Improving statistical registers: innovative integration methods for building and population registers

Damiano Abbatini, Armando D'Aniello, Stefania Lucchetti,

Enrico Orsini, Andrea Pagano

abbatini@istat.it, Italy

armando.daniello@istat.it, Italy

lucchetti@istat.it, Italy

eorsini@istat.it, Italy

andrea.pagano@istat.it, Italy

Abstract

One of the current challenges for statistical institutes is to produce official statistics using administrative archives. In particular, at ISTAT, the construction of statistical registers for different statistical units has recently begun, characterizing and allowing a snapshot of the socio-economic reality of the country.

One fundamental aspect in the construction of registers is to devise processes to create an integrated architecture among different statistical units. In this work, we describe the use of new innovative methodologies for integrating data from the real estate and building registry with data from the resident population registry, with the aim of uniquely placing a family within a dwelling.

These registers are fed with data from the cadastral archive of real estate and buildings, as well as municipal population registers, and both contain information related to addresses with their respective geo-coding. The process begins with geo-coded addresses through geographic coordinates and involves considering two different deterministic methodologies.

The first methodology is based on the ownership of a single dwelling by comparing the proximity between the residential address and the dwelling address with different levels of geographic precision. The second methodology is applied to all families that do not own a dwelling or own multiple dwellings. For these families, a matrix of resident families at an address is essentially constructed for all available properties at that address.

The methodology for resolving this matrix to build a unique household-dwelling association has been implemented by calculating a weight that measures the quality of the association. This weight is calculated based on some variables of both geo-spatial and non-geo-spatial nature. The application of this weight allows the use of the harmonized combinatorial optimization method that solves the assignment problem known as the Hungarian algorithm in polynomial time.

The algorithm solves the unique assignment problem applied to families and dwellings by maximizing the sum of the calculated weights. In practice, all possible family-property pairs are encoded through a bipartite graph, where vertices are the elements to be associated, and edges represent possible pair choices, and for each edge, we have the calculated weight.

The application of these methodologies has allowed placing the entire resident population uniquely in dwellings in the best possible way. And the result of this integration allows calculating various statistical indicators that were previously obtainable only through conducting surveys.

Keywords: dwellings, households, record linkage, statistical registers

1. Introduction and Theoretical framework

Istat is restructuring its production processes towards an Integrated System of Statistical Registers; within this new framework, the Statistical Register of Place (RSBL) provides geographical statistical data to complement the statistical information from other Registers (socio-demographic or economic).

In this paper, we describe the elements used, the issues encountered, and the solutions adopted in the linkage between households and dwellings among the resident population in the 2021 Population Census and the Residential Buildings and Housing Register (RSBL), limited to the housing and building component.

In the traditional census, where information about households was collected through specific survey models delivered to households and then retrieved once completed, the household-dwelling link was contextual since the information of both was contained in the same questionnaire. In the new integrated registry system, the statistical units of interest come from different administrative archives, and the link between the various units must be established through identification codes, when available and in compliance with data protection regulations, and additional information.

Since October 2018, the Italian National Institute of Statistics (Istat) has initiated the Permanent Population and Housing Census (replacing the Decennial Population Census), based on the integration of information available from administrative sources with those acquired from rotational sample surveys conducted in all Italian municipalities (Istat 2022). For the new permanent census, the main reference administrative archive is the National Resident Population Archive (ANPR), which collects and centralizes the contents of municipal registry lists; enriched and corrected by the results of the mentioned annual sample surveys in the territory (Master Sample), the Integrated Archive of Dwellers Habitually in Italy (AIDA), and the information made available by the Basic Statistical Register of Individuals, Families, and Cohabitants (RBI).

Regarding dwellings, instead, the main reference administrative archive is the cadastral archive, integrated with the main information from the 15th General Population and Housing Census of 2011 and the archives of real estate leases.

