Improving statistical registers' quality through attribute-driven spatial matching

Damiano Abbatini, Luisa Franconi, Daniela Ichim

Istat – Italian National Institute of Statistics, via Cesare Balbo 16, 00184, Rome, Italy

abbatini@istat.it, franconi@istat.it, ichim@istat.it

Abstract

In response to the increasing demand for information in modern society, official statistics often relies on multi-source production models. Since statistical registers have lower production costs compared to field surveys, these registers are frequently integrated to derive the estimates of interest. This contribution proposes a data reconciliation method through spatial correspondence to integrate registers of residential addresses and buildings, enhancing their quality and increasing the capability of linkage operations. The method leverages housing ownership conditions, owners' residential addresses, and proximity between buildings and addresses. Implementation and evaluations were performed using an Italian NUTS2 Region as study area. Data from the Integrated System of Statistical Registers of the Italian National Institute were used. About 22 per cent more addresses were found to belong to both addresses and building registers. Our approach is flexible enough to allow different extensions derived from parameterizations.

Keywords: buildings register, addresses register, data reconciliation, spatial matching, ownership

1. Introduction

Following modernisation of official statistics production model, the Italian National Institute of Statistics (Istat) moved to massive usage of statistical registers. The new production model is based on the Integrated System of Statistical Registers (ISSR), a single logical data asset resulting from the integration of survey and administrative data.

The ISSR comprises both annual master and satellite statistical registers (Alleva, 2017). Master statistical registers, i.e. Population Register, Business Register and Register of Places, describe their entire corresponding populations of statistical units. Moreover, each individual, enterprise or place is assigned a unique unambiguous identification code. One of the main objectives of ISSR is to geo-reference all core statistical units involved in official statistics.

All socio-demographic statistics rely on the integration of the Population Register and the Statistical Register of Places. The former includes demographic information, e.g. age, gender, citizenship, on resident population. The latter is a multidimensional and complex register integrating and connecting different components dedicated to different spatial units: addresses, buildings and dwellings, census blocks, grids and administrative and territorial statistical units (Abbatini et al. 2024).

As for the component devoted to addresses, its importance relies on the fact that most statistical units present one or several addresses to describe phenomena. An address is any direct or indirect access from a street to a building unit or units where activities might take place. The usual way to represent an address on a map is via its latitude and longitude coordinates. The aim of the Register of addresses is to give a Unique Identification Code (CUI) to each address stemming from several sources and to assign valid geographical coordinates to each of them. Currently, the Register contains 30 million CUIs; about 88 per cent of them have valid coordinates (Abbatini et al., 2024).

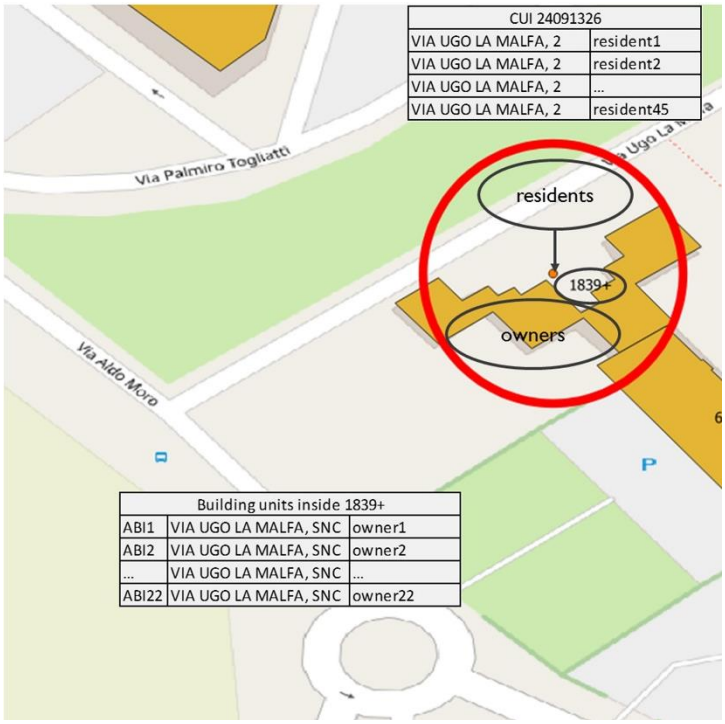
The component devoted to the Italian real estate assets is the Register of buildings and building units, mostly stemming from cadastral administrative data. Again, every building and building units present a (nested) unique identifier. The Register currently collects information on 29 million buildings; 14.4 million of them are residential. About 80 per cent of buildings have valid geographical coordinates, (Abbatini et al., 2024).

