The evolution of the spatial data production model in Istat. New perspectives for the analysis of population socio-economic phenomena

Giancarlo Carbonetti, Raffaele Ferrara, Gerardo Gallo, Mariangela Verrascina

Istat - National Institute of Statistics, Italy

Abstract

Over the last decades, official statistics have shown an increasing focus on the demand for more accurate and timely data with high spatial details. In this regard, the Italian National Statistical Institute (Istat) has undertaken a modernization process that has led to a significant revision of the production processes of official statistics on population. In particular, two important pillars of Istat new production system are outlined: the construction of statistical registers and the Integrated Register System (SIR), and the design and implementation of the Permanent Population and Housing Census (PPHC). The development of statistical methodologies and the adoption of advanced IT solutions have made it possible to systematically integrate administrative data with the results of sample surveys and to geocode information at a very fine level over the territory. This has made it possible to define an integrated and geo-referenced database for the territory, rich in annually renewable information on the main demographic, social, and economic variables related to individuals, household types, and certain characteristics of dwellings.

The new process is continuously evolving to meet ever higher quality standards, release data with high timeliness, and ensure a constant increase in the supply of information at the highest spatial level. These results offer new perspectives for the analysis of the main socio-economic phenomena of the population at the sub-municipal level, both spatially and temporally. It is possible to focus on specific topics through the definition and calculation of appropriate indicators. In this regard, Istat is conducting an experimental project aimed at measuring and representing the socio-economic deprivation of families at sub-municipal level. The possibility of identifying, within the municipal territory, the areas of greatest criticality, represents an extraordinary opportunity for policy makers to define intervention programmes to combat family deprivation in a targeted manner on the territory. Moreover, the possibility of replicating the analyses over time will also make it possible to assess the impact of the interventions implemented in the territory. Finally, Istat is planning to release the sub-municipal data according to a predefined schedule in a structural manner and is designing a user-friendly platform for the dissemination and representation of the results using GIS tools.

Keywords: population census, administrative sources, integration, geo-referencing, territory.

1. Introduction

This paper illustrates how Istat has renewed the population and housing census in Italy, moving from the traditional decennial approach to a combined census, based on the integration of sample data and administrative data, conducted every year since 2018. The new permanent population and housing census (PPHC), together with Istat statistical register system, is the response to the growing demand for more timely, accurate, and detailed data.

The transition to the new census has an important effect on the process of producing sub-municipal output. In the traditional census, observed data on statistical units were immediately geo-coded to the enumeration area (EA). The strategy of the permanent census is that the geo-coding of census information occurs ex-post to the survey, during the integration phase of the statistical registers used in the production process.

Then, after explaining PPHC strategy, the new data production model for the minimum area unit coinciding with the EAs will be presented. This model has led to a diversification of the supply of sub-municipal data and, consequently, of the possibilities for spatial studies. In this regard, a research project that Istat is conducting with the aim of studying the socio-economic deprivation of families at sub-municipal level will be presented as an example. The results of this project will enable municipalities to define more targeted and effective intervention policies on the territory and to follow their effects over time.

Finally, the current census data dissemination system and the new opportunities for the dissemination and valorisation of sub-municipal data for end users will be presented.

2. The need for highly detailed spatial data

The availability of data referable to the sub-municipal scale with high quality level is fundamental for conducting spatial analyses and studies for research purposes, for business objectives and to support decision-making processes on social, economic and environmental issues (Carbonetti et al., 2023a). The main objectives of analysis include: socio-economic transformations; mobility in the territory; urban expansion and transformations; social segregation phenomena (e.g. in schools); housing needs; energy needs for the definition of energy strategies.

The census has always represented the only survey opportunity capable of providing data with a high spatial level that cannot be obtained from any other survey. However, given the increasing specialisation of users - experts in data integration and the use of GIS tools - there is a strong expectation for data with greater timeliness, accuracy and richness of detail.

Official statistics have not remained indifferent to these needs. There are numerous examples of innovation in statistical processes that have concerned the data collection phase, the creation of statistical infrastructures, the development of new methodologies and IT solutions and the design of different ways of disseminating results. In particular, the integration of survey data with administrative data and the geocoding to the enumeration area will enrich the information base available to users for conducting spatial analyses at various levels and for different thematic areas, even according to high quality of data[1].