The geographical element characterizing the individuals and households surveyed is the residential address: a string comprising the type of circulation area identifier (street, avenue, alley, square, etc.), the official name assigned to the circulation area (e.g., 'Marco Polo', 'Julius Caesar'), the house number, and any exponent. In RSBL, each address is coded through a Unique Address Code (Codice Univoco di Indirizzo CUI).

The dwellings in RSBL come from the cadastral archive, and in this case, the geographical elements are the cadastral code of the municipality and the references of sheet, parcel, and subaltern associated with the single property. In the cadastral archive, furthermore, the properties are distinguished into dwellings or others and are associated with one or more addresses. These addresses are also acquired in RSBL, and each of them is associated with the reference CUI. Finally,

for each property, information is available on the natural and legal persons holding rights, the type of right, and the respective ownership shares.

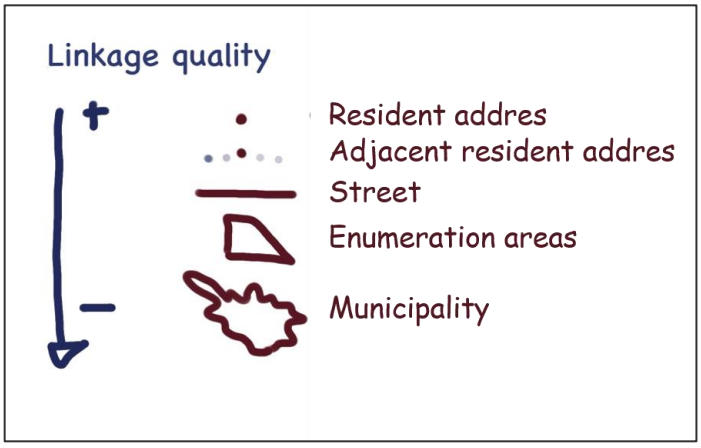
Both the addresses and the properties, but from now on, we will refer more precisely to dwellings, in most cases also have other territorial references, such as geographic coordinates or identification codes of enumeration area in which they fall.

For census purposes, which involve the calculation and dissemination of housing indicators, all households residing in dwellings must be uniquely placed in a single dwelling.

From a theoretical point of view, this is a deterministic record linkage problem based on precise matching variables (Scanu, 2003; D'Orazio, Di Zio, Scanu, 2006); the main matching variable is the address (CUI): households residing at an address are placed in the dwellings located at the same address. The second matching variable is the individual who appears in the census units as a 'resident' while in the housing register appears as the holder of rights to the property.

Figure 1 illustrates the ways in which addresses can link households to dwellings: for each CUI of residence, all associated dwellings are considered, thus forming groups of N households and M dwellings characterized by various qualities of the address linkage, ranging from the best level (identity up to the house number) to the worst level (same dwellings and address within the same municipality). Subsequently, within each group, cases are identified, if present, where the household also owns or rents one of the dwellings at that address.

Figure 1: Qualities of the address linkage



The following section provides a detailed explanation of the methods and procedures used for allocation.

2. Integration methodologies

This chapter outlines the final linking process aimed at establishing a one-to-one correspondence between families and real estate units. The procedure commences with the application of various link types across three distinct households groups. It employs two deterministic methodologies:

1. Uniquely Deterministic Association:

- This methodology ensures exactness by design. It establishes unique associations between families and real estate units.

2. Secondary Deterministic Approach:

- For cases where uniqueness is not straightforward, this methodology processes all non-unique scenarios. Its objective is to provide the best possible unique solution. In summary, this chapter delves into the intricacies of achieving definitive pairings between families and their corresponding residences within the total Italian resident population.

2.1 Definition of the link levels and resident groups

First, resident families are divided into three different types identified by the presence or absence of one or more holders of housing unit within the family unit [table 1].

Table 1: Household types

Household type	Household type descriptions
1	Household with a least one member who owns a housing unit
2	Household with at least one member who has lease agreement for a housing unit
3	Household with all members who are neither owners nor have a lease agreement for a housing unit

The quality of the household–dwelling linkage is defined according to a scale of twelve levels that varies depending on the proximity between the residence address and the address of the housing unit. Table 2 describes linkage levels in descending order of quality.

