Statistical Journal of the IAOS 39 (2023) 123–136
DOI 10.3233/SJI-220113
IOS Press

Enumeration area imputation methods for producing sub-municipal data in the Italian permanent population and housing census

Giancarlo Carbonetti*, Giampaolo De Matteis, Marco Di Zio, Davide Fardelli, Raffaele Ferrara and Fabio Lipizzi
Italian National Institute of Statistics, Rome, Italy

Abstract. Over the years, official statistics have shown an increasing territorial focus on providing detailed and quality information. The Population and Housing Census has always ensured the availability of sub-municipal data useful for social, economic, and environmental decision-making processes. The new Italian Permanent Census focuses heavily on the integration of administrative and sample data and plans to provide more stable and consistent statistical data at the various territorial levels every year. Within this framework, sub-municipal data are derived from the integration of the Base Register of Individuals and the Base Register of Places. Data accuracy depends on the quality of the registers and the procedures adopted to integrate and process the input data. In this regard, Istat is working to improve geocoding information and linking procedures. One of the problems encountered is the presence of non-geocoded units due to problems in the administrative data. Istat has studied a procedure that integrates deterministic and probabilistic approaches to assign the enumeration area code to these critical units. It was conducted an experimental study to assess the quality of the imputation procedure. In this paper, we discuss the approach adopted, the evaluation process, the results obtained, and the impact on data quality.

Keywords: Population census, registers, geo-coding, enumeration area, missing data, geo-imputation

1. Introduction

One of the main objectives of the Census is to provide statistical information at the municipal and sub-municipal levels, which is in high demand by the research community, the private sector, and public administrators. In recent years, there has been a transition in Italy from the traditional census to a strategy based on the integration of statistical registers, built mainly by integrating administrative data, and sample data. This has required an enormous effort on the part of the Italian Institute of Statistics (ISTAT) to identify the best possible methodological and IT solutions to guarantee high levels of data accuracy, consistency, and timeliness [1]. In this context, the reliability of the results depends largely on the quality of the administrative data supplying the registers. This is something on which there is still a lot of work to be done.

In this article, reference is made to the spatial information in the archives that, in some situations, does not allow for the localization of all statistical census units in a sub-municipal area, thus preventing the production of data at maximum spatial detail. In past censuses, the geocoding operation took place during field data collection; the enumerator, when submitting the census questionnaire, recorded the household's home address and the code of the enumeration area[1] (EA) to which it belonged. This now occurs in the data processing phase,

*Corresponding author: Giancarlo Carbonetti, Istituto Nazionale di Statistica (ISTAT), via Cesare Balbo, 16, 00184, Rome, Italy. Tel.: +39 6 4673 4440; E-mail: carbonet@istat.it.

[1]Enumeration areas are geographical areas of defined spatial dimensions assigned to enumerators for population and housing census operations. They draw a partition of the municipalities' territory and may consist of maps or lists of streets or dwellings/buildings.

not without errors of missing or incorrect geocoding of census units (individuals, households, dwellings, buildings) in the EAs.

The accuracy of geocoding can certainly be increased by requiring source owners to pay more attention to quality and by working in the preliminary stages of processing the data once they have been acquired. On the other hand, specific imputation procedures (deterministic and probabilistic) have been developed to retrieve the geocoding of units without EA, exploiting geographical information, past census data, and data contained in archives.

The article will explain the imputation procedures, experimental tests, the analysis process, the results obtained, and will evaluate the impact of the procedures on the results. Finally, the validation process of some census results produced at the sub-municipal level will be illustrated, a preliminary operation to the official dissemination of the final data.

2. The demand for statistical information at the sub-municipal level

Over the years, there has been an increasing demand for sub-domain data from different actors and different purposes. National and international institutions, the academic and research world, and the private sector have a continuous need for highly disaggregated spatial data to conduct spatial analyses, for their business objectives and to make social, economic, and environmental decisions.