3. The Permanent Population and Housing Census in Italy

To replace the decennial census, in 2018 Istat launched the Permanent Population and Housing Census (PPHC): a combined approach which integrates administrative data and sample surveys (with yearly data collection and dissemination of data) (Falorsi, 2017; Gallo & Zindato, 2018). The transition became necessary due to the unsustainability of the organisational apparatus, the statistical burden on households and the reduction in available financial resources. In addition, the development of the new census strategy was facilitated by the wide availability of data from administrative sources and the opportunity to use the statistical registers built in Istat.

The core of the PPHC is the Basic Population Register (BRI). Together with the Basic Register of Places (BRP) and the thematic registers on education and employment, the BRI forms the basis for the production of population census data in a combined census design, with two *ad hoc* sample surveys (Area survey and List survey) conducted annually to collect data for non-substitutable (or only partially substitutable) variables and to collect data useful for the population count.

In the first cycle (2018-2021) of the PPHC, two surveys were conducted annually in self-representative municipalities (i.e. those with a population over 17,800 inhabitants and smaller ones which do not rotate in the sampling scheme of the Labour Force Survey) and every four years, according to a rotation scheme, in non-self-representative municipalities (i.e. all the others). In each municipality[2], a sample of households is selected from the Population Register for the List survey and a sample of addresses from the Address Register for the Area survey.

The adoption of a sampling strategy and integration with administrative data made it possible to reduce the burden on respondents and municipalities (who are responsible for conducting the fieldwork). Furthermore, a multi-mode data collection technique is used, totally paperless with the CAWI mode offered as first option, allows respondents large flexibility. In

[1] The quality of spatial data will be ensured by the use of the Basic Register of Places (see section 4), which will enable correct and detailed geographical localisation of statistical information units. This will result in the consistency of count data on individuals, households, dwellings and buildings. In addition, before dissemination, the data undergoes a careful validation phase with reference information from the previous census or other auxiliary sources.

[2] Every year were involved in the surveys about 2,850 municipalities for a total of about 1,400,000 households (of which 950,000 for the List survey and 450,000 for the Area survey). In 2021 the number of municipalities involved was higher (4,531 out of the total 7,904 municipalities in Italy) as the number of non-representative municipalities was double than the number originally planned (in 2020 the fields surveys were cancelled due to the pandemic therefore the municipalities due to participate in 2020 were 'moved' to 2021). Therefore the households involved in the 2021 surveys were respectively 2,472,400 for the L survey and 776,097 for the A survey. The reference population is the population usually resident in Italy.

addition to the CAWI option, respondents can use the municipal collection centre for assistance in filling out the electronic questionnaire.

Concerning the census outputs, the PPHC produces a: 1) fully register-based population count; 2) census hypercubes estimated by the joint use of information already available in registers and of data collected on the field.

With regard to the population count, in the PPHC original design the capture-recapture model was adopted for direct estimates of the coverage errors of the BRI, with the population register representing the ‘first capture’ and field data being the ‘second capture’. The population count was then obtained by applying correction coefficients for under-coverage and over-coverage errors to individuals in BRI.

In 2020, due to the COVID-19 pandemic the census supporting surveys were cancelled, therefore a fully register-based count was produced for the first time. Thanks to the availability of relevant information originated from administrative sources, the Signs of Life (SoL) method[3] was applied to SoL Archive of integrated administrative data in order to produce the municipal population counts. SoL profiles (individuals who are supposed to have a similar over/under-coverage behaviour) were defined based on experts’ knowledge. In 2021 this change was consolidated and improved, thanks to the availability of survey data and SoL, combining evidence resulting from a statistical model with expert knowledge. In this revised census design, survey data are used also to measure the error of the fully register-based count, instead of being used for correcting the BRI[4].

Regarding census hypercubes, some are the result of integration between sources, while others are the result of an estimation process through the application of specific statistical models (i.e., logistic, loglinear, multinomial), which integrate data from different sources with sample data. The application of these methods led to "building" the micro-data of each statistical unit (individual; household; dwelling) in a complete manner. The absolute frequency referring to any cross cell is obtained by summing the values at the micro-data level relative to the units that belong to that cross cell.

[3] The SoL method is applied to implement the usual residence definition according to European Regulation N. 763/2008. All the administrative sources implemented (organised in a smaller number of statistical registers) are being used to estimate the actual presence of a person at their registered address i.e. to correct the Population Register (ANPR according to the Italian acronym, which is the basis of the census count) through the use of classification criteria applied to individual records in statistical registers.

[4] Modification of the sample surveys to determine the population count error through life signs is also underway. Statistical models have been used to guide the deterministic criteria used on the basis of the SoL to carry out the population count. However, the measure of error of the population count thus obtained has not yet been officially produced.

