United Nations ECE/CES/GE.41/2022/4

Economic and Social Council

Distr.: General
23 June 2022

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-fourth Meeting
Geneva, 21−23 September 2022
Item 5 of the provisional agenda
Transitions in census methodology; plans, experiences and innovations

Evolution of the Italian Permanent Population Census: lessons learned from the first cycle and the design of the Permanent Census beyond 2021

Note by Italian National Institute of Statistics*

Summary

The Permanent Population and Housing Census (PPHC) has been designed according to the modernization programme of the Italian National Institute of Statistics (Istat), which places the integrated system of statistical registers at the core of statistical production. The role of field surveys in this system is to support registers, in the broad sense of assessing their quality and to add information that is missing, incomplete or of insufficient quality. This allows the yearly availability of detailed census statistics.

At the core of the PPHC is the population base register (RBI), whose main sources are the local population registers of Italian municipalities. Two sample surveys (Areal survey and List survey) are conducted annually to evaluate and correct the coverage errors of the RBI and collect the data needed to produce census outputs.

The use of administrative data has been further accelerated because of the pandemic and the subsequent cancelation of the field surveys for the 2020 wave. In order to predict population counts at the municipal level by age, sex and citizenship, a process integrating available data from the past waves and administrative "signs of life" was set up to establish deterministic criteria applied to individual records in RBI. This obligatory push towards use of more administrative data has called for a rethinking of the statistical framework for the quality assessment of the estimation processes of the PPHC and, more generally, of the PPHC

* Prepared by Antonella Bernardini, Angela Chieppa, Nicoletta Cibella, Gerardo Gallo, Fabrizio Solari, Donatella Zindato.

Note: The designations employed in this document do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

design, using survey data to assess the quality of a fully register-based population count estimation.

I. Introduction

1. The Permanent Population and Housing Census (PPHC) has been designed according to the modernization programme of the Italian National Institute of Statistics (Istat), which places the integrated system of statistical registers at the core of statistical production. The role of field surveys in this system is to support registers, in the broad sense of assessing their quality and to add information that is missing, incomplete or of insufficient quality. This allows the yearly availability of detailed census statistics.

2. During the first two waves (2018−2019), due to fieldwork quality issues, the theoretical design was partly modified to include the use of the List survey (originally not intended for estimating coverage errors) to estimate over-coverage errors and the use of administrative 'signs of life' (SoL) to correct the List survey under-coverage.

3. Furthermore, due to the Covid-19 pandemic and the subsequent withdrawal of the sample surveys included in the design of the PPHC, in 2020 Istat opted for the exclusive use of SoL to estimate the coverage errors of the population register.

4. This obligatory push towards the use of more administrative data has called for a rethinking of the statistical framework for the quality assessment of the estimation processes of the PPHC and, more generally, of the PPHC design, using survey data to assess the quality of a fully register-based population count estimation.

II. The first design of the Permanent Population Census

5. At the core of the PPHC is the Population Register (hereafter RBI, the acronym in Italian), whose main sources are the local population registers of Italian municipalities. Together with the statistical base register of addresses (RSBL) and with the thematic registers on education and employment, it provides the basis to produce population census data. Two sample surveys (Areal survey and List survey) are conducted annually to evaluate and correct the coverage errors of RBI and to collect data for variables that are not available (or only partially available) from the registers.

6. The capture-recapture model was adopted for direct estimates of the coverage errors of RBI: the presence in the population register is the 'first capture' while field data represent the 'second capture'. Differing from a typical PES aimed at measuring under-coverage, in the PPHC design, the second 'capture' aims at measuring and correcting for both under-coverage and over-coverage of RBI.

7. More precisely, the Areal survey was used to estimate the number of individuals usually resident in the municipality who are not included in RBI (i.e. RBI under-coverage) while the List survey was used to estimate the number of individuals included in the register who are no longer usually resident in the municipality (i.e. RBI over-coverage). Furthermore, SoL derived from the Integrated Data Base of Usual Residents (hereafter AIDA) and classified according to duration patterns, type and reliability of the source were integrated in the estimation process to correct the survey under-coverage. Individuals in RBI who had not been enumerated were thus considered usually resident if associated with strong (i.e. of at least 8 months) SoL.

