Statistical Journal of the IAOS 36 (2020) 143–149
DOI 10.3233/SJI-190586
IOS Press

The integration of geographic and territorial data sources into the base register of territorial and geographical entities

Fabio Crescenzi* and Fabio Lipizzi
ISTAT, Rome, Italy

Abstract. The National Institute of Statistics (Istat) is moving from a decennial census to a permanent census, which will produce census data every year. The georeferencing of statistical information will allow Istat, through the integrated system of registers, to release statistical and geographical information with a strong territorial detail and will allow users to know more on the structure and composition of the territory of Italy. The new Statistical Base Register of territorial and geographical entities of Istat (RSBL) is a multidimensional register integrating several components: addresses, regular grids and micro zones, buildings and housing units, administrative zones, statistical and functional zones. In the panorama of statistical registers, it represents an innovative way of considering in an integrated way all the most relevant components of territorial data. In this paper, we present some of the main features of the register.

Keywords: Register, territorial data, geocoding, permanent census, addresses, micro zones, buildings, administrative units, functional and statistical zones

1. Introduction

The National Institute of Statistics (Istat) is moving from a decennial census to a permanent census, which will produce census data every year. The georeferencing of statistical information will allow Istat, through the integrated system of registers, to release statistical and geographical information with a strong territorial detail and will allow users to know more on the structure and composition of the territory of Italy. Addresses are of fundamental importance for this purpose, in particular for the realization of the permanent census. For this reason, Istat has intensified the activities to improve the quality of addresses together with the activities to improve the quality of tools for their standardization and recognition.

In the modernization plan of Istat data from administrative sources and statistical surveys are integrated in a system of registers (SIR) covering the demographic, social, economic and environmental domains [1,3,4]. The Base registers are connected by codes and are maintained updated over time using administrative sources and statistical surveys [10,15]. A greater use of administrative and statistical data sources will allow minimizing the burden on respondents and costs of field data collection. RSBL (the Italian Statistical Base Register of Territorial Entities), is one of the pillars of SIR which integrates information from many different sources of geographical data [9], with the aim to improve the georeferencing of data. The aim is to build the register only once, to keep it updated, and to use it in all statistical processes.

In perspective, RSBL has to become the unique complete and integrated source of geographical data to georeference statistical data at different time. We consider this a truly significant innovation in data georeferencing: without this innovation, there could be redundancies in acquisition and processing of geographic data. In particular, when only some of them are used, (addresses, buildings or dwellings) there could be seri-

[*] Corresponding author: Fabio Crescenzi, ISTAT, Rome, Italy. E-mail: fabio.crescenzi@istat.it.

1874-7655/20/$35.00 © 2020 – IOS Press and the authors. All rights reserved
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).

ous consequences for the lack of a cross checked quality control among all data.

RSBL integrates the addresses of the National archive of addresses and urban streets (ANNCSU), made and managed by Istat, the Land Property Registry, and many other sources.

The micro-zone mapping will allow a better and more homogeneous mapping of the territory, not only for the purposes of the census, but also to offer new opportunities to learn more about the territory through spatial data produced by the Institute. It will also be useful for environmental protection policies and, in general, for the organization of local services. It will be enriched by the further territorial subdivision generated by the "regular grids" (areas defined by regular squares of 1 km^2), increasingly recognized as a reference output area by European regulations.

Further components of RSBL are the component "buildings and housing units", and the components on administrative and statistical zones. The "address" component of RSBL has been released and it is now maintained on a current basis. The release of the other four components is expected by 2021.

The elementary units in RSBL are geographical entities having different nature, both with respect to their geometric characteristics, and with respect to their genesis and temporal dynamics. The choice of the right entities may depend on the purpose of the analysis, the type of data, the phase of the statistical process in which the territorial data are used.

Since among these entities there are dependencies and multiple consistency requirements, it is necessary to make them coexist in a unique logical system.

The main elementary units in RSBL are addresses; micro zones and regular grids; buildings and housing units; administrative zones; functional and statistical zones.

