	31.9
Morocco	72	9,1	159,6	19.7	2.9	22.5	46.9	8.0	100.0	36.7
Albania	60	7,5	96,1	16.0	1.4	19.0	40.6	23.0	100.0	48.9
China	34	4,3	107,7	40.2	10.0	15.2	31.3	3.2	100.0	13.2
Ukraine	20	2,6	45,3	10.5	2.3	24.8	56.9	5.5	100.0	40.6
Tunisia	17	2,2	192,7	18.8	6.4	24.0	47.5	3.3	100.0	34.5
Egypt	17	2,2	201,5	22.0	12.4	23.2	40.8	1.6	100.0	29.8
Poland	17	1,9	48,2	7.7	3.4	26.2	59.2	3.5	100.0	38.0
Philippines	15	2,2	78,7	27.2	5.3	22.3	40.2	5.1	100.0	30.5
Nigeria	15	1,9	99,1	24.3	2.4	38.2	34.0	1.1	100.0	29.9
Other cs'	396	49,5	98,1	12.2	3.2	27.1	47.5	9.9	100.0	37.6

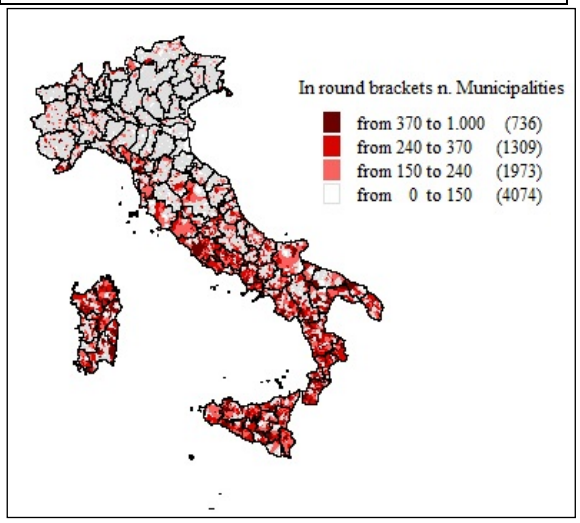
Source: our own elaboration on ISTAT data

Figure 2

Subpopulation H: people recorded in the population register (PR) without signals from other sources

Italian citizens *(Values per 1,000 Italian people)*	**Foreign citizens** *(Values per 1,000 Foreign people)*





Source: our own elaboration on ISTAT data

19. In conclusion, for the purposes of the permanent census, this group is a considerably critical subpopulation. Its amount, the absence of signals in the labor and education registers and the strong concentration in the Southern municipalities are strong risk factors for enumeration. Moreover, due to the ongoing economic crisis, large migration flows to other countries could occur, which, in cases of failure to notify the registry offices, could generate substantial over-coverage in population registers

20 The sub-population composed of people not recorded in the population register (PR) with signals from other administrative sources (Group C) accounted for 1,1 million people. All individuals in this group were not recorded in the PR but have one or more signals coming from one of the other administrative sources. Group C can be further divided into sub-groups: the majority of these signals come from labor and education activity records with a specific pattern of continuity, so that it is possible to take account of the type and duration of the related activities when analyzing this specific population (Table 3).

Table 3

Group C according to type and duration of signals

Sub-groups	Sources and type of signals	Absolute Values
C1	WORKERS	318,159
C2	UNIVERSITY STUDENTS	32,671
C3	PRIMARY/SECONDARY SCHOOL STUDENTS	58,327
C4	WEAK SIGNALS OF PRESENCE	266,763
C5	UNUSEFUL SIGNALS OF PRESENCE OR NO MONTLY INFORMATION	410,242
TOTAL		**1,086,162**