8. Indirect estimation with Small Area Models was then used both to enhance the quality of direct estimates for sampled municipalities and to calculate estimates for non-sampled municipalities. As a result of this process, the population count was finally obtained by applying correction coefficients for under-coverage and over-coverage errors to individuals in RBI (see Gallo and Zindato, 2018).

9. Besides the already mentioned fieldwork issues, other problems related to the design of the first cycle concerned the sample size, not allowing the estimation of the correction weights at the desired disaggregation level, and under-coverage of the Areal survey due to the insufficient quality of the addresses sampling frame.

III. The use of administrative data for the 2020 usually resident population count

10. The use of administrative data within the PPHC design has been further accelerated because of the pandemic and the subsequent cancelation of the field surveys for the 2020 wave. In order to produce the 2020 population count, Istat opted for the exclusive use of SoL to estimate the coverage errors of the population register.

11. This has been achieved through the use of classification criteria applied to statistical registers. More precisely, a process integrating available data from the past waves and SoL was set up to establish deterministic criteria applied to individual records in statistical registers to estimate the population usually resident in Italy. These criteria have been established based on expert knowledge and on a structured and iterative process, in which part of the variables are defined continually as the processing of administrative signals and the data analysis proceed (Chieppa et al, 2018).

12. For the construction of AIDA, the sources relevant to the identification of the usually resident population are selected and ordered hierarchically within the Integrated Microdata System (SIM) that, under compliance with the provisions of the law on confidentiality, collects information from administrative sources to support statistical production processes, both for social and economic statistics. The AIDA integration process involves the processing of data from more than forty administrative archives, each containing basic information on individuals' SoL (events) and covering several years. Among the main sources are the Thematic Registers of Occupation and Education, the tax returns and social security archives, as well as the real estate register. For each administrative event, the information on the location of the event (by means of the province and municipality codes) is also recorded.

13. The AIDA database used for the 2020 population count integrates data at the micro level from the 1st of January 2019 to the 31st December 2020. The longitudinal observation of direct signals over two years allows us to capture specific profiles of presence of individuals in the territory. These profiles of presence in Italy in some cases clearly identify usual residents in Italy, while in others the SoL are of low intensity, or identify seasonal workers (in both cases profiles that cannot be associated with the usual resident definition). Work and study signals, as well as home leases or social welfare benefits from the National Security System, are classifiable as direct SoL with respect to being usual resident in Italy; these records offer considerable detail, i.e. the duration of the activity, its location (municipality and address) and some specific attributes (employment contract, school/course attended, etc.) relevant in assessing the strength of the sign of life (Istat, 2021). Examples of indirect SoL are instead income tax return records (tax declarations, tax return filings, etc.) as well as owning a car according to the Cars Public Register or owning a property according to the cadastral archive, which all provide indirect signals of usual residence in Italy.

14. After processing the direct signs of life and the determination of the municipality where the study or work activity has been carried out, the next step of the process integrates the individuals with direct SoL with RBI. More precisely, this step identifies for each municipality on one hand all individuals with usual residence in Italy (i.e. with direct SoL) who are not recorded in RBI and, on the other hand, individuals recorded in RBI without direct SoL.

15. The next step consists of comparing the indirect SoL (derived by the Tax Registry) related to "dependent family members" and owners of a car or real estate unit with the individuals recorded in RBI who were found to have no SoL in the previous steps.

16. Finally, the last step identifies individuals with neither direct nor indirect SoL, i.e., the over-coverage of RBI. For this population sub-group, however, a further check is performed in order to identify "spouses" of reference persons who have direct SoL. These

spouses of individuals with direct SoL, who would otherwise end up in the set of individuals classified as RBI over-coverage because they lack direct or indirect SoL, are instead considered usual residents.

