The use of the Administrative Data in the production of Italian Permanent Population Census Estimates

Nicoletta Cibella (cibella@istat.it), Antonella Bernardini (anbernar@istat.it), Angela Chieppa (chieppa@istat.it)

Abstract

The Permanent Census of Population and Housing is the Census strategy adopted in Italy from 2018 aimed at integrating administrative data and data coming from sample surveys to obtain the counts of the usual resident population at a reference date, cleaning Census counts of possible coverage errors affecting the registers. The register at the core of the Permanent Census is the Population Base Register (PBR), whose main administrative sources are the Local Population Registers. The Census counts are determined by correcting the PBR with coefficients based on the coverage errors estimated with surveys data but the need for additional administrative sources clearly emerged while processing the data collected with the first rounds of surveys. A thematic register has been set up to exploit all the additional administrative sources useful for the estimation of Census counts; actually linking official population registers to subject-specific administrative sources (Labour and Education registers, Tax Returns register, Earnings, Retired) could help identify groups corresponding to the definition of *usually-resident population*. These records could be considered as *signs of life* of individuals on a specific territory and at a referenced time and are useful to detect possible coverage errors of the PBR, so improving the final population counts estimation also when surveys do not accomplish quality requirements or, like in 2020, surveys data are not available (COVID emergency). In the paper, the richness of administrative sources is exploited to correct the possible under coverage of the surveys in the wave 2018 and 2019 and explore relevant patterns useful for the estimation of population counts, improving the quality of population registers and some possible lacks of the sample surveys. Some recent experimentations on classification of *signs of life* according to duration, type, reliability of source and association among records (households) to detect PBR errors are also presented.

Keywords:
Population Census; Administrative data; Statistical Registers; Coverage estimation

1. The characteristics of the Italian Population and Housing Permanent Census

Administrative data are crucial, when producing official statistics, to comply with cost constraints and timeliness in dissemination, both issues especially relevant for Censuses (ONS 2016). Traditionally in Italy the censuses heavily relied on field-collection and required ISTAT (Italian National Institute of Statistics) to interact with local authorities. The survey covered all the units of the target population, occurring at the same time and every 10 years. The statistical not-sampling errors of a census were calculated at the end of the survey.

The 2011 Census was still a conventional census (i.e. deriving census information from an exhaustive field-collection) though being a register-assisted census, but then, in 2016, ISTAT adopted a modernization programme involving a statistical production based on an *Integrated System of Statistical Registers* (ISSR), combining administrative and survey data. According to this framework, a completely new Population Census strategy has been designed called *Permanent Population and Housing Census* (PPHC). The goal of the PPHC, adopted in 2018, is to produce annual data instead of the previous decennial cycle; the integration of administrative data and sample survey ones is key to obtain the annual counts of the usual resident population as well as other thematic Census outputs. The register at the core of the Permanent Census is the *Population Base Register* (PBR), whose main administrative sources are the Local Population Registers. Since the PBR contains under and over-coverage of residents and, moreover, to produce thematic Census outputs there is a need for variables not included in any ISTAT registers, two specific surveys, an Areal survey and a List survey, have been designed to support and enrich, in terms of coverage and quality, the available registers' data.

Both surveys, with the same questionnaire, are conducted annually in self-representatives municipalities (i.e. with more than 17,800 inhabitants), and once in 4 years, according to a rotation scheme, in non-self-representatives ones, so that at the end of the first cycle (2018-2021) all Italian municipalities will be sampled at least once. Out of a total 7,914 municipalities, about 2,850 are surveyed every year, for a total of about 1,500,000 households (of which 450,000 for the Areal survey and 950,000 for the List survey).

The Areal survey is based on a sample of addresses and/or enumeration areas drawn from the *Base Register of Territorial Units and Addresses*; every usually resident household living in the address/territorial unit sampled has to be enumerated (interview by CAPI). That is to say, households eligible to Areal survey are detected independently from the PBR: this is an important condition for estimation by means of capture-recapture model (see next paragraph).

The List survey, instead, is a sample of households drawn directly from the PBR. Interviews are conducted with a mixed mode technique (CAWI, CAPI, CATI), with a first phase of so-called “spontaneous response”, and a second phase of field follow-up of non-respondents.