European legislation requires the production of data per 1 km^2 cell (Grid); municipalities need statistical information at different sub-municipal territorial levels (administrative areas; enumeration areas); users submit requests to Istat for customized data processing per EAs [2]. Moreover, data users are becoming increasingly demanding as, thanks to the high specialization of the statistical methodologies available, the use of GIS tools, and the possibilities of integration with other statistical sources, they conduct studies or define increasingly advanced projects that require large volumes of data and very fine spatial detail.

The most frequently requested data concern the presence and characteristics of foreigners, the educational level of the population, current activity status, position in the profession, type of occupation or sector of economic activity of the employed, family types, and structural characteristics of dwellings.

The Census of Population and Housing is the only statistical survey able to guarantee the availability of data at such a high territorial detail. The transition from the traditional decennial Census in Italy to the new permanent Census has required a careful review of the strategy to ensure census results that allow comparability with the past and adequate availability of sub-municipal domain data. The statistical methodologies and IT procedures used in the Permanent Census strategy are constantly evolving, as is the supply of sub-municipal data in quantitative and qualitative terms.

3. The permanent census strategy

Since 2018, Istat has been conducting the Permanent Census of Population and Housing. The census has moved from a traditional to a combined approach that integrates data from administrative sources and two sample surveys on households appropriately designed for the objectives of the annual census production: the population count and the calculation of contingency tables related to the main census variables.

Appropriate methodological and computer architectures enable the integration of information of register data and data collected on the sample households to produce census results at all geographic levels, from national to municipal.

The production of census data at the sub-municipal level requires the integration of two registers prepared by Istat: BRI – Base Register of Individuals[2] and BRP – Base Register of Places[3] [3]. Through the operation of "linkage" between those registers, it is possible to locate individuals and households on the territory and, in particular, in EAs [4]. In the past, this operation took place in conjunction with field data collection; with the new strategy of the permanent census, the geocoding of statistical units at EAs is the result of the integration of BRI and BRP, and of course missed geocoded units[4] may arise. This critical issue is mainly due to the unavoidable structural deficiencies in the territorial archives.

To be able to produce data at the sub-municipal level, Istat has developed some solutions applicable in a previous stage to the linkage operation to improve the outcome of the integration process between BRI and BRP,

[2]The BRI contains information on some demographic variables such as sex, place and date of birth, citizenship, and place of residence from administrative data.

[3]The BRP contains addresses, enumeration areas and, if possible, geographic coordinates.

[4]4.4% of BRI units (as of 31/12/2019), about 2.6 million, were non-geocoded.

Table 1
Geocoded and non-geocoded population by type of error

Population	Number (in millions)	%
Geocoded	56.6	95.6
Non-geocoded	2.6	4.4
Without house number in population registers	1.1	1.9
With an unrecognized house number (non-existent or wrong)	1.1	1.9
Without coordinates or enumeration areas code	0.1	0.2
Street non-existent or not recognized	0.3	0.5
Total	59.2	100.0

Source: Base Register of Individuals (BRI) as of 31.12.2019.

and specific procedures, applicable at an ex-post stage, aimed at processing the (residual) units that result without geocoding. This made it possible to define a process that led to the complete geocoding of all units (households and individuals) to EAs. The paper illustrates the proposed imputation methodologies and the experimental study that allowed the definition of a procedure for the recovery of the EA of non-geocoded units.

4. Critical issues for the production of sub-municipal data

As mentioned in Section 3, BRP plays a crucial role in the process of producing census results as it allows the geographical localization of statistical units belonging to Istat registers and surveys.

The construction of BRP is a very complex and innovative process [5]. It consists of four components:

- Census geographical database (enumeration areas and micro-zones);
- Addresses and geographical coordinates;
- Residential buildings and housing;
- Administrative and Statistical Territorial Units.

All these components contribute to the geo-coding of units to produce the sub-municipal data of the permanent census. However, public administrations, as providers, still pay little attention to spatial information in the data sources used. In particular, the quality of the "address string" of individuals residing in the Population Registers (PR) of the municipalities, the main source supplying the BRI, is not high.