17. Table 1 shows the comparison between individuals with SoL according to the AIDA archive and individuals registered as usual residents in RBI. AIDA identifies almost 62 million individuals with administrative SoL but of these only 59.2 million can be considered as usually resident in Italy. On the other hand, the population correctly registered in RBI amounts to 58.7 million. The national under-coverage of RBI is almost 325,000 individuals, while the over-coverage of population registers is just over 1 million.

Table 1
Population counts and total population at 31 December 2020 as a result of integration between RBI and AIDA

Description of outcomes	*Type of register or archive*	*Total population counts*	*Population census counts*
Population correctly placed in RBI	RBI and AIDA	58,713,660	Yes
Under-coverage at national level	Only in AIDA	324,932	Yes
Uncertain units	Only in RBI	197,621	Yes
Over-coverage at national level	Only in RBI	1,005,908	No
Uncertain units	Only in AIDA	288,211	No
Population not included in the count	Only in AIDA with not useful SoL	1,410,497	No
2020 Census population	AIDA + RBI	59,236,231	Yes
Total population	AIDA + RBI	61,961,252	

Source: Istat, 2022

18. Furthermore, as shown in the table, in RBI there are almost 200,000 individuals whose usual residence status cannot be clearly established based on SoL (i.e. with weak or no SoL) and for whom a conservative choice was made, i.e. they were included in the final population census count. On the other hand, those with uncertain SoL (over 1.4 million people with weak or not well-localized administrative signs) who were not usual residents according to the RBI have not been considered in the census count. Finally, it has to be mentioned that the misplacement error of the population register (individuals registered in a municipality who are usually resident in a different one) has not been evaluated so far.

19. As in 2018 and 2019, the purpose of the 2020 count was to produce population size estimates by correcting the coverage errors of RBI. This goal was achieved by identifying individuals recorded in RBI as usual residents but not found in the other administrative sources (RBI over-coverage), on one hand, and individuals found in the administrative data as usually resident but not recorded in RBI (RBI under-coverage) on the other hand. This process could also be described as the identification of over-covered individuals in the integrated database RBI-AIDA on the basis of SoL.

20. This was indeed a significant innovation, ensuring the correspondence between the census count and the individual records, distinct from 2018 and 2019, when coverage-error correction was achieved by applying weights to RBI records (Istat, 2020; 2021).

21. 2020 represents a starting point for a more intensive use of administrative data for the population size estimation. Istat is working to improve the use of SoL in the new cycle (post-2021) of the PPHC. The acquisition of new sources (e.g. utilities archives such as energy consumption - smart meters data) that will most likely provide objective assessment elements for the actual place of usual residence will represent a turning point (Albert and Rajagopal, 2013).

22. In the second round of the PPHC starting in 2022, the surveys will retain a crucial role to collect data for variables non-replaceable (or only partially replaceable) by administrative data (List survey) and to provide quality measures of the fully register-based population size estimation. A formalization of the population size estimation process fully based on

administrative data is under study, including the choice of quality measures and the definition of the survey (audit survey) supporting the new estimation process.

VI. The Permanent Population Census beyond 2021

23. For the 2021 wave of the PPHC the main goal is to produce population counts fully based on SoL derived from administrative data. New techniques are being investigated that use survey data to improve SoL profiles (latent class models and other data science methods) related to populations groups not easily identifiable by deterministic criteria alone. At the same time, survey data will be used to produce estimates of the error of such population counts.

24. For the second round of the PPHC, which will start in 2022, the proposed architecture is based on the definition of an Extended Population Register (EPR) resulting from an integration process involving the Population Register (RBI) and administrative archives containing SoL. In Italy, the EPR will come into effect as an integration process between RBI and AIDA.

25. It is assumed that the EPR can only be affected by over-coverage (i.e., it is not affected by under-coverage). SoL profiles are then defined to identify subpopulations whose individuals are supposed to have a similar over-coverage behaviour, and each individual in the EPR is assigned to only one specific SoL profile h, h=1, …, H. SoL profiles can be defined using either information provided by experts or evidence resulting from statistical models, or a combination of both.