Table 2: Linkage quality

Linkage quality level	Linkage quality description
1	Household with residence address matching the address of housing unit
2	Household with residence address matching one of the building's addresses
3	Household with residence address adjacent to the address of housing unit (maximum distance eight house numbers)
4	Household with residence address adjacent to one of the building's addresses (maximum distance eight house numbers)
5	Household with residence address not matching the address of housing unit but matching the street within the same census section
6	Household with residence address not matching the address of housing unit but matching one of the building's streets within the same census section
7	Household with residence street not matching the street of housing unit but within the same census section

8	Household without a residence address or housing unit without an address
9	Household without a residence address and housing unit without an address
10	Household with residence address not matching the address of housing unit but matching the street in different enumeration areas
11	Household with residence street not matching the street of housing unit in different enumeration areas and with a maximum distance of 200 meters
12	Household with residence address in the municipality of the housing unit

In the case of holders (owners or tenants of a property), the bond between the family and the housing unit is very strong, so all levels of linkage can be used.
For non-holders (neither owners nor tenants of any housing unit), the only components for the link considered reliable are the address and the street, which is why only the first four levels of linkage are applied.

2.2 First deterministic integration methodology

The first deterministic methodology is the strongest association between families and properties, and it is based on two highly restrictive conditions. The first condition is that families who are the legal owners of the property (groups 1 and 2) must be considered before any other families. The second constraint is the uniqueness of associations. This means that the families who are the owners of the property possess a single property, while the families who are not the owners of the property possess a single property that is not occupied by the owners. The associations are constructed through an algorithm that verifies the existence of the two constraints and applies the best possible link to the data. Consequently, for the holders of the title, all the different types of link are considered, whereas for those who are not holders of the title, only the four best types of link are considered, as described in paragraph 1.3 (table x).
Families associated with a specific type of link are excluded from subsequent link types. Furthermore, each produced combination may result in the creation of new potential combinations, thus necessitating the reiteration of all link types until the outcome of the combinations is exhausted. The overall result of this initial association methodology is a preliminary output of families associated with a single residence. This output is then further classified into two categories, a family and an apartment (1:1), multiple families and an apartment (n: 1).
The latter category defines the proportion of cohabiting families.
The remaining unmatched families are those for which the aforementioned constraints have not been met, namely, there are no unique linking conditions. In general, there are numerous sets of N families associated with M properties, which we refer to as K. The resolution of this set is achieved through the second deterministic methodology.

2.3 Second deterministic integration methodology

The second deterministic methodology was designed to solve the set K with the criterion of having the best possible unique associations.

For this purpose, each single association between family and property was assigned a score between 0 and 2 where 0 indicates the minimum quality of association and 2 the maximum quality of association.

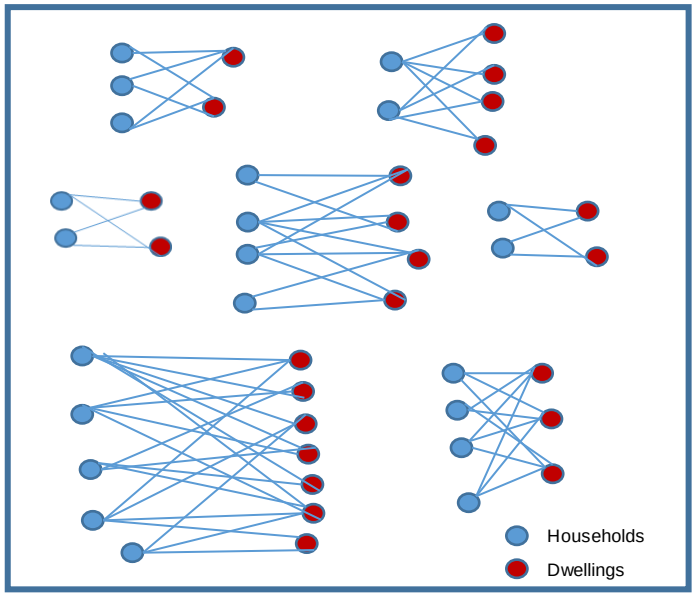
The score is constructed using four variables with different weights, the share of ownership of the property, the distance between the coordinate of the residence address and the coordinate of the centroid of the building of the property and the ratio between the number of rooms and the surface area of the property compared to the number of members of the family unit as described in Appendix.