In addition to providing users with the codes and the location of each type of unit, RSBL aims to characterize and classify.

RSBL will improve:

- Geo referencing in support of the statistical data production process;
- Spatial data production (e.g. surfaces, altitudes, distances, contiguities, statistics on buildings, etc.) also allowing to measure, characterize and classify the territory at a certain moment in time, or to evaluate its dynamics over time.

RSBL will make the following activities easier:

- The access to the codes of the geographical entities of each reference period to be used in the statistical processes;
- The correct georeferencing to geographical entities in the integration of data from different sources;
- The calculation of geo-statistics (surfaces, distances, contiguity, accessibility, etc.);
- The statistical evaluations of population concentration, spatial trends, classifications and/or clustering of territorial units;
- The planning of surveys and improvement of estimation phases (e.g. spatial sampling designs, small areas estimation designs, etc.);
- Data mapping.

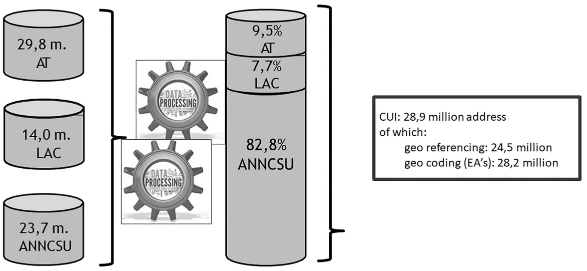
The main actions to improve quality of the RSBL are:

- Crosschecking of each components of the register with the others (e.g. the presence of housing units in micro-zones or the presence of occupied dwellings in buildings);
- The acquisition of "new" sources from which extract data (e.g. remote sensing images);
- The contribution of survey data to enrich the register, e.g. in the collection of coordinates, or in the detection of land use changes.

2. The address component of the RSBL

An address is any direct or indirect access, from a street to a housing unit or to other units where economic activities take place. If the point coordinates are available, it is possible to put the addresses on a map. The register will therefore allow to georeference units of a different nature (individuals, local units, buildings, etc.) that refer to the same address through an unified address code (CUI) allowing to make sure that the address is correctly written in each archive and source. This archive is essential for georeferencing data in a standardized manner in all systems, including information from the data contained in administrative files and in statistical surveys.

The Law of May 30, 1989, n. 223 establishes that each municipality compile and update the list of streets and the list of addresses according to the rules issued by the National Statistics Institute. In addition, the Law of 17 December 2012, n. 221, provides for the institution of the National archive of addresses and urban streets (ANNCSU), made and managed by Istat and the Land Property Registry. The ANNCSU is the reference database on streets and addresses, which contains information for the entire country in digital format. Each address is geocoded to the Census Enumer-



CUI: 28,9 million address
of which:
geo referencing: 24,5 million
geo coding (EA's): 28,2 million

Fig. 1. Address Register before and after data processing. EA: Enumeration Area; CUI: Address Code.

ation Areas of the Census Mapping. The National Digital Agency recognizes ANNCSU as a core database for the great impact on the transition to the National Register of Resident Population (ANPR) as well as on the many other uses of public interest.

As part of the data processing of the 2011 population census activities, municipal offices verified the misalignment of data and, when necessary, corrected, integrated and validated them.

The addresses component of RSBL is obtained using the data from ANNCSU as primary source and the data from others administrative and statistical sources. Point coordinates are essential for locating units in spatial analyses and producing statistics on regular grids, as required by the modernization of European statistics.

Figure 1 shows the results obtained integrating the addresses taken from the following sources:

- ANNCSU, the National archive of addresses and urban streets,
- LAC, the municipal registers of resident population,
- AT, the register of the tax Agency.

3. Micro zones and regular grids

Micro zones and regular grids are the two layers that we consider as the lowest geographical level for data dissemination [7].