26. As of 2023, a further goal is to produce a list of over-covered individuals in the EPR. To this end, a SoL profile-based indicator function is defined (Bernardini, Cibella and Solari, 2022), according to which all the individuals in each profile are classified to be either included or excluded from the population count. This choice results in biased register-based estimates and it can be seen as a dichotomization approach of fractional counting proposed by Zhang (2019) where instead over-coverage profile probabilities in [0,1] are predicted.

27. Furthermore, in addition to the over-coverage error, an EPR can be affected by misplacement with respect to the local areas, i.e., individuals can be correctly included in the EPR but assigned to the wrong location area. In order to assess the misplacement error, an enhanced version of the EPR should be used, in which records refer to pairs ‘individual-address’ instead of only individuals (see Zhang, 2021c and Bernardini, Cibella and Solari, 2022). This would prevent adding bias selection components in survey estimates and, therefore, in the quality measures associated with register-based population size estimates (see Bernardini, Cibella and Solari, 2022).

28. Quality measures related to population size estimation could be assessed by comparing register-based estimates and survey estimates (for details see Zhang, 2022) or following the evaluation coverage approach proposed by Zhang (2021b) in which confidence interval coverage is chosen as an accuracy measure. This approach implies using surveys to assess quality measures instead of producing estimates and it is known as audit survey approach. As for the PPHC, an audit survey is planned to be conducted starting from 2023.

29. Following the audit survey approach, survey inefficiencies such those encountered in 2018 and 2019 affect only the estimation of quality measures and not the estimation of the population size.

A. Definition of SoL Profiles and of a SoL Profile Based Indicator Function

30. As mentioned above, SoL profiles of the EPR are defined to identify subpopulations whose individuals are supposed to have similar over-coverage behaviour. To this aim, a crucial point is the choice of the methodology for the definition of the SoL profiles. Classification rules can be derived by a machine learning approach (Michalski, 1983) or by using propensity score methods (see e.g. Stuart, 2010). Another possible approach is to

explore the information in administrative archives according to a Knowledge Discovery from Databases process (see Chieppa et al, 2019).

31. Statistical models can be used to predict the over-coverage rates θ_h for each profile h. To this aim, the best options seem to be trimmed log-linear and latent class models.

32. Instead of setting the SoL based indicator function in each SoL profile to be a constant value equal to 0 or 1, starting from some specified predicted probabilities θ_h , random values 0 or 1 can be generated. This would allow (model) unbiased register-based estimates.

33. In order to have a stable list of over-covered individuals over time, instead of generating independent random numbers every year, permanent random numbers can be assigned to each individual.

B. The main features of the Audit Survey

34. As previously mentioned, an audit survey will be conducted in 2023 to provide quality measures of the register-based population size estimation.

35. The most appropriate sampling design for the audit survey consists in sampling individuals from the EPR or pairs of individuals and addresses from the enhanced version of the EPR. The audit survey sampling approach allows a cost reduction with respect to areal surveys.

36. Furthermore, the overall audit survey sample size can be smaller than that of a traditional frame-based survey aiming at estimating the population size.

37. For some sets of individuals, there might be incomplete or unreliable information about their potential usual residence address. In this case, alternative sampling techniques need to be used. Possible solutions are area sampling (if those individuals are supposed to be concentrated in specific portions of territory), or indirect sampling, (for instance, by selecting their work or study SoL addresses to try to collect useful information on their true address).

38. Unlike traditional surveys aiming to estimate one or more population parameters, the definition of the sample allocation for an audit survey is not simple and intuitive. Two proposals based on power allocation defined by Bankier (1988) are presented in Bernardini, Cibella and Solari (2022). The basic concept is that SoL profiles h for which the over-coverage rate θ_h is not close to 0, or 1, should be partitioned into K_h profiles hk, k = 1, . . . , K_h , for which the corresponding values over-coverage rate θ_{hk} is closer to 0, or 1, than θ_h. In order to achieve this goal, additional auxiliary information is needed, and a possible solution to collect the required information is oversampling critical profiles.

