Economic and Social Council

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Seventeenth Meeting
Geneva, 30 September to 2 October 2015
Item 3 of the provisional agenda
Innovations planned for 2020 census round, and results of tests

The combined use of multiple data sources in the population census

Note by the National Institute of Statistics of Italy

Summary

The new strategy for the Italian population and housing census implies a greater use of available administrative and statistical sources with a limited use of sample surveys, in line with the objective of the modernization plan of the National Institute of Statistics of Italy (Istat) to move towards an integrated system of registers. Lighter but continuous field work is expected to produce on-going improvement and gains in experience. Positive are also the expected effects on financing; the demand of public financial resources would be reduced and diluted over time. The constant production of data would allow for much more effective and approachable liaisons with users too.

This paper discusses some of the most important methodological, technological, and organizational requirements to get these results. Istat suggests to make the cooperation among countries stronger on methods concerning the use of multiple sources with special reference to: i) models making integrated use of data from multiple sources; ii) checks of coverage of population registers to produce statistical counts; iii) pooling data from available surveys; iv) measuring quality of data obtained by integration of multiple sources.

Please recycle

I. Introduction

1.   Combining available data sources, including surveys, census and administrative registers is becoming a necessity more than a choice for the next round of censuses. A decennial census of population fully based on field enumeration is a too complex and no more sustainable operation and stakeholders such as politicians, administrators and researchers ask for regularly updated and harmonized demographic, social and economic data, georeferenced at the highest level of geographic detail. The high costs of traditional censuses and their operative burden push towards a more effective use of administrative data sources.

2.   If multimodality in data collection characterized the last round of censuses, the use of many administrative and statistical sources will most likely characterize the next one and on these issues effective cooperation among countries is to be set.

3.   Individuals continuously leave traces in administrative registers, where potentially there is an enormous amount of geo-referenced data available for statistical use. These data, however, may be not updated and are affected by coverage errors, due to unharmonized classifications and definitions, which might compromise their usability.

4.   In the revised version of the UN Principles and Recommendation the alternative methodologies to conduct a census are represented in a matrix where the rows describe data collection through field enumeration and the columns represent use of administrative and/or population registers as data sources. The matrix columns contain many types of registers; administrative registers, created and used mainly for administrative purposes outside the national statistical authorities, statistical registers and base registers.

5.   In the new Italian census, the *"censimento permanente",* the challenge is to move towards a more intensive use of administrative and statistical sources with a limited additional use of ad hoc sample surveys in order to produce census outputs every year. The aim is to move towards a new kind of census which uses statistical methods and new technologies to produce census outputs every year in a sequence of well-engineered integration operations starting from available administrative and statistical sources. We are thoroughly exploring methods which generalize the use of a single source and the conceptual issues underlying the simultaneous use of many sources.

II. The innovations of the multimode data collection in the 2011 census.

6.   The 2011 census was approached in a completely different way from the previous ones. Administrative population registers ("Anagrafi") were used for mailing out questionnaires to families. Respondents could choose the way in which they preferred to complete and return the questionnaire in a mixed-mode return approach:

- online, using the password provided with the questionnaire;

- at any post office in Italy;

- at one of the municipality census collection centers, where specialized assistance was available;

- directly to a municipality enumerator.

7.   Sampling techniques were used for the measurement of some of the variables of interest. This was carried out through the use of two versions of the questionnaire, a short

form, for two thirds of the families in the larger municipalities, and a full form for the remaining families.

8. A web-based system was developed to manage multimode response and support interviewer recovery of nonresponses. The system was designed to automate back-office work and enabled the status of every individual questionnaire to be updated almost in real time. It also permitted production of census progress reports, allocation of census areas to the enumerators and monitoring of their work, targeted recovery of non-respondents and unregistered individuals, comparison between the census and municipal population records, and production of relative accounts of found and not found persons. The availability of constantly updated information on the status of each questionnaire enabled enumerators to be directed only to households to which the questionnaire had been sent but not yet returned and in this way to systematically recover non-response.

9. In the largest municipalities, an additional support list was based on the pre-census Address Number Survey, containing information on potentially inhabited housing units for which there was no corresponding family in the municipality records. The use of other auxiliary lists available from central and local administrative sources, containing information on the presence of individuals not registered in the municipal records, enabled their targeted and systematic recovery.