Microdata integration of the Population Register and the Register of places through resident address should allow production of aggregated data consistent with official statistics quality standards (Zhang, 2012). In fact, the Population register contains the address of residence of each individual; at the same time the building (unit) where the individual lives present an address in the Register of buildings. The matching of this same address, belonging to both registers, and its associated geographical coordinates (Register of addresses) makes it possible to position accurately the population in their home on the territory.

However, the sole information provided by addresses is not sufficient to fully link registers. In fact, different administrative data sources may contain slightly different information that causes the failure of matching methods. An example is reported in Figure 2 where the same address is correctly and completely listed in the population and in the addresses register whereas street number information is missing in the Register of buildings not allowing the correct association of individuals to their dwelling/home.

This paper builds on the spatial matching of addresses and buildings registers using attributes stemming from the population register. Scientific literature mentions data reconciliation as a first step in any data integration process. Indeed, when two or more data sources are matched, data reconciliation identifies which entities refer to the same real entity (Bakhtouchi, 2022). The aim of the paper is to illustrate preliminary results of a spatial reconciliation method applied to address and building registers. The method relies on information from the population register to borrow strength. In this setting, the reconciliation involves only addresses where population lives (resident addresses).

Figure 1: Correctly identified residential CUI in the Register of addresses, top table; missing information on the street number of building with identifier 1839+ in the Register of Buildings (bottom table) not allowing correct association



The study area is the Italian NUTS2 region of Emilia-Romagna (Nomenclature of Territorial Units for Statistics, Eurostat, 2020). Table 1 shows the initial situation when matching by CUI: it contains the percentages of CUIs of residential addresses that do not match any unit in the Register of buildings when joining them by CUI (i.e. street type, name and number). There are several administrative and technical explanations behind these mismatch rates. The mismatch rate indicates that registers should undergo a data reconciliation procedure before jointly using them in any statistical process (e.g. calibration, estimation, dissemination, etc.).

The spatial matching method for reconciliation of addresses and buildings registers by means of population attributes observed in the population register is described in the next section. The section Results illustrates some preliminary findings obtained in Emilia-Romagna NUTS3 regions. Finally, the last section draws main conclusions and sketches possible research directions and improvements.

Table 1: Percentage distribution of residential CUIs not included among addresses of buildings by NUTS3 regions of Tuscany description

NUTS3	Residential CUIs without an associated building (%)
Piacenza	56.70
Parma	36.16
Reggio nell'Emilia	38.47
Modena	34.61
Bologna	28.59
Ferrara	28.72
Ravenna	25.43
Forlì-Cesena	30.48
Rimini	46.64
Total	**33.59**

2. Data and Methodology

2.1 Data

In this work, the Registers of buildings and addresses are integrated. They are both components of the Istat Statistical Register of Places, see Abbatini et al. (2024). The illustrated analyses focus on residential addresses and residential buildings in the NUTS2 Emilia-Romagna region in Italy. The region is located in the central part of Italy, it comprises of nine NUTS3 regions (provinces), see Figure 2.

Figure 2: NUTS3 regions of Italy (red) and of the Emilia-Romagna NUTS2 region (yellow).