2. The role of administrative data for the estimation of the population counts in the first cycle of the Permanent Census

Population counts, i.e. total amounts of usual resident population for each Municipality, are the primary Census outputs. In the framework of the PPHC, these counts are based on Population Base Register (PBR) data, corrected by means of survey results and, possibly, integration with other administrative data.

The PBR may be affected by errors of:

- over-coverage: inclusion in the register of individuals who are no longer on the territory
- under-coverage: non-inclusion in the register of individuals who are on the territory.

The estimated coverage errors determine the *correction factors* that have to be used as weights for each individual record in the PBR so to obtain the population counts by municipality adjusted for over- and under-coverage of the register.

In a traditional census, a Post Enumeration Survey (PES) is often used to measure the census undercount by means of the capture-recapture model (Wolter 1986 and Pfeffermann, 2015), with the PES being the *second capture* while the census itself is the *first capture*. In the case of the PPHC, the same estimation model is used to estimate the quality of the PBR: the 'first capture' is the presence in the PBR, while the annual sample surveys and administrative data represent the 'second capture'. Furthermore, for the PPHC the 'second capture' could be used to evaluate and correct both under-coverage and over-coverage of the PBR. For first waves 2018-2019, this estimation model was adopted for direct estimation of coverage errors in sampled municipalities; then indirect estimation with Small Area Models (Fay-Herriot, 1979) was used to enhance the quality of direct estimates for sampled municipalities and to calculate estimates for non-sampled municipalities.

In the original theoretical design for the estimation of coverage errors, the 'second capture' was planned to be derived mainly from the Areal survey data, but during the processing steps of the first wave 2018, this design was partly modified due to fieldwork quality issues and also administrative data were used to support the direct estimation of PBR coverage errors. The raw rate of PBR under-coverage has been computed as the ratio between the newly enumerated individuals (i.e. individuals not expected according to PBR) and the total number of individuals enumerated by the Areal survey. For measuring the over-coverage error of PBR, the List survey has been used together with data coming from a thematic register based on administrative sources, called *AIDA*,*Integrated Archive of Usual Resident Population*.

The AIDA register has been set up in 2015 to exploit all the administrative sources and to find relevant patterns useful for population counts estimation. For this purpose, administrative data are classified according to duration patterns, reliability of the original source and the association with other individual records (e.g. household relations) with the aim of predicting whether a presence in the PBR corresponds to a usual residence. These derived data are called *Administrative Signs of Life (SoL)* of a specific individual.

AIDA signs of life and surveys data are both a source for improving the estimation of coverage of PBR data: SoL from AIDA were used to adjust raw number of units for the over coverage estimation, taking into account the possible survey under coverage. All untraceable people (sample units from PBR not respondent to survey) associated with strong SoL in AIDA were recovered and considered usually resident.

The raw estimate of the over coverage rate for each municipality is given by the ratio of the number of sample units in PBR classified as untraceable to the total number of sample units (eligible). The raw estimate of the over-coverage rate was 6.10 (at national level) without using the signs of life and it becomes equal to 2.73 when using both survey and AIDA.

Small area models improve raw estimation of coverage errors and produce correction factors for each municipality and citizenship: the correction factors are weights attached to each individual in PBR. Therefore, if PBR, for a given domains, is affected by neither over-coverage nor under-coverage errors (or if the two errors compensate each other), the weight applied to residents in the register will be equal to 1: each individual represents him/herself and therefore will be counted as 1 in order to get the population count at the municipality level. If the estimated under-coverage of PBR is greater than the estimated over-coverage for a specific domain (municipality/citizenship), the corrector applied to each individual of PBR will be higher than 1. Consequently, the total population will result higher than that of PBR, i.e. the weight to be applied to each individual record in PBR will be inflated of a given quantity in order to obtain the correct population count at the municipality level.

The first two waves of the PPHC have highlighted what are the strengths and areas for further study in producing estimates of population counts, adjusted for under- and over-coverage. Istat produced the first prototypes of registers in 2018 and in the same year the annual surveys specifically designed for Census purposes were started up with very high response rates.