10. To produce signals of people not enrolled in the municipality records and in order to make spatial information available at the unit level, either natural or legal person, administrative or statistical sources were integrated in an Integrated System of Micro data (SIM). Data from many sources such as business, tax, education, employment and other relevant registers were linked by individual unique codes. One of the main sources used during this process was the list of permits to stay for non-communitarian citizens. Among the topics included in the SIM are: household characteristics, place of usual residence, location of place of work, location of place of study (school, college or university), status in employment, educational characteristics, dwellings and housing arrangements, etc. In order to manage potential under-coverage, data from sources such as the revenue agency or the foreigners' permit-to-stay archives were used to set up a list of persons not included in the municipal population registers but potentially residing in each municipality.

11. Despite the big innovations of the 2011 census, a stable and enduring balance between census costs and benefits was not achieved yet. In fact, costs remain high and too concentrated in time, while the use of administrative data as they are is still not suitable to satisfy the demand of census-like data. Moreover, Census data becomes quickly outdated and the supply of highly detailed geographic data remains only decennial. For these reasons the development of a different approach is necessary.

III. Beyond 2011

12. Summarizing the experiences of the previous population and housing census round, we are now planning to use new methodologies for producing census data. The crucial principle of providing detailed statistics at the lowest geographical level remains of utmost importance. The use of registers – primarily population registers - in combination with other sources is being considered for the purpose of producing detailed small areas statistics on population and housing, as well as the application of continuous surveys methodology for the same purpose.

13. Techniques, methods and organizational solutions implemented for the 2011 round are now reconsidered and combined with new ones in an innovative framework which makes them consistent with more advanced strategic goals.

14.   There are many reasons for exploring such an alternative approach. Among them: i) the need to produce more frequent and timely statistics; ii) the budget limitation for the census; iii) the reluctance of the population to participate in the census; iv) the increased technical capacities to handle data sources.

15.   Information on individuals and households are collected from available administrative or statistical sources, namely, different kinds of registers, of which the following are of primary importance: individuals, households and dwellings and existing survey data linked at the individual level. Ad-hoc sample surveys are used to provide information on census topics not available from administrative sources or to adjust poor quality register data.

16.   Statistical methods will be employed to achieve the two main objectives of the population census, which are:

- counting the usual residents producing key data on demographic structure of population and households (C objective);

- producing socio economic census data (D objective).

17.   In the C objective, the yearly estimates of population counts broken down by demographic structure (sex, age, marital status, citizenship), the most serious issues to face over the next years will be the count of aging population and migrants.

18.   In the D objective, in order to meet national and international requirements to estimate socio economic data of households and individuals, the main aim is to collect through surveys only variables which need to be collected in compliance with UE Regulations and cannot be estimated combining information available in population registers, administrative sources and statistical surveys. A study concerning the substitutability with socio-economic data from administrative sources is currently carried out. The aim is to substitute data traditionally collected through the census with estimates obtained from available information: administrative sources and pooling of sample survey already existing. This requires strengthening the current studies on local and central administrative sources. We would meet user needs by providing more frequent updates, so avoiding the decline in accuracy over the decade. Most notably, individuals and households will benefit through statistical offices' use of available information that would otherwise need to be collected from them again through costly and redundant surveys. Administrative data and big data can help also in increasing the amount of small area data and allow the construction of new indicators for policies. Finally, the increased use of administrative data will enhance the public offices ability to build evidence on which to evaluate the effectiveness of their programs and policies. Efforts will be put into place to identify areas which will be more fruitful to cover with new information.

IV. Projects launched by Istat

19.   In the Modernization plan of Istat, the challenge for the Italian statistical system is to move towards an integrated system of statistics, in which the focus is on new statistical methods, new technologies which enable a more effective use of administrative and statistical data source to minimize the burden on respondents and costs. The pillar for such an integrated system covering the economic, environment, demographic and social domains are integrated and harmonized business and population base registers and frames. A coordinated system of samples of statistical units for the various surveys will be developed along with tailor-made questionnaires for target population groups. The unique identification code used for record integration in the SIM will also be used within the national statistical agency to facilitate the combination of data from different sources, the

comparison over time and to stimulate multi-use of the data. Statistical methods are employed to correct data. Data should be collected from persons, households or enterprises only if not obtainable from base registers, administrative and available statistical sources.

20.   A key role is given to the ANNCSU (Italian acronym for the national register of streets and addresses) with its system of geo-referencing of streets, and house numbers. Such a geographical data base is in fact essential to locate units in the information system. A crucial tool to improve geocoding of Data from Administrative Sources. Each house number will be geocoded to the census enumeration areas created by Istat.

21.   The pillars of the digital agenda of Italy are laid down in a late 2012 Law. Beside the "censimento permanente" and ANNCSU, the ANPR (national archive of the resident population) will be the unique national population register in which all the municipal registers will be transferred. To produce small area data from the population register the integration of these three pillars is crucial.

