Changing census by combining administrative sources and sample surveys

Fabio Crescenzi*
National Institute of Statistics, Rome, Italy – fabio.crescenzi@istat.it

Abstract

Istat, the National Institute of Statistics of Italy, to overcome the main gaps of a decennial population census, a too big and no more sustainable survey with too many sunk costs, is moving to a *censimento permanente,* a new kind of census which, in order to produce census outputs every year, makes a more intensive use of administrative and statistical sources in a sequence of well-engineered operations and requires a limited additional effort in "ad hoc" sample surveys.
We are making in-depth studies on methods and new technologies that can be employed to get these goals and on the underlying conceptual issues. According to the European Regulation on population and housing censuses, the 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. Considering that intentions to live are not registered in administrative sources, and also that the identification of residents who lived longer than a year is quite difficult, this definition is tough to apply from existing sources. Moreover the actual usual resident population may differ from the registered population because of undocumented or not canceled migrations or improper registration for fiscal convenience.
From the methodological point of view, we are investigating on the predictive capacity of models having usual residence as response variables and data from multiple administrative and statistical sources as independent variables, and among them, latent class models and several generalization of capture – recapture methods to the use of multiple lists. To achieve the goals set to the *censimento permanente* is extremely important the increase of the availability of administrative data sources and a more extensive use of new technologies for data collection.

Keywords: Population census; Administrative data; Population counts; Small area data.

1. Introduction

Users require a more timely and frequent availability of census data. Actors of political and social life as politicians, administrators, researchers ask for regular updated and harmonized demographic, social and economic data, georeferenced at the highest level of spatial detail. The high costs of traditional censuses and their operative burden push toward a most effective use of administrative sources and to spread over years the census fieldwork.
Considering how often a person leaves traces in administrative information systems, there is an enormous potential amount of spatial data available in registers for statistical analyses. These data, however, may be not updated and affected by coverage errors, obtained by non-harmonized classifications and definitions, which might compromise their usability.
Up to and including 2001, the Italian population and housing census was conducted with the traditional "door-to-door" survey method, with the same economic, human and organizational resources thus allocated to every household.
Despite the big innovations of the 2011 census, a stable and enduring balance between census costs and benefits has not been achieved. In fact, costs remain high and too concentrated in

time, while the use of administrative data is anyway not suitable as they are to satisfy the potential demand. 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 seems to be necessary.

2. The innovations of the 2011 census round

The 2011 census was approached in a very different way from the previous censuses with significant changes in the survey methods and techniques and standardized solutions adopted in relation to municipality size.
Existing records of the municipality registers were used to create census mailing lists of households, after normalization and geocoding of each address. The questionnaires were customized with name, addresses and place of return after completion. The mailing out of questionnaires to the 25 million households was performed by the postal service.
Municipalities were mapped in terms of enumeration areas, localities and sub-municipalities census areas.
A unique archives of address numbers geocoded to enumeration area were built, containing information about the characteristics of each address number.
Sampling techniques were employed for the measurement of some of the variables through use of two versions of the questionnaire, the full form and a short form.
Households could choose the way in which they preferred to complete and return the questionnaire in a mixed-mode return of questionnaires:

- online, using the password provided with the questionnaire;
- at any post office in Italy;
- at one of the municipality census collection centers, at which specialist assistance for the completion of questionnaires was also available;
- directly to a municipality enumerator.

All phases of the census were managed through an advanced census management system, accessible online to all the field forces involved, with accesses authorized on the basis of locations and geographic areas. The system was designed to automate back-office work and enabled the real time check of every questionnaire. It also permitted production of allocation of areas to enumerators, census monitoring reports, targeted recovery of non-responders and unregistered individuals.
In order to manage potential under-coverage, data from different sources (such as the data from the revenue agency or the archives of foreigners' permit-to-stay) were used to set up a list of persons not included in the registers but potentially residing in each municipality.
In the largest municipalities, an additional list was based on the pre-census survey of the address numbers, containing information on potentially inhabited housing units for which there was no corresponding entry in the municipality records. The availability of constantly updated information on the status of each questionnaire and the use of the auxiliary lists from central and local administrative sources containing information on the presence of individuals not registered in the municipal records enabled their targeted recovery. To produce signals of people not enrolled in the municipality records and in order to make spatial information available at the unit level administrative and statistical sources were integrated in an Integrated System of Micro data (SIM). Data from different sources such as business, tax,

education, employment and other relevant registers were linked by individual unique codes. Among the topics included in the SIM are: household characteristics, place of usual residence (location of place of work, school, college or university), status in employment, educational characteristics, dwellings and housing arrangements, etc.
The availability in the census management system of a supplementary municipality list, containing data on new residents and changes of address enabled real-time crosschecking between people responding to the census and those registered in the municipality records on the date of reference, allowing the earlier release of the census results.

3. Beyond 2011

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.
Techniques, methods and organizational solutions implemented are now reconsidered and combined with new ones in an innovative framework which makes them consistent with more advanced strategic goals.
There are more than a few reasons for exploring such an alternative approach, and among them: i) the need to produce more frequent and timely statistics, ii) the budget limitation for census taking, iii) the reluctance of the population to participate in the census, iv) the increased technical capacities to handle data sources.
This change fits and it is part of the "Modernization plan of Istat" launched in late 2014. The traditional model based mainly on full field data collection is often burdening the respondents, and this reduce response rates and quality of data. This boost statistical offices to use statistical registers obtained from administrative sources and maintained continuously updated by the use of new technologies.
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 integrate and harmonized business and population base registers and frames. Coordinated system of samples of statistical units for the various surveys will be developed along with made to measure questionnaires for sections of the target population.
The integrated system will be organized to facilitates the multiple use of micro data for a variety of different statistical products. At several clearly defined stages in all the production processes, the micro data will be stored in data depositories for use in other processes with standardized meta data.
The unique identification code 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.
This means that all information, which can be derived from administrations and registers should be used and only additional information should be collected from persons, households or enterprises.