The Permanent Census permits to disseminate results, respecting cost constraints and timeliness in dissemination but even minimal variations in response rates and in List survey outcomes, as a result of very small sample numbers, produce very substantial impacts on the population counts, both in absolute and relative terms. At the same time, some gaps in the overall strategy have emerged as the complexity and the strong dependencies from the survey network and the need to leverage administrative records to reduce costs and the non-ignorable survey no response. Moreover the need, given the sample sizes and the sampling variability, to apply the correctors at the municipal level only with reference to the total population, separately for Italian and foreign citizenship is not sufficient to satisfy the request of more detailed data and, also, local authorities need data that are immediately usable (e.g. no need to round data like in case of probabilities) and to ‘delete’ records with 0 as correction factor from the PBR.

Census results, also, in Italy still has legal value (determining funding levels, availability of services): in the vast majority of cases the differences between register of population and census population estimates are of a few dozen units that lead to exceeding or being below the thresholds of interest to have or to lose funding.

3. The use of the administrative data in the 2020

A deterministic approach to produce census counts on the basis of only SoL is the shortcut used for the production of census counts in absence of surveys for the exceptional circumstances of year 2020, due to COVID Pandemic. Deterministic rules are a way to rapidly setting up a new estimation frame making the most efficient use of available administrative data and of expert knowledge. Results on calculating PBR coverage for year 2020 using deterministic rules based on SoL are currently under validation: these results will also be extremely relevant for designing the post-2021 Census round, for which there will be a serious reduction in census budget.

Associations between administrative signals, PBR data and, when available, survey data from first two waves are crucial to define patterns of individuals and criteria to compute aggregates of under- and over-coverage in the register. To extract knowledge from these integrated data is extremely important, considering the complexity of the phenomenon and that a massive use of administrative data for population statistics is an innovative solution. The overall objective of this investigation is to find the best classification of signs of life and presence in PBR and predict usual residence in Italy at individual level.

All data available have been processed and loaded in a database useful for investigation objectives (Chieppa et al 2018 and Bernardini, Cibella et al 2019) so to:

- pre-processing survey data to eliminate biased data, considering also quality indicators deriving from field-operations;

- loading more than 50 different archives from various administrative sources, managing time reference of SoL with proper time-window on duration of signal: different period are tested, from a longitudinal window on more years to at least 12 months; duration of a signal over time is needed to determine continuity patterns and ‘strength’ of signal)

- detecting patterns associated with high uncertain prediction for the usual place of residence (critical profiles).

When looking at the joint distribution of PBR presence and administrative signs, figures shows that steady/continuous signs of work or study coming from AIDA are the 52% of total PBR; strong signals of retired people are about 23% of PBR records. But there are people in PBR with no associated signals that have to be investigated to identify which part correspond to overcoverage of PBR: the percentage on total of individual in PBR is 2% for Italians, but more than 5% for foreigners. Moreover, there are about 500 thousands of individuals that have at least an administrative signal over the period considered (2018-2019) and that are not recorded in the PBR: these are the records that have to be investigated to detect which part is really undercoverage of the Population Register.

If we consider response rates of first two waves of surveys to better evaluate administrative profiles, multidimensional analysis results show that: administrative data on work/study have to be considered together with family ties and relationships among household, in order define signs of life useful to predict usual residence; citizenship is a discriminant factor in identifying patterns, together with the demographic size of the municipality where signals are located. Usually in largest towns response rate is lower for every profiles (workers, students...) except than for retired persons (always above 90%); moreover in the surveys some profiles are very hard to count (e.g foreign workers living alone in some suburban areas) while in administrative data capture less of the movement of Italians on the national territory.

After the phase of exploratory analysis, of which some results have just been reported above, the steps of the estimation process with deterministic criteria for producing 2020 census counts have been:

- building new variables (individual, family, travel routes, typology of municipalities) based on expert knowledge and exploratory analysis;
- formalization of the criteria by means of a grid on which experts could define deterministic rules for the classification of each individual records on the basis of the variables identified in the previous points - this formalization is also very useful in view of the engineering of the process;
- application of deterministic rules and classification of individuals coming from AIDA and/or BPR in one of three possible classes: under-coverage of PBR, over-coverage of PBR, correctly registered;
- correction of PBR based on individual classification (previous point) for over and under coverage and production of population counts;
- analysis of the results and comparison with the disseminated data in the waves 2018 and 2019.

The results of the estimation through deterministic criteria for 2020 counts are relevant also to design the future cycle of the Census, ensuring an optimal trade-off between costs and quality of estimates under the assumption of investigating a reduced number of sample municipalities and maximizing the administrative information available.