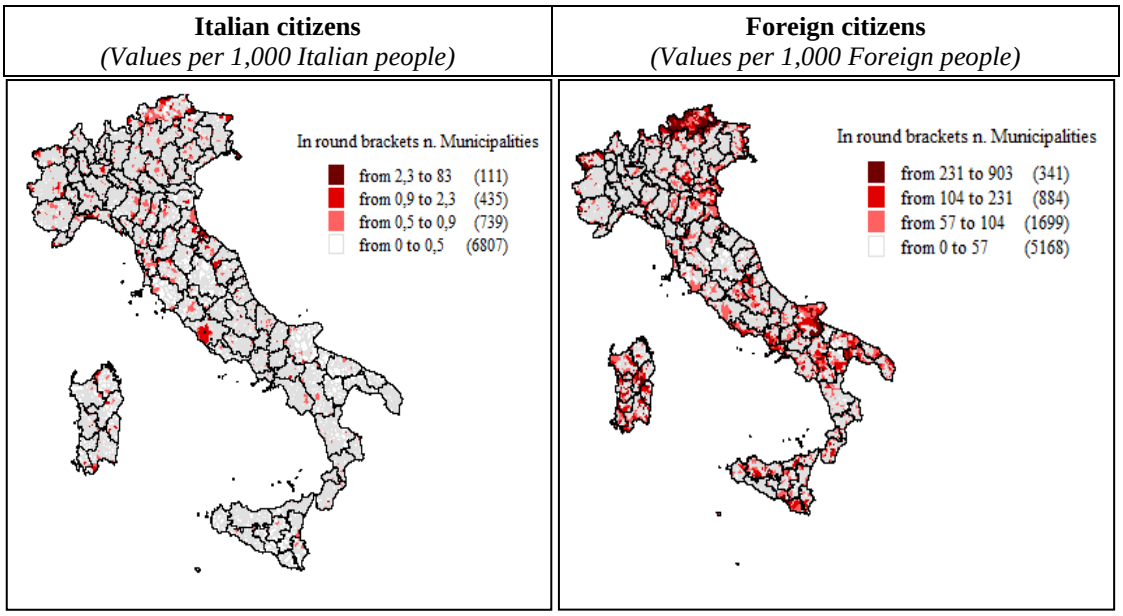
Source: our own elaboration on ISTAT data

21. An explorative analysis combined with expert supervision suggests an appropriate classification consisting of three main types of steadiness patterns. Firstly, there are “strong steady signals”: profiles from 1 to 5 can be considered the most relevant with respect to the official definition for usual residence, given that they show at least twelve months of presence, including December 2012. Then there are “weak not-steady signals”: profiles 6 and 7 of the monthly scheme, which include seasonal signals and non-seasonal, intermittent signals for a total duration of less than twelve months. The remaining profiles, from 8 to 12, are defined as “unuseful type of signals” in this trial, because they do not correspond to the EU definition of usual residence: among these are signals that disappear before December 2012 or signals that appear after December 2012 or completely random signals before December 2012. The latter needs a specific analysis in the future, because these activity patterns could, nevertheless, be related to some usually-resident people.

22. When classifying people from Group C by signal type and duration, five sub-groups emerged (Table 3). There are about 410 thousand cases which are not useful for analysis (Group C5) because they are individuals with “unuseful type of signals” or individuals with no monthly information at all. Strong-steady signals total 409,157 individuals: 78 per cent of them (318 thousand) are workers (Group C1) as the signal originates from the labor register; the remaining 22 per cent is composed of primary or secondary school pupils (58 thousand) and university students (about 33 thousand). These groups probably represent most of the under-coverage in the PR: that is to say, a very critical population in the official population counts. The cases with “weak not-steady” signals (Group C4) count 266,763 individuals: although the signals have a total duration of less than the minimum period requested for enrolment on the official PR, these cases would require further analysis, because seasonal or intermittent patterns could be related to the specific type of labor or education activity and not to residence characteristics. Group C2 and C3 represent two specific clusters of individuals: each of them correspond to a specific register and the related temporal pattern is the attendance of at least one academic course. It is important to emphasize that about 90 per cent of the individuals in these groups are foreigners, not Italian people: Group C3 represents foreign children attending Italian schools without being recorded in the official population register; Group C2 represents foreign university students that have not been recorded in the population register yet. Group C1, with strong-steady signals emerging from the labor registers, could correspond to more than one specific cluster of individuals, because their signals originate from many different registers related to a specific job type; moreover, it includes a wide range of country of citizenship and age (from 18 years). The geographical distribution of people in Group C1 according to citizenship shows slight levels of Italians in the municipalities along the Northern borders, around the capital city of Rome and the Republic of San Marino (Figure 3)