For the purposes of this work, the main variables related to buildings are cadastral address and personal identifiers. The latter identifies individuals or legal entities holding rights (ownership, rental, etc.) on the building unit. The spatial information part of the statistical register of buildings pertains to geo-graphic coordinates of their centroids. About 10 per cent of buildings have missing geographic coordinates. This may be due to updating, projection or digitalisation issues, etc. see Comparetti & Raimondi (2019), Mora et al. (2022). As for the residential addresses, only their identifiers and spatial coordinates are required by the proposed data reconciliation methodology.

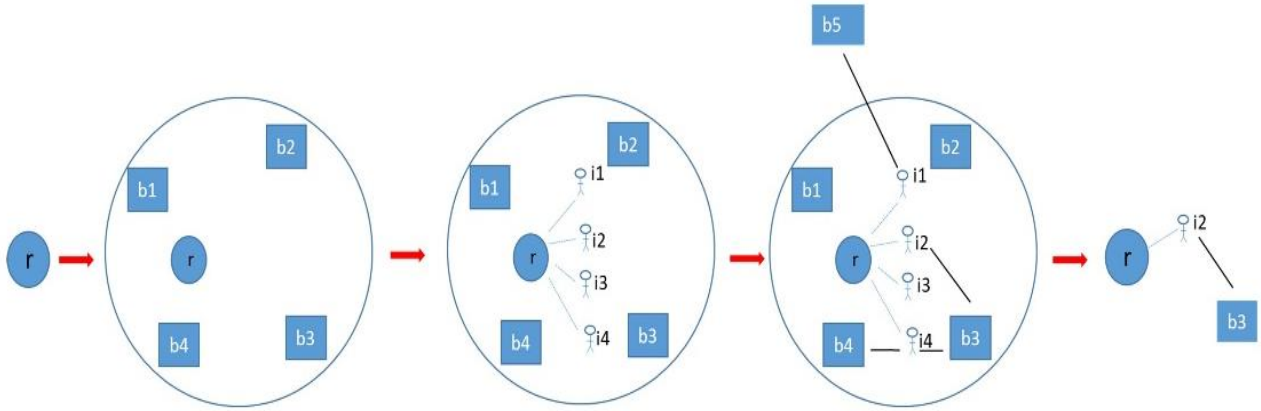
2.2 Methodology

The proposed GIS-based address matching is derived from the idea that owners of a single dwelling usually live in their own house. Moreover, it is supposed that the geographical coordinates of a building centroid are in the proximity of the point representing the building address. The GIS-based linking process is illustrated in Figure 3.

Given a residential address *r*, the spatial matching method between coordinates first identifies all nearby buildings. A pre-defined buffer area around each residential address allows the operationalization of proximity between buildings and the residential address of the owner. The Euclidean distance is used to measure distances between buildings and addresses. In this application the used buffer size equals 30 mt.

Subsequently, the attribute-driven part of the algorithm analyses residents in r and selects the buildings b whose owners are single-owners residents in r. Thanks to the uniqueness property of the single-residents, the procedure should identify a single building. Such building b is then matched to *r*; in case the building had a missing address identifier, the reconciliation of addresses takes place (Curriero et al., 2010).

Figure 3: GIS-based address matching (*r-b*) derived from a sequential selection process starting from geometric buffers. Large circle indicates the buffer, blue circles represent residential addresses, squares represent buildings, symbols labelled i1, i2, i3 and i4 represent individuals, dashed lines indicate a residential relationship, full lines indicate ownership relationships.



3. Results

The main results of the work carried out using the adopted methodology are reported in Table 2. The first column shows the percentage of the resident population at addresses not linked to buildings, calculated at time t0 before the matching procedure. In the second column, the percentage of matched population relative to the total population of the first column is provided.

In general, the initial proportion of unallocated population amounts to a quarter of the total (25.11 percent), as indicated by the regional data in the table 2. At a higher level of territorial detail, the data shed light on a notably diverse initial landscape across the region, with certain areas encountering more pronounced challenges compared to others. For instance, Piacenza and Reggio nell'Emilia provinces exhibit relatively higher percentages of the population residing at addresses not linked to properties in the register (38 and 37.64 per cent, respectively) than the rest of the region. Indeed, they contrast provinces like Bologna and Ravenna, where the percentages are notably lower (18.11 and 18.48 per cent, respectively).