- Micro zones are polygons characterized by a significant internal homogeneity. They are a subdivision of the enumeration areas of the 2011 census mapping. Micro zones will be sampling areas in field surveys.
- Regular grids, especially the 1 km^2, are layers on which statistical data are increasingly required, also according to European guidelines and regulations [3,4].

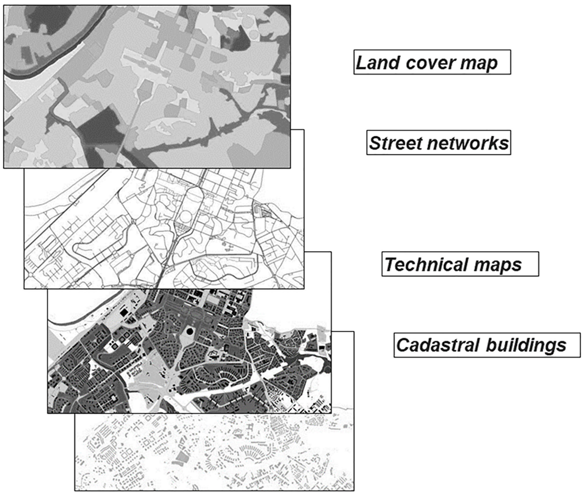


Fig. 2. Micro zones geo processing trough map overlay.

Here some of the main activities of the project:

- Improve the geometric quality, classification and detail of micro zones (starting from enumeration areas)
- Reuse of already available sources (Farm Register, Land Register)
- Acquire new cartographic sources to support the activities
- Develop a web-editing platform for check control of municipalities
- Improve land cover homogeneity
- Integrate data with open data sources

Figure 2 shows a scheme describing the process of splitting enumeration areas in micro zones.

4. Buildings and housing units

The buildings and housing units are a further component of the RSBL. The main reasons to include this component in RSBL are the following:

- To produce data on buildings and housing units for the purposes of the census according to the provisions of the Eurostat framework regulation 763/2008 and subsequent amendments;
- To provide benchmarks on buildings and housing units for national accounting estimates;
- To provide data on buildings and housing units required for dissemination at national and international level.

The data must guarantee the availability of data on buildings and housing units and, only for residential