Figure 3

Subpopulation C1: people not recorded in the population register (PR) with signals from other sources

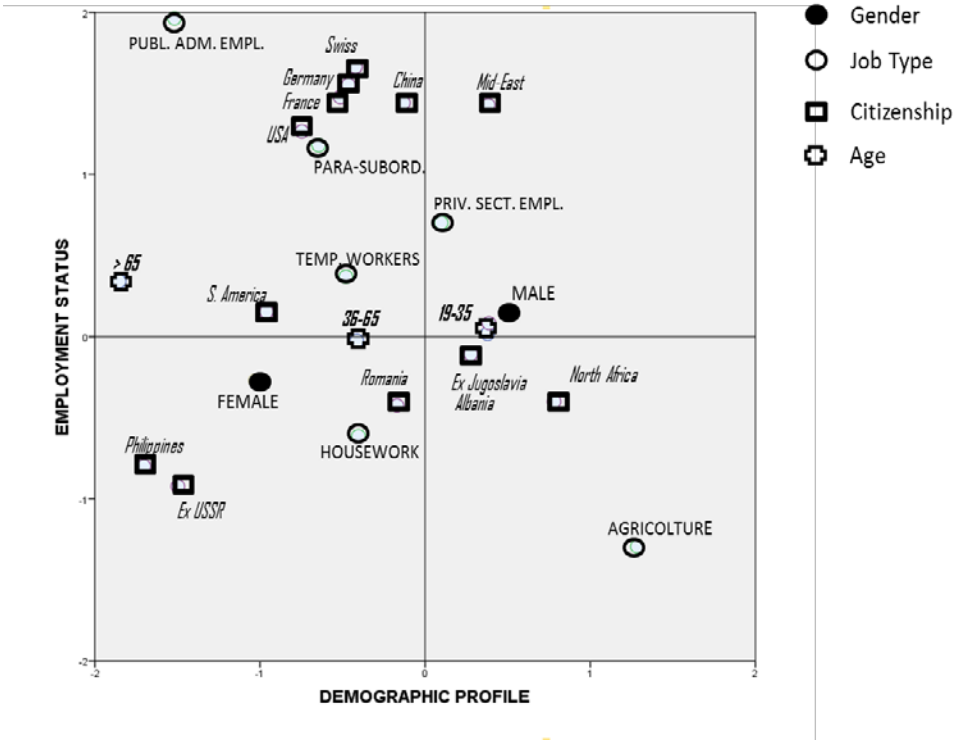


23. Foreigners are well represented in the Trentino-Alto Adige region and in the municipalities in the agricultural areas of the Po Valley, Puglia, Campania, Sicily and Sardinia.

24. Multiple Correspondence Analysis is used to find associations between categories of available variables that can offer useful insights into specific clusters. The variables considered in this step are gender, age, citizenship, province and type of labor. Two important discriminant factors emerge (Figure 4): the horizontal axis in the Correspondences Plot could be interpreted as the demographic profile, with age increasing from right to left and with the left quarters characterized by women's presence and the right ones by men's; the vertical axis seems more related to the employment status and the labor sector: the housework and agriculture sectors are at the bottom of the graph, while other workers in the public or private sectors are plotted in the upper area. The plot suggests the existence of different clusters: 1) young men working in agriculture, mainly from northern Africa and partly from Albania and former Yugoslavia, whose signals are located in Southern Italian provinces where agriculture is the main labor sector; 2) another group is mainly composed of women employed as caregivers or housekeepers, aged over 35 years, natives of the Philippines, Eastern Europe (ex-USSR) or South America, with signals located in the provinces where big towns are present; 3) a group consisting of individuals from Western and Central European countries, who are employees, self-employed, consultants or employed by temporary employment agencies, with the presence of people aged over 65 years and with signals located in the Central and Northern provinces of Italy; 4) Chinese people are represented in the same quadrant as Western European people; 5) Romanian people do not show gender or age specificity, but are associated with agriculture and housekeeping labor sectors; 6) Italian, French and Swiss people in this group are associated with provinces in the North, on the border of Italy, suggesting that they are probably border workers, usually living abroad but working in Italy and therefore not recorded in the PR even though they show strong signals in labor registers.