The main operation performed to improve the quality of address strings is to subject them to a normalization process with commercial software. The software also returns the geographical coordinates of house numbers, if any, contributing to the generation of the BRP. This process allows the addresses to be geocoded and the two archives BRI and BRP to be linked through a specific "unique address key" (CUI).

In this context, the sources of possible errors are mainly due to:

1) the sources supplying the registers;
2) the register generation process;
3) the linkage process between BRI and BRP.

These critical issues are also reflected in the statistical units of the BRI, causing cases of missing or incorrect geocoding of individuals.

Table 1 shows the absolute values and percentages of the census population as of 31.12.2019. It focuses on approximately 2.6 million non-geocoded individuals. The errors for this residual group are due to non-geocoding. The main unresolved cases include individuals with an address without a house number from the PR or with a house number not recognized by the normalizer because it does not exist or is incorrect. In both cases, the number of individuals is about 1.1 million, or 1.9% of the total number of individuals. Finally, the last two lines concern individuals without geographical coordinates or enumeration area, i.e. 0.2%, and those who, although they have an address, the address is not recognized by the normalization software or is non-existent. The percentage of these cases is 0.5% of the total number of individuals.

A statistical process of checking and correcting the "consistency" of units of different populations (individuals, households, dwellings, buildings) was recently started in Istat. This should make all statistical units that belong to the same sub-municipal domain geographically consistent. Thus, each individual finds a household residing in the dwelling of a single building and this building should be correctly geocoded. These considerations contrast, however, with what has already been said about possible sources of error. A further critical point is the quality of the cadastral source used by the buildings register. To reduce these problems and to improve the final quality of the register, additional sources from the open world and from the Italian regions that produce cartography are also used.

5. Process improvement actions

To overcome the illustrated criticalities that do not allow the production of sub-municipal census data with acceptable levels of quality, Istat has defined several interventions and made some methodological choices to improve the statistical process that concern:

- the coverage and accuracy of sources, including ad-hoc surveys of municipalities;
- the process of address recognition and CUI assignment;
- the process of assigning geographic coordinates and EA to house numbers, including through the implementation of other open sources;
- the improvement of the record linkage operation between BRI and BRP;
- the integration of housing and building sources.

To improve the quality of spatial information in the archives, the list of unrecognized and non-geocoded addresses was sent to the municipalities. Municipalities were asked to correct errors, complete missing data, and update the information in their archives.

Procedures to improve the address recognition phase and the linkage operation between the BRI and BRP components were also implemented using other sources [6]. This made it possible to increase the level and quality of geocoding.

Nevertheless, there still remains a percentage of non-geocoded units for which specific EA retrieval procedures have been defined. These techniques described in the following sections will be implemented in the future in the final sub-municipal data production process.

6. Imputation procedures for non-geocoded units

The literature on the imputation of missing addresses (also known as geo-imputation) is sparse. Applications on Census data and health studies can be found in [7–9] and references therein. Generally speaking, there are two types of geo-imputation strategies: Stochastic and deterministic. Cases are deterministically assigned to locations using deterministic processes, which follow a set of principles. They are adopted when there is a high degree of reliability on rules determining the imputations. Their advantage is the computational efficiency, and the clarity at the basis of the imputations. On the other hand, when decisions on imputations are not clear enough, it is better to resort to probabilistic imputation methods. In this setting, uncertainty can be taken under control and evaluated by resorting to the tools typical of a probabilistic setting.

In practice, as in other NSIs' imputation procedures, the two approaches are combined. Generally, the first step is carried out by resorting to deterministic methods, then the remaining missing data are imputed through probabilistic methods.

This is the approach adopted for the EA imputation in our application. In particular, the results of the deterministic procedures are jointly used as follows; the deterministic imputed EA is the value with the highest frequency obtained by the deterministic procedures. For instance, if two or more procedures give the same EA, this will be the imputed value. When all procedures provide discordant EAs, the value corresponding to the most reliable method is selected for imputation.

In the following sections, we detail the deterministic and imputation methods, and the experiment performed to assess their quality and to design the overall imputation procedure.