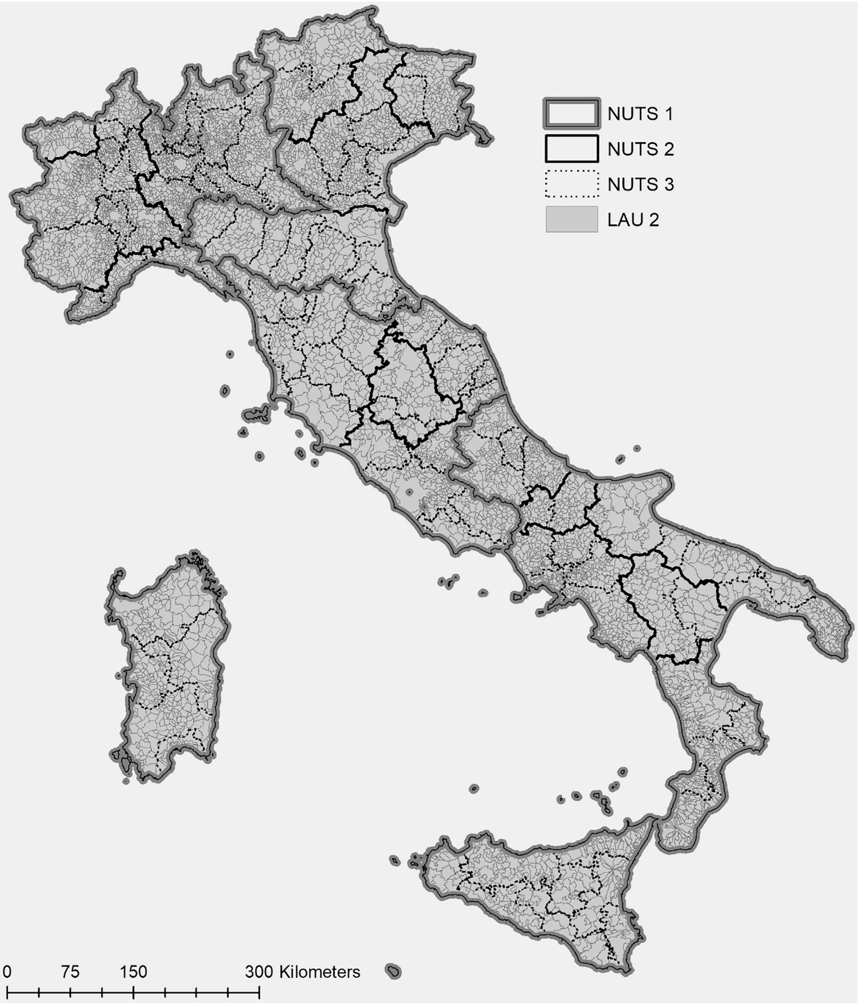


Fig. 3. Administrative borders.

buildings, on other structural characteristics (year of construction, number of floors, etc.).

Buildings contain housing units. A unique code identifies each building or housing unit. Each housing unit is associated with the code of the building that contains it.

Each unit of the building and housing unit of RSBL is associated with one or more CUI to allow the link to the information of the RSBL address component and to the statistical units of the other base registers.

If for the associated CUI point coordinates are available we refer units to geographical points, otherwise we refer units to areas through correspondence tables of address to micro-zones.

Since there is no difference between the statistical definition of "building" used by Istat and the cadastral definition of "building" used by the Land property register, the cadastral data of are taken as the main source for the census of buildings and housing units. The cadastral data are checked with other sources, including open source data and field control surveys.

5. The administrative units

The Constitutional Law, as well as the regulations on local authorities, establish the following territorial units recognized as territorial administrative units:

- Municipalities (LAU2);
- Unions of Municipalities;
- Metropolitan Provinces and Cities;
- Provinces (NUTS3);
- Regions and Autonomous Provinces (NUTS2).

With the exception of the Unions of municipalities, which are object of further study, ISTAT assigns to all of them an identification code. The municipalities are coded within the province and represent the minimum territorial unity which, through successive aggregations are obtained the unions of municipalities, provinces, metropolitan cities, regions and autonomous provinces (Fig. 3).

The main activities on the administrative units are:

- The ongoing maintenance of the database (the dynamics of changes of some of these units is very high and uneven on the territory);
- An in-depth study on the the coding rules to be adopted to guarantee the consistency and sustainability of historicizing data also in the long term;
- The enrichment of the territorial typologies to be compliant with the European TERCET Regulation [5].

6. Functional and statistical zones

Functional and statistical zones are areas whose boundaries are determined by institutional and/or socio-economic organization processes of the territorial governmental actions and/or principles of homogeneity with respect to environmental statistical or morphological characteristics. With reference to statistical zones, stakeholders and users require Labour market areas (areas beyond the administrative boundaries defined for purposes of compiling, reporting and evaluating employment, unemployment, workforce availability and related topics), industrial districts for socio-demographic and economic evaluations as well as for policy reasons. In [6] for each type of territorial zones are described:

- The needs that the zone meets,
- The definition process,
- The input sources needed for definition or updating,
- The constraints to which this process must undergo,
- The basic variables and the extended variables.

7. RSBL and quality

In a traditional survey, the addresses and other geographical features are attributes of the survey units, which are used to links units to the place where units are located. Conversely, from the opposite point of view, the addresses themselves and other geographical entities are statistical units which are affected by errors.

It is therefore necessary to choose a methodology for assessing the quality of the geographical entities in RSBL and to take appropriate action if the quality is bad. Bad quality could have as effect a wrong geographical location of the units which can produce errors in the sums of data for small areas.

The quality of geographic data would require procedures to measure several components, including logical consistency, completeness, positional accuracy, temporal accuracy and thematic accuracy.

For a more detailed discussion of the subject in question, see [11–14], where the main definitions of the components of quality are reported.