Figure 4

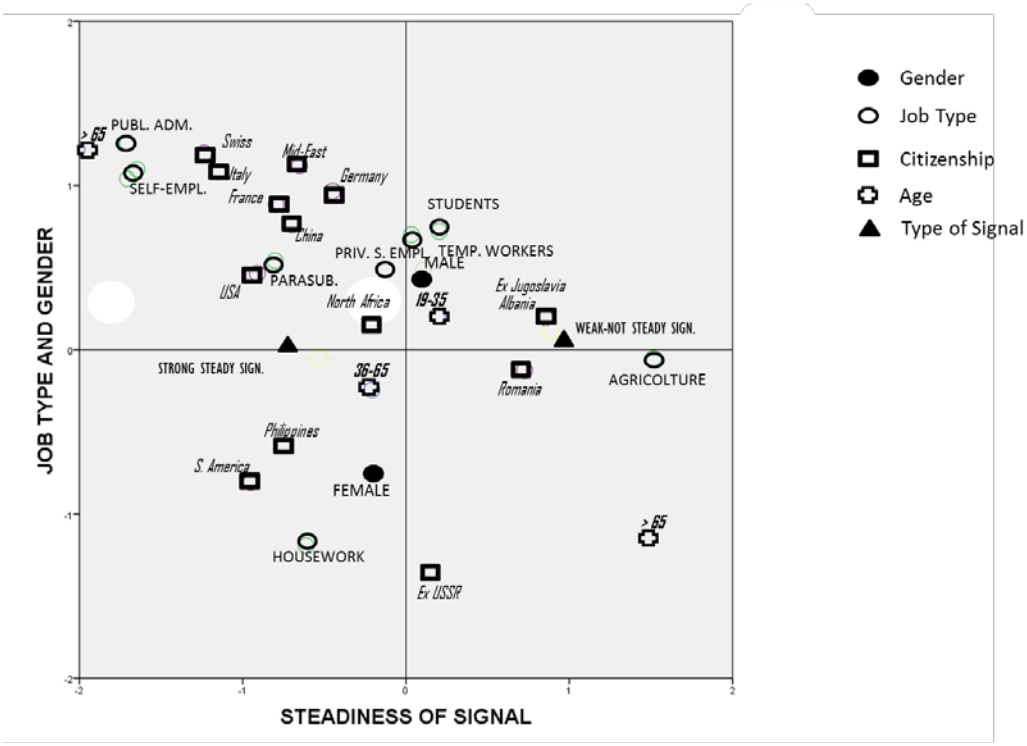
Subpopulation C1: Foreign workers with strong-steady signals and their association between the categories



25. The same kind of analysis has been carried out on Group C4, who are people with "weak not-steady signals": the output shows the same factors and the same associations between categories as Group C1 (Figure 5). The analysis of both Groups C1 and C4, with strong and weak signals, at the same time, seems appropriate. Among the relevant factors, besides demographic profile and employment status, steadiness of signal emerges as a new significant dimension characterized by the contrast between the strong-steady signals in one direction and the weak, not steady signals in the opposite direction. Clusters of individuals are much the same as mentioned above, but it is possible to identify a strong association between not-steady signals and specific type of labor and citizenship: weak, not steady or intermittent patterns are strongly related to agriculture and to young people coming from Romania, Albania and other countries of the former Yugoslavia.