4. The sub-municipal data production process

In the context of the new permanent Census, the process of producing population and housing data at sub-municipal level is based on the link between the Basic Population Register (BRI) and the Basic Register of Places (BRP). BRI is the statistical register of population, which contains the individuals who are annually identified by the Census as usual residents in Italy; BRP is the statistical register of the territory, which contains all the elements that can be linked to it: addresses, dwellings, buildings and the new 2021 territorial bases (enumeration areas and other territorial areas). The link between the two registers makes it possible to relate individuals and households to their dwellings and buildings, as well as to establish an univocal and consistent spatial geo-coding for all the statistical units of interest for the Census, with the consequent possibility of providing, even for very detailed territorial levels, single variables and cross-tabulation of considerable importance for census dissemination.

4.1 The linkage process between BRI and BRP

The linking process between BRI and BRP is rather complex (it is carried out by population groups and with different processing steps) and is mainly done by combining the addresses of the population with those of the dwellings and/or with information from the real estate register and/or rental registers (Carbonetti *et al.*, 2023b). Once the link between households and dwellings has been made, the geocoding processing phase of all statistical units begins:

- ➢ an enumeration area (EA) is attributed to the population through the geographical coordinates (x,y) of their addresses;
- ➢ an enumeration area is assigned to the buildings from cadastral sources through a spatial joining operation of them on the territory and in the new land bases;
- ➢ a comparison between two EAs is carried out and inconsistencies between the two geocodings are resolved.

More analytically, the following combinations occur:

1) same EA for addresses of individuals/households and buildings;
2) presence of only the EA associated with population addresses;
3) presence of only the EA associated with buildings;
4) different EA for addresses of individuals/households and buildings;
5) absence of EA on either side.

In cases falling under the first three points listed, geocoding is assigned automatically, whereas for the last two situations, geocoding must be decided. In the case of different EAs between population addresses and buildings (item 4), empirical analyses have shown that when the distance between the coordinate of the address and the centroid of the building is large, the correct geocoding is often the one associated with the building; conversely, the correct EA is most often the one associated with the population address. When, on the other hand, information is missing on both the individual/household and the building side (point 5), the EA is assigned either using the information associated with the individuals/households in the 2011 spatial bases (the centroid of the latter is re-processed on the new 2021 spatial bases), or randomly among the EAs of the municipality with empty dwellings.

4.2 The result validation phase

Analyses and checks are carried out on all population and housing data, aggregated by EA, in order to validate them. This operation is preliminary to the dissemination of the results.

The main checks concern EAs with anomalous numbers. For example: EAs of urban centres with no population and no dwellings; EAs of scattered houses with a lot of population and a lot of dwellings; 'special' EAs (i.e. EAs with particular portions of territory, such as squares, urban greenery, cemeteries, harbours, etc.); EAs with a large number of dwellings. Finally, special attention is paid to the localities (EA aggregations) of municipalities that have exceeded relevant population thresholds in the transition from one census to the next.

The publication is accompanied by quality metadata concerning both the geocoding process and the validation outcome.

5. Perspectives for spatial studies

The availability of census and administrative data each year offers new possibilities for conducting studies and territorial analyses. The linkage of administrative data to the census database makes it possible to construct specific indicators to study socio-economic phenomena as never before. In addition, the geo-coding of information to the enumeration areas and the annual replication of the available information base allows for comparisons on both spatial and temporal dimensions. This represents a new frontier for the study of the dynamics of socio-economic population phenomena.

5.1 A case study: the study on family deprivation

Istat is currently conducting, together with a number of municipalities, an experimental project to study family deprivation at sub-municipal level (Biasciucci *et al.*, 2023). The realisation of the project is based on the richness of information from censuses, administrative sources and statistical registers held by Istat. Once the concept of family deprivation[5] ('real deprivation' and not 'exposure to risk') was defined, an exploration of the sources present in Istat was carried out to identify the most appropriate ones - in terms of data availability and quality - useful to calculate specific elementary indicators (referring to households or individuals) at the highest territorial detail (EAs). These sources, in order to be used functionally for the project objectives, must have the following characteristics: adequate level of quality (accurate); reference to a well-defined time interval (timely); renewable from year to year (updatable); allow geo-reference to the enumeration areas (geo-codifiable).

Nine indicators related to different components of deprivation (economic, occupational, educational, housing) were defined and calculated (by EA) using data from both census and administrative sources. These indicators were then synthesised through a non-compensatory composite index (De Muro *et al.*, 2011; Mazziotta & Pareto, 2016; Mazziotta & Pareto, 2017) to produce an "Index of Family Deprivation" (IFD) calculated at EAs' level.