22.   In order to make administrative sources useful for statistical purposes it is essential to ensure their compliance with quality requirements through a strategy of continuous data quality control and correction. ARCOLAIO is a project launched in Istat aiming to control the quality of statistical use of administrative sources. We are thoroughly studying also methods that can be employed to handle these aspects.

23.   According to the European Regulation on population and housing censuses (European Parliament, 2008), the census target population is the set of usual residents, defined as those who have lived or intend to live for a period of more than 12 months in their place of usual residence. Since intentions are not registered in administrative sources, and the identification of residents who lived in the same place more than a year is quite difficult, this definition cannot be easily applied in the new approach. Moreover the "real" usual resident population may differ from the registered population because of undocumented migrations or fake registrations for fiscal convenience.

24.   ARCHIMEDE (ARCHIve of Micro data Economic and DEmo-social) is the innovative infrastructure that will produce micro data obtained by administrative sources and statistical surveys after their quality check and treatment. Data, geocoded to enumeration areas, will be released respectfully of confidentiality, in an effective and transparent way. The outputs of ARCHIMEDE will be defined strictly considering users' needs, to provide public administrations, researchers and users with general purpose collections of data referred to base-level statistical units and geographical units (down to census enumeration areas).

25.   From the methodological point of view, we are investigating the predictive capacity of models having usual residence as response variables and data from multiple administrative and statistical sources as independent variables. We are considering among them latent class models (Biemer Paul P., 2011, p. 258) and several generalization of capture – recapture methods to the use of multiple lists. New sources added in the system of integrated micro data such as lists of utility contracts from gas and electricity companies, and lists of users of local services can greatly increase the predictive capacity of these models.

26.   A pilot of the censimento permanente surveys started in March 2015 involving 150 Municipalities and 160.000 households. The aim is to test, beside the predictive capacity of the above mentioned statistical models, the organization of the surveys in some possible variants with a focus on the new technologies of data capturing and IT architectures, with special attention to:

- the IT architecture and the development of an integrated management system for multimode data collection, which allows to work using many different tools (tablets,

mobile, PC, etc.) with or without a network connection and regardless of hardware and operating system used;

- abolish paper questionnaires in the field survey data collection aiming to different organizational solutions, based on local patterns in returning census questionnaires and differences in ICT usage;

- check of streets names and house numbers in the ANNCSU register.

V. Conclusion

27. Up to now and including 2011, census have been undertaken in Italy every ten years. The new Italian strategy for population and housing census will join a greater use of available administrative and statistical sources with a limited use of sample surveys.

28. Continuous operations would bring significant improvement in fieldwork efficiency and many benefits in terms of quality. A local permanent fieldwork would allow expertise to be retained and developed over time; a lighter but continuous field work is expected to produce ongoing methodological improvement and gains in experience.

29. Positive are also the expected effects on financing; the demand of public financial resources would be diluted over time and continuous operations might make service contracts more attractive and possibly cheaper than in a "one shot" operation. The constant production of data would allow for much more effective and approachable liaisons with users too.

30. This paper discussed some of the most important methodological, technological, organizational requirements to get these results.

31. We suggest to make the cooperation among countries stronger on methods concerning the use of multiple sources with special reference to:

- Models making integrated use of data from multiple sources;

- Checks of coverage of population registers to produce statistical counts;

- Pooling data from available surveys;

- Measuring quality of data obtained by integration of multiple sources.

References

- Biemer Paul P., 2011 Latent class analysis of survey errors, Wiley Series in Survey Methodologies

- Baffour, B., J.J. Brown, P.W.F Smith, 2013, An investigation of triple system estimators in censuses. *Statistical Journal of the International Association for Official Statistics*, 29, 53-68

- Bakker, B.F.M., Rooijen, J. van & Toor, L. van (2014). The system of social statistical datasets of Statistics Netherlands: an integral approach to the production of register-based social statistics. Statistical journal of the United Nations ECE, 30(4), 411-424

- European Parliament, 2008, Regulation (EC) No 763/2008 of the European Parliament and of the council of 9 July 2008 on population and housing censuses, In: *Official Journal of the European Union*, 13.8.2008, pp. L 218/14-L 218/20

- Schulte Nordholt, E. (2005), The Dutch Virtual Census 2001: a new approach by combining different Sources, Statistical Journal of the United Nations Economic Commission for Europe 22, 25–37

- 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

- Wallgren, A. and Wallgren, B. (2007), Register-based statistics - administrative data for statistical purposes, John Wiley and Sons, Chichester

- Zhang, L-C. (2012), Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica

---