Figure 5

Clusters of strong steady signals and weak not-steady ones: association between categories



26. Group D counts about 350,000 individuals and it was identified only by the signals derived from the Permits to Stay source (Table 1). Therefore, the “uncertain residents” in Group D are all foreigners from non-EU countries and not registered in the PR. Managing the information contained in this source, the duration of the permit to stay can be used in the monthly scheme which was processed for the signals contained in the other administrative sources. This transformation enabled the identification of foreigners who had the requirements to stay in Italy, on 31 December 2012, for at least twelve months. Finally, about of 61.5 per cent of individuals in Group D could be eligible as usual residents even though they have no labor or education activity signals, or other information to prove their presence on the Italian territory.

IV. Conclusion

27. The use of Knowledge Discovery from Databases proved very effective in exploiting the wealth of information contained in the administrative sources. However, further support for ISTAT's decision-making process might be to integrate the Knowledge Discovery from Databases process with the use of predictive models. The selection of specific administrative sources from all those available and the identification of their hierarchical order are essential elements: labor and education registers rank higher than the other sources examined, particularly in terms of the amount of more detailed information that they provide on the territorial level and duration of activity.

28. Signals of presence on the territory are extracted from the sources and can be used effectively to improve the quality of population registers. However, as the signals can also

represent a temporary or occasional presence, it is necessary to carry out a characterization process by constructing derived variables which then make it possible to identify cases of permanent presence that correspond to the usual residence definition in the international regulations.

29. The results of the trial show that the non-continuous signals of presence are specific to certain labor sectors (in particular, seasonal, agricultural work). Therefore, the classification of signals must be carried out on the basis of more accurate and multidimensional criteria, where the weight of the labor sector does not cause any distortion of the usually-resident population estimate. "Indirect" signals originating from the relationships between individuals can be derived from administrative sources. These are useful for evaluating the presence on the territory of individuals who show no direct signals, as they do not work or study and are not recipients of other income (for example, dependent family members).

30. In the trial, ISTAT defined a preliminary workflow to integrate the use of administrative sources and the official population registers in order to calculate the usually-resident population. Using this workflow, it is possible to define a group of individuals eligible to be included in the usually-resident population of Italy at a given reference date. This group of possible usual residents totalled 62.6 million individuals in 2012 (on December 31) and 62.4 million in 2013. From this group of individuals, it is possible to identify three main sub-groups:

a) The subpopulation present in the population register, without signals from other sources (3.0 million individuals in 2012);

b) The subpopulation present in the population register that showed signals in other sources (58.1 million individuals in 2012);

c) The subpopulation not present in the population register but with signals from other sources (1.5 million individuals in 2012).

31. Demographic variables, especially those of gender, age and country of citizenship as well as the location of the signal on the territory have proved to be very significant variables for defining specific sub-population profiles. A more detailed analysis of the characteristics of sub-populations at risk of over- and under-coverage has highlighted some important elements and specific clusters of individuals. The division into sub-groups and clusters is of great interest: on the one hand it allows the identification of subpopulations requiring further thematic analysis (for example, the typical foreign communities that elude the registers, but perform specific labor activities, or people who frequent certain territories); on the other hand, these same groups can form the basis for defining a census strategy formulated on the use of mixed techniques that combine specific surveys and appropriate statistical models.

32. Regarding the 3 million individuals who represent the potential over-coverage, three out of four have Italian citizenship, and the foreigners are on average more than six years younger than the Italians. This sub-population mainly represents people of working age (15-64 years). Apart from housewives, because of its geographical distribution, this sub-population is related to the geographical areas where unemployment is higher such as the municipalities in the South and in some central areas of the country.

33. With reference to under-coverage, the analysis shows the presence of several specific clusters. The continuous signals showing stable presence on the territory are related to just over 400 thousand people, which probably represents under-coverage in the population register: these are mainly foreign nationals. Geographical location and specific citizenship are essential for identifying the cross-border workers, for whom it is admissible to be absent from the population register. The analysis of weak (not continuous) signals has

highlighted the fact that some of them may still be associated with individuals with stable presence on the territory, and for this reason a better characterization of the direct signals would seem appropriate.