6.1. Deterministic methods

Different deterministic methods are proposed that exploit information from the household, the last traditional census, the real estate properties, and the geographical spatial coordinates.

6.1.1. Family reconstruction (FR)

This method retrieves the EA of individuals found to be missing considering the family to which they belong. Within a household, the EA of the geocoded members is assigned to the non-geocoded household components. This procedure is not applied when the geocoded members have discordant EAs, and households have an anomalous high number of individuals.

6.1.2. Spatial interpolation (SI)

This method aims to assign the EA to the population that lacks it but has an address, including house number, for which the location on the territory is unknown. There are several causes attributable to this problem, among the most frequent is that of new established addresses by municipalities, for which geographic coordinates are not yet available within the databases used.

SI bases its foundation on the concept of geographic proximity. Specifically, an address is assumed to have a high probability of being in the same EA as a neighbouring address. In Italy, addresses have house numbers, and for the population without EA, the distance in terms of house numbers of the address without a geo-

graphic location was measured with the nearest address whose location is known. The latter value is used for the imputation of EA. We notice that, since house numbers are even and odd numbers according to the side of the street (right and left side of the street), the distance between even and odd numbering was also taken into account when measuring distance.

6.1.3. Address strings from the 2011 census (AD11)

This method is based on a comparison of the address strings associated with individuals without an EA, with the address strings in the 2011 Census data. When the linkage takes place, the EA associated with the address in the 2011 Census archives is retrieved, after verification on the municipality to which it belongs.

More in detail, the retrieval by strings of the 2011 Census addresses, as a whole, is carried out in three successive steps. The logic of the retrieval involves gradually less restrictive linkage constraints.

The first step of the retrieval consists of recognizing, within the 2011 Census archive, the EA associated with the address string. Retrieval is made possible only for addresses found to be associated in the municipality with only one EA (uniqueness of correspondence between the address and EA).

The second step introduces the individual's identification code into the linkage key. If the individual is the same, lives in the same municipality as in 2011, and at the same address, it is still reasonable to impute the same EA as in 2011, even when the linkage between the text strings of the two addresses is not fully and completely realized.

The third and last step is similar to the first one but makes use of the individual code as a linkage key. Less detailed information of the address string is used It is composed of the "name" of the street and the "house number" of the address, but the "species" element is excluded. In this way, it is possible to overcome failures of recognition related to:

1) the lack of the element of "species" in either of the two textual addresses considered (e.g., "WITHOUT SPECIES TURRITA LOC. SCATTERED HOUSES 67" is considered equivalent to "FR. TURRITA LOC. SCATTERED HOUSES 67");
2) to the presence of different ways of abbreviating the "species" (e.g., "CONTRADA SAN SALVATORE 77" is considered equivalent to "CDA SAN SALVATORE 77," to "C.DA SAN SALVATORE 77" and to "CONTR. SAN SALVATORE 77");
3) to temporal changes of "kind" in the address (e.g., "PIAZZA GIUSEPPE MAZZINI 11" is considered equivalent to "PIAZZETTA GIUSEPPE MAZZINI 11").

6.1.4. Real estate property (REP)

This method involves the use of information about the real estate property that the individual is reported to own. The source of information on real estate properties in Italy is the *"Catasto delle Unità Immobiliari dell'Agenzia delle Entrate" (Cadastre of Real Estate Units)*. Istat places the real estate from this administrative data source in an EA. Owners are imputed with the EA associated with the house, provided that some conditions are fulfilled. The real estate unit must be owned by the individual in the data year and the municipality of residence or tax domicile. In cases the individual owns more than one real estate unit in the municipality for the given year, de-duplication steps are used.

The result is extended to family members who still lack an EA. The extension of EA to family members is relevant numerically, as not all members of a family own the housing unit in which they live.

6.1.5. Real estate rentals (RER)

This method proceeds similarly to the previous one. The difference is only in the relationship between the individual and the property. In the previous one, the property is owned, in this case, it is a rented property. Again in the presence of more than one lease, de-duplication is used for the above considerations.