4. Current investigations and future perspectives.

The Permanent Population and Housing Census in Italy was designed for integrating registers, administrative data and sample surveys to obtain the counts of the usual resident population at a reference date, including the estimation of the coverage errors. The Covid pandemic and the subsequent complete changes in our way of living and working represent a new challenge to be faced by those who must disseminate official statistics figures and different scenarios for improving the quality of population counts with budget constraints and for overcoming some weaknesses of the actual strategy need to be produced in which administrative data are certainly an asset.

The case of the deterministic rules used for the estimation for the 2020 counts show the usefulness of mixed solutions based on classifiers that also take into account criteria entered by thematic experts.

The Census is based on a perspective of continuous improvement (design and implementation) of the overall process. The thematic experts, the statistical and the IT experts are working in a synergistic way to improve estimation processes used for the first waves of the PPHC. Alternative methodological solutions for the estimation of the population counts are currently being evaluated, to make the most efficient use of administrative data of the AIDA register and of the data collected from census surveys.

Data gathered for the years 2018-2019 are a valuable source to test and evaluate different predictive models (e.g. supervised and unsupervised models classification trees, latent class models) for usual residence and coverage errors in PBR. Moreover, the overall statistical framework for the estimation of quality of counts is currently investigated, through the availability of multiple lists that enumerate the population of interest (Zhang et al 2017 and Zhang 2019).

Current analysis on the integrated database with all available data are useful for training, testing and evaluating alternative models for estimation and detection of relevant variables/patterns of individuals. The results on patterns of critical subpopulations (with uncertain prediction) could have a strong impact on a new design for surveys data, that could be used to fill the gaps emerging in the registers and in the administrative sources.

Best solution implies not only methodological aspects and model assumptions, but also surveys costs and feasibility issues as well as expertise in designing and managing databases workflows: an interdisciplinary approach is needed to develop an improved estimation process for population counts. Moreover, decisions on this improved solution have to be shared with local authorities and continuous exchanges with them and data users contributed to improve the quality of the process.

References

1. Bernardini A., Cibella N., Gallo G., & Al. (2019), “Empirical evidence for population counting: the combined use of administrative sources and survey data”. Paper presented at ESS Workshop on the use of administrative data and social statistics, Valencia, June 4th 2019 , https://ec.europa.eu/eurostat/cros/system/files/gerardo-gallo_empirical-evidence-population-counting_istat.pdf
2. Chieppa A., Gallo G., Tomeo V. & Al. (2018), “Knowledge discovery for inferring the usually resident population from administrative registers”. In Mathematical Population Studies. International Journal of Mathematical Demography, Published online: 27 Jul 2018. http://www.tandfonline.com/loi/gmps20
3. Kim J.K., Rao J.N.K. (2012), Combining data from two independent surveys: a model assisted approach, Biometrika, Vol. 99(1), 85-100.
4. ONS (2016). Annual assessment of ONS’s progress towards an Administrative Data Census post-2021, downloadable at https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments.
5. Pfeffermann D. (2015). METHODOLOGICAL ISSUES AND CHALLENGES IN THE PRODUCTION OF OFFICIAL STATISTICS. Journal of Survey Statistics and Methodology 3, 425–483.
6. Wolter K.M(1986),Some Coverage Error Models for Census Data, Journal of the American Statistical Association,81,338-346.
7. Zhang, L., & Dunne, J. (2017). Trimmed dual system estimation. In D. Bohning, P. G. M. van der Heijden, & J. Bunge (Eds.), Capture Recapture Methods for the Social and Medical Sciences (pp. 239-259). (Chapman & Hall/CRC Interdisciplinary Statistics). CRC Press.
8. Zhang, Li-Chun (2019) A note on dual system population size estimator. Journal of Official Statistics, 35 (1), 279-283.
9. UNECE (2018), Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses.
10. UNECE (2020), New frontiers for censuses beyond 2020, https://unece.org/statistics/publications/new-frontiers-censuses-beyond-2020
11. UNECE(2021) KEEPING COUNT: CONDUCTING CENSUSES DURING THE COVID-19 PANDEMIC, https://unece.org/sites/default/files/2021-01/KeepingCount_0.pdf