Specific spatial analyses are still being carried out to identify, within the municipality, critical areas of concentration of family deprivation. These areas will be drawn as "clusters of contiguous enumeration areas", through a specific algorithm, around the most critical EAs (with the highest values of IHD) according to statistical (homogeneity of internal areas) and geographical (shape and extension of areas) rules that are being defined.

The aim of the project is to utilise the information content of permanent census data and administrative data in order to conduct studies on the territory for socio-economic phenomena at a more detailed scale. The results, in terms of maps and indicators, will be a useful tool for municipal administrators for the planning and evaluation of local policies. Moreover, the possibility of replicating the study each year makes it possible to follow the dynamics of the phenomenon and to assess the effectiveness of the policies implemented in the territory. The dissemination of results will also make use of the technical solutions that will be adopted more generally for the dissemination of sub-municipal data (see 6.2).

6. New opportunities for dissemination

The implementation of the permanent census in Italy has offered new opportunities for the release of results, both in terms of data and dissemination tools. In the past, analyses of fine territorial detail were only possible every ten years, after the release of the results derived from

[5] Currently, the definition of family deprivation chosen is the following: *"a condition in which families and individuals experience difficulties in adequately meeting their basic needs due to a shortage or insufficiency of economic, employment, educational, social and housing resources and opportunities"*.

the census rounds. The analysis of the data available every ten years revealed variations over a very long time interval, not allowing to capture those recorded in the short term, now possible with the results of the permanent census. The annual publication in fact allows to identify and monitor subgroups of the population or portions of the territory in difficulty, critical situations, and greater vulnerability from year to year to promptly initiate local interventions on certain phenomena. With the new census strategy, a remodelling of the dissemination was also envisaged, with the possibility of associating the permanent census with new forms of output release. The need arises, on one hand, to define new products to be disseminated that can exploit the information potential of both the census and the other information resources available in the institute, with the aim of continuing to enrich the information assets of which Istat has makes available and, on the other hand, to introduce more technologically advanced tools.

6.1 The current system of output dissemination

The results of the annual census surveys are disseminated both on the Permanent Census Datawarehouse and on the new Data Browser platform[6] for browsing and visualising municipal Census data. The first channel represents tradition, maintaining the approach used for the 2011 Census data dissemination. The second channel represents innovation. Data Browser makes it easy to browse through the data and quickly select an area of interest. It is an area-oriented and user-friendly way of meeting knowledge needs and expanding the audience of census data users. A third tool that traditionally accompanies the dissemination of census results is cartographic representations. For several decades, Istat has been publishing geographical data from the system of territorial bases (1991, 2001, and 2011)[7] of a set of partitions and zoning, which compose the mosaic of the entire Italian territory.

Up to now, Istat has made available to users, on one hand, a tool (BT.Carto) that allows them to consult and export thematic maps referring to census indicators calculated at municipal, provincial and regional level, adapting them to their own reporting needs and, on the other hand, a WebGIS application (BT.Viewer) dedicated to the visualisation and consultation of geographic data of the territorial bases and census variables. The latter is connected to GISTAT, Istat Geographic Information System, through which the Institute geographically-based information assets are shared and made available to users. Currently, the tools made available are mainly used by expert customers.

[6] The Data Browser is a dedicated thematic platform for browsing and querying statistical data, based on cutting-edge architecture and technology, with high performance in terms of response time, a flexible layout and suitable for the dissemination of large amounts of data at municipal level.

[7] The partitions indicated refer to: enumeration areas, sub-municipal areas (boroughs, districts, etc.) and localities.

At a finer territorial level (enumeration areas and sub-municipal areas), to the traditional dissemination illustrated above, we wish to add a tool that takes advantage of some of the possibilities offered by the technological evolution that has taken place in recent years, both in terms of tools and computer security.

6.2 The dissemination of sub-municipal data

Istat started a reflection to enhance sub-municipal data dissemination and to make geographical tools more easily accessible to less skilled users. New forms and tools for disseminating sub-municipal data are therefore being studied, with the aim of visualising complex topics in a simple way and illustrating territorial and local differences at a glance.