Comparison of the processed data with those coming from a cartographic source of greater accuracy may allow the evaluation of some of these components. The most accurate source may not be unique. For the process of integration, it is possible to use to this scope a plurality of available sources [2].

One of the criticism concerns the logical consistency. To evaluate this element and possibly correct the data, in the Census 2011 two instruments was used [7]. In particular, we mention the GIS geoprocessing tools to aid topological editing and the GIS geoprocessing tools that allow deterministic control of the data and the customization of the working environment used for the production of data.

Topological editing aid buttons simplify the acquisition of geographical data. The deterministic control buttons, on the other hand, check the consistency of the attributes, the control of the ranges and the validity of the attributes relating to the acquired geographical data. The control procedures requires verification of all the observations, at "micro" type level. The error localization algorithms implemented have proved to be particularly efficient in correcting any anomalies entered by the operator during the data acquisition phase. We will to adopt this procedures in the micro-zoning project.

For the assessment of positional accuracy, i.e. the difference between the location of the geographic coordinates relative to a certain object on the earth's surface and its representation in a database, the standard EMAS/ASPRS test (Engineering Map Accuracy Standard American Society of Photogrammetry and Remote Sensing).

The test requires the selection of some points of a layer and their counterparts identified on the control layer. The test is repeated on the single geographical coordinates x and y.

Let $x_i'(i = 1, 2 \ldots N')$ be the "true" coordinate of the point (where the coordinate is detected with a process that ensure negligible error, for example a measurement obtained from a GPS) and $x_i(i = 1, 2 \ldots N'')$ the corresponding coordinate in the register to be checked The error $ex_i(i = 1, 2 \ldots \min(N', N'))$ is calculated, defined by the difference of the two coordinates, its mean $E(e_x)$ and its variance $\mathrm{Var}(e_x)$. Assuming that the errors come from a normal distribution, of mean and variance specified above, it is possible to identify the test statistic through the related estimators, $M(e_x)$ and $S^2(e_x)$ respectively, whose expression is as follows:

$$X_x^2 = (n-1)S^2(e_x)/\sigma_{x0}^2$$

where σ_{x0}^2 is the acceptable error limit in the x dimension. Established σ_{x0}^2 for example at 8 or 10 meters, the statistic will be subjected to the test check to see if the difference between the two coordinates is random or not. The test is then repeated also on the y coordinate.

For the evaluation of the other elements of the geographic data quality, please refer to [12].

Concerning addresses, they are first of all normalized and recognized. Addresses are processed by an automated software E-GON.[1] E-GON service has been specially designed to ensure that all the information contained in addresses database is correct and univocal.

The evaluation of quality of the address requires indicators of all the above mentioned components of the quality of geographic data, including:

- Positional accuracy – in case the x and y of point coordinates of the addresses or of the coordinates of the small areas deviate from the "true" position.
- Syntactic accuracy of the address – in this case it is a matter of measuring the: "false non-matches: some records of the two databases refer to the same unit but the merge is not able to identify them since at least one key variable is affected by some error" (it is one of the possible causes of over-coverage);
- False combinations: some records can be combined even if they actually refer to different units (it is one of the possible causes of under-coverage);
- Time accuracy – when attributes have different reference periods.

The most important indicators are the following:

Table 1
Georeferenced and non-georeferenced addresses

Addresses	Number	Percentage
Georeferenced	26.392.405	91,0
Non-georeferenced	2.603.548	9,0
Total	28.995.953	100,0

- Over-coverage – percentage of addresses not existing in the territory included in the RSBL
- Under-coverage – percentage of addresses existing in the territory not included in the RSBL

We started trying to estimate the under-coverage rate, further estimates will follow.

Two distinct groups were considered, the first is composed by the addresses not linked, and the second by the total number of Egon addresses. The ratio of units in these two groups is about 8% which is an overestimate of the undercoverage rate since the syntactic accuracy of the address affects non recognized cases.