39. Finally, a small-scale areal survey could be performed in order to evaluate possible under-coverage of the EPR. An alternative way of measuring under-coverage consists in introducing into the audit survey a reverse record check component (i.e., crosscheck for records related to non-respondents). Though integrating the different surveys could pose organizational issues and fieldwork problems, from a bare statistical point of view they can be linked by using graph sampling (see Zhang, 2021a).

V. Conclusions

40. To replace the decennial census, in 2018 Istat launched the PPHC, based on the integration of administrative data with information collected from two sample surveys (Areal survey and List survey) conducted annually in self-representative municipalities and every four years, according to a rotation scheme, in non-self-representative municipalities (Falorsi, 2017). According to this design, register data are supplemented by field data.

41. The use of administrative data has been further accelerated because of the pandemic and the subsequent cancelation of the field surveys for the 2020 wave.

42. This obligatory push towards a larger use of administrative data called for a rethinking of both the statistical framework for the quality assessment and the overall estimation

processes of the PPHC design, using survey data for the quality measurement of a fully register-based population count estimation.

43. To this aim, the processing of 2021 Census data, currently ongoing, will be of great importance, given the availability of both survey and administrative. Comparisons among different estimation models, the integration of administrative and survey data, and the evaluation of fieldwork quality are all important areas of investigation to improve the design of the future PPHC cycles.

References

Albert, A. & Rajagopal, R. (2013), Smart Meter Drive Segmentation: What Your Consumption Says About You. IEEE Transactions on power systems: 4019−4030.

Bankier, M. D. (1988), Power allocations: determining sample sizes for subnational areas, The American Statistician, 42:3, 174–177.

Bernardini A., Cibella N. & Solari F. (2022). A Statistical Framework for Register Based Population Size Estimation, Technical Report, Istat Advisory Committee on Statistical Methods 2022 Springtime Meeting, Roma, 19−20 May 2022.

Chieppa A., Gallo G., Tomeo V., Borrelli F. & Di Domenico S. (2019). Knowledge discovery for inferring the usually resident population from administrative registers, Mathematical Population Studies Mathematical Demography, 26:2, 92−116. https://doi.org/10.1080/08898480.2017.1418114.

Falorsi, S. (2017), The Italian experience on the Population and Housing Census: the Master Sample, UNECE Meeting, October 4−6 (2017), https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/Day2_1130_Italy_falorsi_presentation.ppt__1_.pdf.

Gallo, G. & Zindato, D. (2018). Annex H. Italy case study, in UNECE, Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses, Geneva, https://unece.org/guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0.

Gallo, G. & Zindato, D. (2021). Italy: The combined use of survey and register data for the Italian Permanent Population Census count in UNECE, Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses (endorsed by the 69th plenary session of the Conference of European Statisticians), https://unece.org/statistics/publications/CensusAdminQuality.

Istat (2020). Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età cittadinanza, grado di istruzione e condizione professionale per gli anni 2018 e 2019, https://www.istat.it/it/files/2020/12/NOTA-TECNICA-CENSIPOP.pdf.

Istat (2020). Nota tecnica sulla produzione dei dati del Censimento Permanente: la popolazione residente per genere, età, cittadinanza e grado di istruzione al 31.12.2020, https://www.istat.it/it/files//2021/12/NOTA-TECNICA-CENSIMENTO-POPOLAZIONE_2020.pdf.

Zhang, L.-C. (2019). On provision of UK neighbourhood population statistics beyond 2021, Report for ONS, https://arxiv.org/pdf/2111.03100.pdf.

Zhang, L.-C. (2021a). Graph Sampling, Chapman and Hall/CRC.

Zhang, L.-C. (2021b). Proxy expenditure weights for Consumer Price Index: audit sampling inference for big-data statistics. Journal of the Royal Statistical Society, Series A, 184:2, 571–588.

Zhang, L.-C. (2021c). Discussion at ISI session ‘Population statistics using administrative data instead of census’, Virtual 63rd ISI World Statistics Congress, 11–16 July 2021.

Zhang, L.-C. (2022b). Complementarities of survey and population registers, in N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels (eds.), Wiley StatsRef: Statistics Reference.