Archimede (ARCHIve of Micro data Economic and DEmo-social) is the innovative infrastructure that will make available 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 strongly considering users' needs to provide public administrations, researchers and users with general collections of data referred to integrated elementary-level statistical units and territorial units (up to census and enumeration areas).
A key role is given to the national register of streets and addresses (ANNCSU) which, with its system of geo-referencing of "streets", and "house numbers", represents a cornerstone of the system. Such a data base is essential in order to locate units in the information system, a crucial tool to improve geocoding of data from administrative sources. Each house number is geocoded to the areas of the census mapping of Istat.
The pillars of the digital agenda of Italy are laid down in a Law of late 2012. Beside the *censimento permanente* and ANNCSU, the nascent national register of residing population (ANPR) held by the Ministry of the interior will be the unique administrative register in which will be transferred all the existing municipal registers. To produce small area data from the population register the integration of the information of these three pillars and SIM is crucial.
In order to make administrative sources useful for statistical purposes it is essential to ensure their compliance with the quality requirements by a strategy of continuous data quality control and correction. Administrative data are obviously free from sampling errors, but are subject to non-sampling errors. It is therefore essential to the conceptualization and measurement of the statistical accuracy of the statistics produced from administrative sources, which allows to apply concepts such as bias and variance, which we are using when the data are produced by survey. Istat launched a project aiming to the quality of administrative sources in their statistical use (ARCOLAIO - engineered operations for quality assessment of data from multiple administrative sources). The basic idea of ARCOLAIO is to overcome the traditional quality assessments mainly related to inference from sample field data collection, building a new framework focused on quality assessments obtained by the employ of multiple sources.
Information on individuals and households are collected from existing administrative or statistical sources, namely, different kinds of registers, of which the following are of primary importance: individuals, households and dwellings and existing surveys 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 data which are of poor quality in registers.
Statistical methods will be employed to achieve the two main objectives of the population census, which are:

- counting usual residents population producing key figures on demographic structure of population and households (C objective);
- producing socio economic census data (D objective).

In the C objective the yearly estimates of population counts referred to 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.
In the D objective, in order to accomplish national and international requirements to estimate socio economic data of households and individuals, the main aim is to collect by survey only

core topics variables which need to be collected in compliance with U.E Regulations and are not possible to estimate using combining information available in population registers, administrative sources and statistical surveys. In order to identify the topics to collect, a study concerning socio economic data included in registers is carried out. The aim is to substitute data acquiring by survey with estimates obtained by available information. This requires strengthening the studies on local and central administrative sources already initiated for this census. We would meet user needs providing more frequent updates to remove the decline in accuracy over the decade. Most notably, individuals and households will benefit through statistical offices' use of existing information that would otherwise need to be collected from them again through costly and duplicative surveys. Administrative data and big data can help also in increasing the amount of small area data and allow the construction of new measures for policies. Finally, the increased use of administrative data will enhance the 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.
According to the European Regulation on population and housing censuses (European Parliament, 2008), the 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. Considering that intentions are not registered in administrative sources, and also that the identification of residents who lived longer than a year is quite difficult, this definition is tough to apply from existing sources. Moreover the actual usual resident population may differ from the registered population because of undocumented or not canceled migrations or improper registration for fiscal convenience. The intention is to build a conceptual map of the main differences with the goal to produce an estimate of each of them.
We are deepening, in addition to the underlying conceptual issues, also on methods that can be employed to handle these aspects. We are investigating on the predictive capacity of models having usual residence as response variables and data from multiple administrative and statistical sources as independent variables, and 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 from administrative sources as list of utilities from gas and electricity companies, and lists of users of local services can greatly increase the predictive capacity of these models.
A pilot survey of the *censimento permanente* 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 *censimento permanente* surveys in some possible variants with particular regard to the new technologies of data capturing and IT architectures, with special attention to:

- the IT architecture for multimode registers and survey data collection and processing;
- 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;
- develop a multi-modal approach based on the use of many different acquisition tools (tablets, mobile, PC, etc.), allowing data to be stored locally for subsequent upload to the server many areas in Italy not covered by data transmission networks for mobile devices;
- check of streets names and house numbers in the ANNCSU register.

4. Conclusion

Up to now and including 2011, census have been taken every ten years; The new Italian strategy for population and housing census will join a greater use of existing administrative and statistical sources with a limited use of sample surveys to get census data every year.
Continuous operations would bring significant growth of fieldwork efficiency and many benefits in terms of increased 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.
Positive are also the 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 "one shot" operation. The constant production of data would allow much more significant and approachable dealings with users too.
In this paper we discussed on some of the most important methodological, technological, organizational requirements to get these ambitious results. Conceptual and definitional aspects are of crucial importance, together with the evaluation of the predictive capacity of models having usual residence as response variables and data from multiple administrative and statistical sources as independent variables, considering among them latent class models and several generalization of capture – recapture methods to the use of multiple lists. The increase of the availability of new data sources and a more extensive use of new technologies for data collection are of paramount importance to achieve the new goals of the census.

References

Baffour, B., Brown J.J., Smith P.W.F, (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., (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.

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

Crescenzi F., (2010). Beyond the 2010 census round: plans for the 2020 round, Note by the National Institute of Statistics, Italy (ECE/CES/GE.41/2010/5).

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, Official Journal of the European Union, 13.8.2008.