Solutions that go beyond the traditional static display of maps (non-editable images) are being evaluated. It will be possible to download data in CSV format to be associated with shapefiles to allow users to obtain their own cartographic product (as is already the case for all the sub-municipal data of the population and housing census), but more technologically advanced tools based on the geographic information system (GIS) will also be introduced that will allow users to visualise and navigate geographic data to create interactive maps. With this tool, users will be able to: find a location; draw the area on a map; save the drawn area; export the outline of the area as an image or spreadsheet (CSV).

This system will make it possible to navigate and query the data, to move around the map, to make selections on the map, to consult the map table to be navigated and to customise area and data selections. The aim is to make it easier, even for non-expert users, to navigate the interactive thematic maps at a minimum territorial level, and to enhance the informative richness of the sub-municipal data that will be made available, progressively in greater quantity and quality than in the past, not only from the point of view of content (thanks to the integration of permanent census data and registers) but also from the point of view of historical series (thanks to the continuous, i.e. annual, survey).

Therefore, the new supply of highly detailed spatial information will create new opportunities for spatial analyses focusing on both spatial and temporal dimensions.

7. Conclusions

The work represents the response of the National Statistical Institute (Istat) to new challenges in the production of spatial data. There is a growing need for accurate, timely and geo-referenced data at small spatial domains on demographic, social and economic issues. The availability of such information is crucial to study the main phenomena characterising the dynamics of resident population and households at the highest territorial detail.

Methodological advances and the development of IT solutions have facilitated the integration of administrative data with sampling data, georeferencing down to the address of residence and the construction of the integrated system of statistical registers.

The results of the permanent population and housing census, data from the administrative sources and statistical registers of Istat represent a rich repository of information useful for producing spatial indicators with a high information content.

The new information base will allow spatial and temporal analyses to be carried out as never before, on various topics and with reference to very fine spatial levels. Finally, the use of GIS tools will enhance and facilitate the exploitation and dissemination of results.

References

Biasciucci, F., Carbonetti, G., Ferrara, R., Mazziotta, M., Quondamstefano, V. (2023). New scenarios for measuring household deprivation in sub-municipal areas. ASA Conference 2023: Statistics, Technology and Data Science for Economic and Social Development, Bologna, Italy, 6-8 September 2023. (Forthcoming in the *Book of Short Papers*. Firenze University Press).

Carbonetti, G., Ciccarese, A., Roncati, R. (2023a). An analysis of the demand for sub-municipal data from the Population and Housing Census. (Forthcoming on the *Review of Official Statistics*. Rome, Istat).

Carbonetti, G., De Matteis, G., Di Zio, M., Fardelli, D., Ferrara, R., Lipizzi, F. (2023b). Enumeration area imputation methods for producing sub-municipal data in the Italian permanent population and housing census. *Statistical Journal of the IAOS, vol. 39, no. 1*, 123-136. IOS Press. DOI: 10.3233/SJI-220113.

De Muro, P., Mazziotta, M., Pareto, A. (2011). Composite Indices of Development and Poverty: An Application to MDGs. *Social Indicators Research, 104*, 1-18.

Falorsi, S. (2017). Census and Social Surveys Integrated System. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting, Geneva, Switzerland, 4–6 October 2017.

Gallo, G., & Zindato, D. (2018). Italy case study. In Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses, UNECE New York and Geneva, 2018.

Istat, 2022 - *Popolazione residente e dinamica demografica. Anno 2021.*
https://www.istat.it/it/files//2022/12/CENSIMENTO-E-DINAMICA-DEMOGRAFICA-2021.pdf
https://www.istat.it/it/files//2022/12/Nota-metodologica-censipop-_2021.pdf

Istat, 2021 - *Popolazione residente e dinamica demografica. Anno 2020.*
https://www.istat.it/it/files/2021/12/CENSIMENTO-E-DINAMICA-DEMOGRAFICA-2020.pdf
https://www.istat.it/it/files/2021/12/NOTA-TECNICA-CENSIMENTO-POPOLAZIONE_2020.pdf

Istat, 2020 - *Il Censimento permanente della popolazione e delle abitazioni. Prima diffusione dei dati definitivi 2018 e 2019.*
https://www.istat.it/it/files/2020/12/REPORT_CENSIPOP_2020.pdf
https://www.istat.it/it/files/2020/12/NOTA-TECNICA-CENSIPOP.pdf

Mazziotta, M., & Pareto, A. (2017). Synthesis of indicators: the composite indicators approach. In: Complexity in Society: From Indicators Construction to their Synthesis. *Social Indicators Research Series*. Filomena Maggino Editors, Springer, 159-191.

Mazziotta, M., & Pareto, A. (2016). On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. *Social Indicators Research, 127*(3), 983-1003.