Once the scores were assigned, it was possible to apply the Hungarian algorithm, a combinatorial optimization method that solves the assignment problem in polynomial time (Kuhn, 1955). The method was developed by Harold Kuhn in 1955 and today is implemented by an R package (R Core Team, 2023), (Papadimitriou and Steiglitz, 1982), (Hornik, 2005).

In practice, the method, starting from an N x M matrix, determines the unique associations by maximizing the sum of the total scores.

The set K essentially contains many bipartite graphs, i.e. all the possible associations between families and properties for a given type of link as shown in figure 2.

Figure 2: Set K representation



The application of the Hungarian algorithm in our specific case is bound to two conditions, the first is that N must be equal to M or there must be sufficient homes for the families and the second is that the bipartite graphs are independent of each other, that is a family and a property cannot be in multiple graphs otherwise uniqueness would not be achieved.

To satisfy these conditions, fictitious properties with a score of 0 where M>N were added and new macro graphs were constructed to bring together all the graphs in correlation by family and property.

Thanks to these measures it was possible to run the algorithm on the entire set K, but clearly the pairings of families in fictitious properties were eliminated and constituted a new residue.

For this residue, a new K1 set of graphs was constructed, obtained by associating all the families residing in a specific census section of residence in all the homes left vacant in that section. Once the score was recalculated according to the criteria described above, the Hungarian algorithm was reapplied.

Overall, this second output made it possible to complete the univocal association between family and home for the total Italian resident population.

3. Results and Conclusions

The main results of the analysis conducted using the adopted methodology are summarized in Table 3. The two procedures outlined in the previous paragraph allocate nearly the entire considered population, with a final residue of less than one percent (0.95%).

The first method links three-quarters of the reference population (75.36%), exhibiting significant territorial variations, with the Central-Northern regions showing the highest percentages: Veneto (84.04%), Friuli-Venezia Giulia (82.82%), Marche (81.35%). Conversely, four Southern regions (Calabria 61.22%; Campania 62.26%; Sicily 71.30%; Sardinia 72.67%) and one Central region (Lazio 72.15%) record lower values. Campania, Calabria, and Sardinia also show significantly higher residual shares compared to other territorial entities (respectively 3.31%, 2.92%, and 1.95%).

Table 3: Population percentages by type of link and NUTS2 region

NUTS 2	Uniquely Deterministic Association	Secondary Deterministic Approach	Residual Processing Approach
Piemonte	77.65	22.17	0.18
Valle d'Aosta	75.11	24.78	0.11
Lombardia	78.79	20.80	0.40
Trentino–Alto Adige	74.96	24.39	0.65
Veneto	83.04	16.57	0.39

Friuli–Venezia Giulia	82.82	17.00	0.18
Liguria	74.31	25.50	0.19
Emilia-Romagna	79.28	20.41	0.31
Toscana	79.50	20.04	0.46
Umbria	77.04	22.24	0.73
Marche	81.35	18.40	0.25
Lazio	72.15	26.51	1.33
Abruzzo	75.07	24.34	0.59
Molise	76.90	22.59	0.51
Campania	62.26	34.82	2.92
Puglia	78.14	20.93	0.94
Basilicata	74.43	24.15	1.41
Calabria	61.22	35.47	3.31
Sicilia	71.30	27.58	1.12
Sardegna	72.67	25.38	1.95
Total	**75.36**	**23.70**	**0.95**

Appendix

For the second deterministic methodology. we developed a scheme to assign scores to family-real estate pairs. This scheme assigns 0 as the minimum score and 2 as the maximum score to indicate the highest quality of association.