The results obtained are then extended to family members still without an EA using the method that retrieves the EA through the family. The extension of the EA to family members is also important here since not all members of a household hold a lease related to the housing unit in which they live.

6.2. Probabilistic imputation

At the end of the deterministic imputation procedures, a certain number of units are still without EA. For those units, a step of probabilistic imputation is performed.

The probabilistic imputation is composed of a sequence of donor imputation steps mainly characterized by different imputation cells. Donor imputation steps are performed by choosing at random an observed household in the imputation cells and then by imputing the corresponding EA to a household with missing information in the same cells, that is with the same characteristic of the variables determining the imputation cells. At each step, the imputations are applied to the units not imputed in the previous one.

EA is imputed to the household following the principle that all the individuals of a household should be in the same EA.

The sequence of imputations steps is:

Table 2
Imputations by methods and concordances with observed EAs and ADAs

Deterministic methods	Absolute and percentage frequencies of imputations		Absolute and percentage frequencies of imputations concordant with observed EA		Absolute and percentage frequencies of imputations concordant with observed ADA	
RER	325,505	9.3%	179,232	55.1%	255,195	78.4%
REP	1,772,574	50.8%	1,468,139	82.8%	1,612,729	91.0%
AD11	2,922,393	83.8%	2,884,880	98.7%	2,899,427	99.2%

Source: the results of the preliminary experiment conducted on four municipalities. Istat. Reference date: December 31, 2019.

I1. Donor imputation within cells determined cross classifying Street, EA in the 2011 Census;
I2. Donor imputation within cells determined by Street;
I3. Random choice of an EA belonging to the street of the non-geocoded household;
I4. Donor imputation within cells determined by the EA in the 2011 Census;
I5. Donor imputation through a random choice of an EA attached to an observed household.

The variable "*street*" reports the name of the street where the household is resident. We remind that this is an important information, but it is not sufficient to determine the EA because it lacks the exact location in the street (house number) and it can cross two or more EAs.

Step 3 is a bit different, in fact, a household in a "street" is imputed taking at random an EA where the street can be located. The random draw is weighted with the frequency of addresses in the EAs. For instance, in case a street crosses two EAs, but in an EA there are 2 addresses (streets and house number), and in the other 8 addresses, the weight of the random imputation is given by those two frequencies, which means that the first EA will be chosen with a probability equal to 2/10 and the other 8/10.

The characteristic of those methods is that of reproducing the frequency distributions of the EA observed within the imputations cells [10]. For example, in step 1, for a household that is in a specific *street* and that was in a specific EA in the 2011 census, the method reproduces the behaviour (the frequency distribution) of the units that are in the same street and that were in the same EA in 2011.

7. Experimental study of the imputation procedures

Some experimental studies are carried out to design the overall imputation procedure, which should combine a set of deterministic and probabilistic imputation methods. The procedures are applied to a preliminary version of actual data from some Italian municipalities to assess their quality.

The evaluation of the methods is a crucial and critical point since we do not know the true EA value. Hence, we need to resort to some indicators providing an indirect measure of their accuracy.

For deterministic procedures, a subset of units considered particularly reliable in terms of observed EA is selected. Deterministic procedures are applied to these units and the imputed EA is compared with the observed one. In addition, concordances between imputed and observed values are also assessed for particular aggregations of EAs, i.e., administrative areas[5] (ADAs). Thus, both concordances between EAs and ADAs are analyzed.

The municipalities selected are *Rome*, *Genoa*, and *Palermo*, which are particularly important, and on the opposite side a very small municipality, *Lacco Ameno*. Table 2 reports the imputations by method and the concordances with observed EAs and ADAs.

REP and AD11 provide good results both for EA and ADA. RER results in good imputation for ADA, but is quite poor for EA.