If the address is correctly normalized and recognized it is marked by a flag of georeferenced address and the software associates the coordinates and/or Enumerations Area, otherwise the address is flagged as no georeferenced. It is clear that if we were too inclusive we could produce an excess of bad geographical locations, on the contrary, if we were too little inclusive, we would risk losing too much geographical location of units. A first ingredient used in this evaluation is the "string-completeness of the address" which concerns the completeness of all the information needed to fully describe an address: street-type (the type of street, e.g. via or piazza), name (the name of the street), address number (the progressive number which identify the address inside the street).

A second ingredient is the completeness of the attributes, which are relevant for geographical location. These are point coordinates, the administrative codes, the enumeration area codes, the regular grid codes.

In the Table 1 below the number and percentage of georeferenced addresses and non-georeferenced addresses in RSBL at the time of writing this paper.

8. Concluding remarks

The Statistical Base Register of Territorial Entities is built with the aim to standardize geo referencing both administrative and survey data. In this paper, we presented some of the main features of RSBL and some preliminary results on the data contained in it.

Here the issues that we consider of particular importance for the next steps:

[1] See https://www.wareplace.com/products/.

- To complete and refine the strategy to maintain update over time geographical entities in the base registers also in relation to the needs of the permanent census.
- To increase the number of quality indicators of geo referencing of RSBL data.
- To reduce the "Non-georeferenced" addresses of Table 1 using techniques of probabilistic recognition.
- To increase the use of the GIS for statistical purposes also in the data production processes.

References

[1] Bakker, B.F.M., Rooijen, J. van & Toor. 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, 2014, 30(4), 411-424.

[2] Cerroni, F., Di Bella, G., Galiè, L. Evaluating administrative data quality as input of the statistical production process, RIVISTA DI STATISTICA UFFICIALE N. 1-2/2014, 2014 Istat.

[3] Commission Implementing Regulation (EU) 2017/543 of 22 March 2017 laying down rules for the application of Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns, C/2017/1728.

[4] European Parliament, Regulation (EC) No 763/2008 of the European Parliament and of the council on population and housing censuses, *Official Journal of the European Union*, 13.8.2008, L 218/14-L 218/20.

[5] Eurostat 2018, Methodological manual on territorial typologies – 2018 edition © European Union, 2019.

[6] Franconi, L., Ichim, D., D'Alò, M., Cruciani, S. Guidelines for Labour Market Area delineation process: from definition to dissemination, Released: August 2017, 2017. https://ec.europa.eu/eurostat/cros/system/files/guidelines_for_lmas_production08082017_rev300817.pdf.

[7] Lipizzi F. Innovazioni di processo e di prodotto nelle fasi di aggiornamento delle basi territoriali 2010–2011. Istat Working Papers, n.2/2013, 2013.

[8] Mugnoli, S., Lipizzi, F., Esposto, A. New ISTAT "micro zones" layer: a new way to read land cover statistics, J-Reading-Journal of Research and Didactics in Geography, 2 2018.

[9] Schulte Nordholt, E. The dutch virtual census 2001: A new approach by combining different Sources, Statistical Journal of the United Nations Economic Commission for Europe, 2005, 22, 25-37.

[10] UNECE, 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, 2007.

[11] Veregin, H. Data quality parameters. In Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., eds., Geographical information systems, 2nd Ed., Vol. 1, Chap. 12, New York, NY: Wiley, 1999, pp. 177-89.

[12] Veregin, H. Data Quality Parameters. In Longley, M.F.G.P.A., Maguire, D.J., Rhind, D.W., eds., New Developments in Geographical Information Systems: Principles, Techniques, Management and Applications, Hoboken, NY: Wiley, 2005, pp. 177-189.

[13] Wallgren, A., Wallgren, B. Register-based statistics – administrative data for statistical purposes, John Wiley and Sons, Chichester, 2007.

[14] Zandbergen, P.A. A comparison of address point, parcel and street geocoding techniques Computers, Environment and Urban Systems, 32(3), 2008, Elsevier.

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