The scheme is based on four variables related to different weights:

- Combination of link type. quality of the coordinate of the residential address and. if present. the distance between that coordinate and that one of the building centroid of the real estate. For the distance. it is considered whether this is the minimum among all the family-real estate pairs with the same residence CUI (Table A.1).
- The ownership share of the property (Table A.2).
- The ratio of number of rooms-and-spaces of the real estate to the number of household members (Table A.3).
- The ratio of surface area of the real estate to the number of household members (Table A.4).

Table A.1: Scores by link type. coordinate quality. distance

Link Type	Coordinate quality	Distance	Score
1	1 – 2	min	0.75
1	3	min	0.65
< > 1	1 – 2	min	0.55
< > 1	3	min	0.45
1	1 – 2	< > min	0.65
1	3	< > min	0.55
< > 1	1 – 2	< > min	0.45
< > 1	3	< > min	0.35
1	no coordinate	-	0.45
< > 1	no coordinate	-	0.25

Table A.2: Scores by percentage of ownership

Ownership percentage	Score
[0.75 . + [	0.25
[0.50 . 0.75 [	0.20
[0.25 . 0.50 [	0.15
] 0 . 0.25 [	0.10

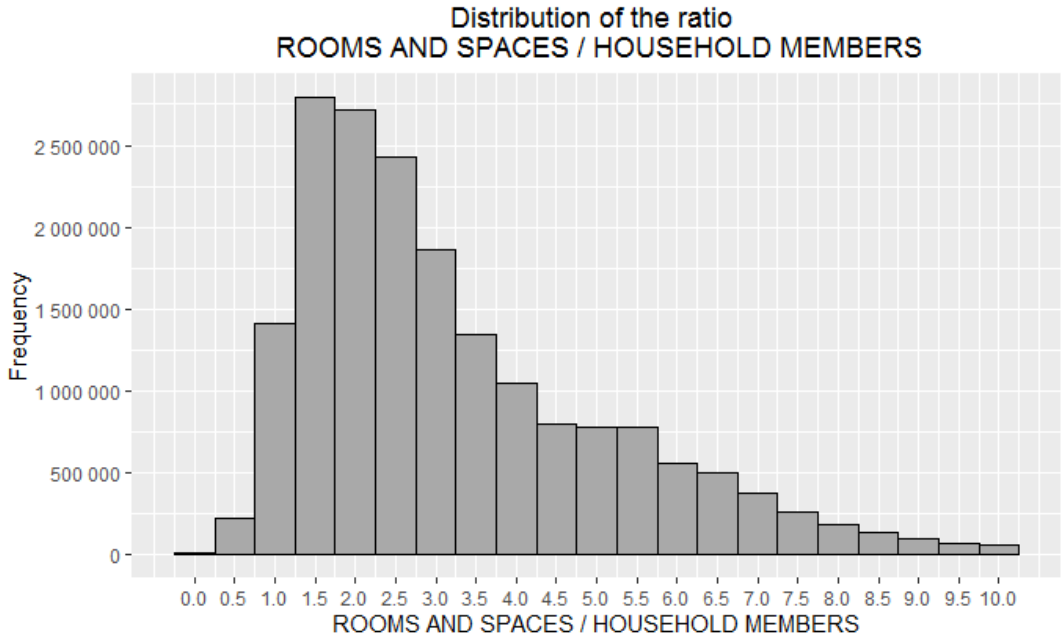
Table A.3: Scores by rooms and spaces

Property rooms and spaces / household members	Score
[1.3 . 1.7]	0.125
[1 . 1.3 [.] 1.7 . 2]	0.100
] 2 . 3]	0.080
] 3 . 5]	0.070
] 5 . 7]	0.060
[0.5 . 1 [.] 7 . 10]	0.050