For the spatial approximation method SI, it is not possible to conduct the same type of analysis. Therefore, attention is focused on the subset of units where there is a prevalence among the predictions of four methods SI, AD11, REP, and RER; the prevalence being the case where there is a majority of methods that reconstruct the same EA or ADA. The behaviour of the spatial approximation method is analyzed assuming that the prevailing value is with higher confidence the correct value of the EA. On this subset of units, it is computed the number of cases where SI provides the prevalent value (SI included) and the number of times the SI imputation is not the chosen value (SI not included).

[5] Sub-municipal areas, that correspond to an administrative geographical classification and consist of the aggregation of contiguous EAs.

Table 3
Classification table of SI with threshold less than 10 house numbers

	EAs		ADAs	
	Freq.	%	Freq.	%
SI included	2,233	93.6	2,855	100.0
SI not included	153	6.4	0	0.0
Total	2,386	100.0	2,855	100.0

Source: the results of the preliminary experiment conducted on four municipalities. Istat. Reference date: December 31, 2019.

Table 4
Classification table of SI with threshold greater than 10 house numbers

	EAs		ADAs	
	Freq.	%	Freq.	%
SI included	265	77.7	566	99.1
SI not included	76	22.3	5	0.9
Total	341	100.0	571	100.0

Source: the results of the preliminary experiment conducted on four municipalities. Istat. Reference date: December 31, 2019.

Table 3 and 4 show the frequencies for the SI method with thresholds of up to 10 house numbers and over. We observe that the totals for EAs and ADAs change because the reference subset is different if the prevalence analysis concerns EAs or ADAs.

In both cases, SI has a very high level of correct classification of ADA, while with the threshold over 10 house numbers, although still quite good, has a non-negligible level of misclassification of EA.

For the analysis of probabilistic imputation methods, it is necessary to evaluate the variability of the imputation uncertainty. This step is essential since the choice underlying their use in this procedure is because there is not a high confidence in a deterministic reconstruction method. Consequently, the choice of a probabilistic method has the advantage of not "forcing" the choice in some directions, thus possibly introducing a bias. However, this entails the need to evaluate the reconstructions in probabilistic terms, that is, to make an evaluation taking into account all the possible solutions and considering their probability.

It is important to evaluate summary measures such as the coefficient of variation and the confidence intervals using the variability of the counts of people for each EA. To this end, it was adopted a replication approach, i.e., the imputation was repeated 100 times. On the 100 possible results, the coefficient of variation (CV) for the count of people for each EA and their confidence intervals are computed.

Table 5 shows the summary statistics of the distribution of the CVs over EAs. In the same table, the average width of confidence intervals (CI) computed over EAs is reported. CVs are in general almost very low, that is, the people count estimates in the EAs have a very high level of precision. However, there are still some high maximum (max) values, e.g. 0.66 (66%). These values are not always worrying because the number of individuals in EAs in some cases is very small and the variation is reduced to a few units, sometimes it is even due to decimal differences (these can exist because the algorithm averages several integer values).

For Genoa, the EAs with a CV greater than 5% are 9 out of a total of 3317. Out of them, the largest confidence interval is [18, 22], thus showing good performances for those units.

For Rome, EAs with a CV greater than 5% are 164, corresponding to 1.4% of the total number of EAs. Table 6 shows summary statistics of the distribution of the width of their confidence intervals.

Also in this case, the performance seems good in terms of count estimates. Similar behaviour is observed for the other two municipalities.

In conclusion, based on these first assessments, deterministic methods give good reconstructions, particularly concerning to ADA. Probabilistic methods allow for the assessment of allocation uncertainty, and this assessment indicates low variability, due essentially to the low number of units to be reconstructed probabilistically.

8. Overall imputation procedure and results of its application

In the second part of the design of the procedure, according to the preliminary experiment described in Section 7, an overall imputation procedure is defined and applied to a larger and more updated set of data.