34. The analysis of signal strength based on its continuity over time is only the starting point in the use of longitudinal data that can be processed with administrative sources. The main objective of the future trials should be, therefore, the study of longitudinal models, over several years, to produce subpopulation estimates which are more stable in relation to the fluctuations linked to the labor market, from which signals are derived.

References

Argüeso A., Vega, J.L. (2014). A population census based on registers and a "10% survey" methodological challenges and conclusions. Statistical Journal of the IAOS. 30(1): 35-39

Bonifazi C., Martini C. (2014). The Impact of the Economic Crisis on Foreigners in the Italian Labour Market. Journal of Ethnic and Migration Studies. 40(3).

Cibella N., Gallo G., Pezone A., Tuoto T. (2015). The integration between the 2011 census Post Enumeration Survey Data and Administrative Data. The Analysis on Hard-To-Count Population. Paper presented to the Population Days Conference. Palermo: 4-6 February.

Citro, Constance F. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology. 40(2): 137-161

Crescenzi, F.,Sindoni, G. (2015). The Combined Use of Multiple Data Sources in the Population Census. Paper preseted to the Unece Group of Experts on population and Housing Censuses. Geneva: 30 September – 2 October.

Di Bella G., 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.

EMN European Migration Network (2012). Practical responses to irregular migration: the Italian case. Edited by EMN National Contact Point, Idos, Rome, 2012. See: http://ec.europa.eu/dgs/home-affairs/index_en.htm

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/MIT Press: Cambridge

Gallo G., Paluzzi E., Benassi F. (2014). The 2011 Italian experience towards supported-census for measuring migration. Paper preseted to the Unece Wok Session on Migration Statistics. Chişinău Republic of Moldova: 10-12 September.

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.

Jensen, P. (1983). Towards a register based statistical system- some Danish experience. Statistical Journal. 1(3): 341-365

Lanzieri, G. (2013). On a New Population Definition for Statistical Purposes. Paper presented to the Fifteenth Meeting of Group of Experts on Population and Housing Censuses. Geneva: 30 September – 3 October 2013.

ONS Office for National Statistics (2003). Census Strategic Development Review— Alternatives to a Census. Review of International Approaches. Information Paper. London: United Kingdom. Office for National Statistics

Poulain M., Herm A. (2013). Le register de population centralisé, source de statistiques démographiques en Europe. Population. 68(2): 215-247.

Statistics Canada (2011). Preliminary Report on Methodology Options for the 2016 Census. http://www12.statcan.gc.ca/strat/Preliminary%20Report%20on%20Methodology%20Options%20for%20the%202016%20Census.pdf (accessed April 17, 2016).

Statistics Denmark (1995). Statistics on persons in Denmark – a register-based statistical system. Eurostat: Luxembourg.

Statistics Finland (2004). Use of registers and administrative data sources for statistical purposes– best practices in Statistics Finland. Handbook 45. Statistics Finland: Helsinki.

UNECE (2007). Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics. United Nations Publication, ISBN 978-92-1-116963-8.

UNECE (2011). Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices. United Nations Publication

UNECE (2013). Population Definitions at the 2010 Censuses Round in the Countries of the UNECE Region. Paper preseted to theFifteenth Meeting of Group of Experts on Population and Housing Censuses. Geneva: 30 September – 3 October 2013.

Wallgren A., Wallgren B. (2011). To understand the Possibilities of Administrative Data you must change your Statistical Paradigm! Proceedings of the Survey Research Methods Section. American Statistical Association, Invited Papers.

Warners A.M., King R., William A.M. and Patterson G. (1999). The well-being of British expatriates retirees in southern Europe. Ageing and Society. 19(6): 717-740

Zhang, Li-Chun (2012). Topics of statistical theory for register-based statistics and data Integration. Statistica Neerlandica. 66, (1), 41-63.