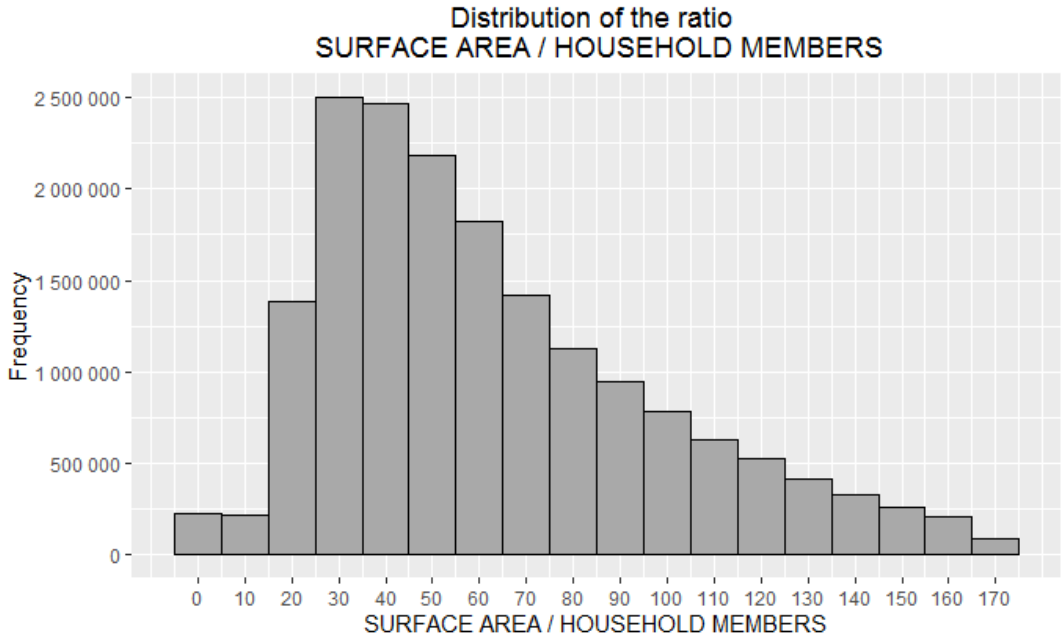
Table A.4: Scores per surface area

Property surface area / household members	Score
[30 . 40]	0.125
] 40 . 60]	0.100
] 60 . 90]	0.080
[20 . 30 [.] 90 . 110]	0.075
[10 . 20 [.] 110 . 140]	0.050
] 140 . 170]	0.025

To determinate the scoring intervals for rooms-and-spaces and surface area. we analysed the distributions of the ratios of rooms-and-spaces and surface area to the number of household members for the family-real estate pairs already validated in previous stages of the process (Picture A.1. Picture A.2).



Picture A.1: Distribution of the ratio of rooms-and-spaces to household members



Picture A.2: Distribution of the ratio of surface area to household members

Acknowledgment (11pt. bold)

Although the article is the result of a discussion and collective work of the five authors, it is possible to attribute paragraph 1 to Stefania Lucchetti, paragraph 2.1 to Andrea Pagano, the introductory part of paragraph 2 and paragraph 2.2 to Enrico Orsini, paragraph 2.3 to Armando D'Aniello, paragraph 4 to Damiano Abbatini and the Appendix to Armando Daniello, Enrico Orsini and Andrea Pagano.

References

Hornik, K. (2005). "A CLUE for CLUster Ensembles." Journal of Statistical Software, 14(12). doi:10.18637/jss.v014.i12.

Istat, (2023), https://www.istat.it/it/files//2023/12/NOTA-TECNICA-CENSIPOP_2022.pdf

Kuhn, H.W. (1955). The Hungarian method for the assignment problem, in Naval Res. Log. Quart, vol. 2, 83-97

Papadimitriou, C. and Steiglitz, K. (1982). Combinatorial Optimization: Algorithms and Complexity. Englewood Cliffs: Prentice Hall.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>