Table 2: Resident Population subject to spatial matching and success rate by Province

NUTS3	Population on Addresses not associated with Buildings* (%)	Success rate** (%)
Piacenza	38.00	28.22
Parma	24.02	23.13
Reggio nell'Emilia	37.64	43.19

Modena	26.51	51.15
Bologna	18.11	51.25
Ferrara	21.80	44.53
Ravenna	18.48	26.23
Forlì-Cesena	21.75	32.60
Rimini	29.12	45.75
Total	**25.11**	**40.87**

* *Percentage calculated on the Total Resident Population*
*** Percentage calculated on the amount of Population on Addresses not associated with Buildings*

Regardless of the starting situation, the procedure did not yield uniform results across all provinces. A significant variability in success rates is observed, with the lowest matching rate recorded in Parma (23.13 per cent) and the highest in the province of Bologna (51.25 per cent), which also had the most favorable initial situation compared to other provinces.

Even Piacenza and Reggio nell'Emilia although starting from comparable situations, achieved very different results. In Reggio nell'Emilia, the achieved success rate achieved equals 43.19 per cent, whereas in Piacenza the success rate was is less than 30 per cent.

4. Conclusions

The work presents preliminary results of a spatially based matching procedure. The achieved matching rates are extremely encouraging and provide a sound foundation for further development. There are many opportunities to improve the proposed approach. For example, territorial aspects of the method can be subject to various modulations, such as varying buffer radius size around coordinates to explore additional recovery possibilities. Additionally, it will be essential to analyse the observed territorial differences more deeply to identify the variables that may influence success rates. Furthermore, we can enhance the starting context by optimising buildings geographical position. These potentialities offer important directions for future work, which will focus on improving and refining the methodology.

Acknowledgment

Although the article is the result of a discussion and collective work of the three authors, it is possible to attribute paragraph 1 to Luisa Franconi, paragraph 2 to Daniela Ichim, and paragraph 3 to Damiano Abbatini; the conclusions are common to all authors.

Istat is not responsible for any views expressed in this paper.

References

Abbatini, D., Clary, T., Chiocchini, R., Fardelli, D. et al. (2024). Statistical register of places: opportunities for sustainable and climate change related indicators. *RIEDS - Rivista Italiana di Economia, Demografia e Statistica - The Italian Journal of Economic, Demographic and Statistical Studies*, 78(1), 85-96.

Alleva, G. (2017). The new role of sample surveys in official statistics. ITACOSM 2017, The 5th Italian *Conference on Survey Methodology*, Bologna, Italy. Available at: https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf (accessed December 2023).

Bakhtouchi, A. (2022). Data reconciliation and fusion methods: a survey. *Applied Computing and Informatics* 18 (3/4), 182–194. https://doi.org/10.1016/j.aci.2019.07.001

Comparetti A., Raimondi S. (2019). Cadastral models in EU member states. *EQA - International Journal of Environmental Quality* 33, 55-78. https://doi.org/10.6092/issn.2281-4485/8558

Curriero, F.C., Kulldorff, M., Boscoe, F.P., Klassen, A.C. (2010). Using Imputation to Provide Location Information for Nongeocoded Addresses. *PLoS One*, 5(2):e8998. https://doi.org/10.1371/journal.pone.0008998

Eurostat. (2020). Statistical regions in the European Union and partner countries. NUTS and statistical regions 2021.doi:10.2785/850262

Mora-Navarro, G., Femenia-Ribera, C., Velilla Torres, J.M., Martinez-Llario, J. (2022). Geographical Data and Metadata on Land Administration in Spain. *Land*, 11, 1107. https://doi.org/10.3390/land11071107

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66: 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x