8.1. Integrated imputation procedure

The results of the previous experiment led us to design and test the overall imputation procedure composed of the following steps:

1. Family reconstruction (FR) is applied.
2. SI, RER, REP, and AD11 methods are independently applied.
3. Missing units are imputed with the prevailing value obtained in step 2 (including the case of only one available prediction).
4. Units without a prevalent EA are imputed with the value obtained by the methods ranked according to the following hierarchy:
   a. SI with less than 10 house numbers

Table 5
Summary statistics of the CVs and width of CIs for the count of people computed over EAs

	Summary CV						Summary CI width 95%					
	Min	1st qrt	Median	Mean	3rd qrt	Max	Min	1st qrt	Median	Mean	3rd qrt	Max
Genoa	0	0	0.001	0.003	0.002	0.099	0	0	0	0.65	1	14
Rome	0	0.001	0.001	0.005	0.004	0.662	0	0	1	1.47	2	25.52
Lacco Ameno	0.004	0.006	0.008	0.017	0.012	0.069	5	9.8	13.5	12.5	46.2	18.6
Palermo	0	0	0.001	0.003	0.002	0.476	0	0	0	0.8	1	36

Source: the results of the preliminary experiment conducted on four municipalities. Istat. Reference date: December 31, 2019.

Table 6
Summary statistics of the distribution of width of CIs for the EAs with CV $>$ 5% in Rome

Min	1st qrt	Median	Mean	3rd qrt	Max
0	1	4	3.9	5.6	16.5

Source: the results of the preliminary experiment conducted on four municipalities. Istat. Reference date: December 31, 2019.

Table 7
Distribution of the Italian municipalities by geocoding rate classes of individuals

Geocoding rate	Municipalities		
(classes)	List	Selected	%
$\geqslant$ 95 %	5,543	32	0.6
90% $\vdash$ 95%	1,107	4	0.4
85% $\vdash$ 90%	494	5	1.0
80% $\vdash$ 85%	249	3	1.2
$<$ 80%	521	19	3.7
Total	7,914	63	0.8

Source: the result of the linkage of the Base Register of Individuals (BRI) and the Base Register of Places (BRP). Istat. Reference date: December 31, 2019.

b. AD11
c. RER
d. SI with threshold house number between 10 and 50
e. REP

5. On the remaining units, probabilistic imputation as described in Section 6.2 is performed.

8.2. Selection of municipalities for testing

To assess the impact of the general imputation procedure on the data, is applied to a larger subset of municipalities. From the list of 7,914 Italian municipalities, a set of municipalities representative of the different population sizes, the different geographical areas of Italy (North-West, North-East, Centre, South, and Islands) and the different geocoding rates of the missing EAs are selected.

The data set is composed of 63 municipalities with around 12 million people, about 20% of the Italian population. All municipalities (25 in total) for which administrative areas have been established are included (see Note 5).

Table 8
Distribution of the Italian municipalities by different geographical areas of Italy

Geographical	Municipalities		
area of Italy	List	Selected	%
North-west	2,996	17	0.6
North-east	1,397	14	1.0
Centre	971	10	1.0
South	1,783	11	0.6
Islands	767	11	1.4
Total	7,914	63	0.8

Source: Istat.

Table 9
Distribution of the Italian municipalities by population size classes

Population size	Municipalities		
(classes)	List	Selected	%
$\geqslant$ 100,000	45	26	57.8
50,000 $\vdash$ 100,000	103	5	4.9
10,000 $\vdash$ 50,000	1,080	11	1.0
2,000 $\vdash$ 10,000	3,199	12	0.4
$<$ 2,000	3,487	9	0.3
Total	7,914	63	0.8

Source: Istat. Reference date: December 31, 2019.

In the first phase, the number of municipalities to be selected from each geocoding level class was defined (five classes: $>$ 95%; 90%–95%; 85%–90%; 80%–85%; $\leqslant$ 80%) to better represent those with the lowest percentages of geocoded units (Table 7). Subsequently, the municipalities were selected for experimentation to fairly represent the different Italian geographical areas (Table 8) and population size (Table 9), with the constraint of including all municipalities with administrative areas.

The set of 63 municipalities selected for the experimentation includes 62,588 EAs. The outcome of the linkage between BRI and BRP (with the population as of December 31, 2019) did not geocode 285,035 units. The database for the 63 municipalities was prepared and the tests of the complete procedure were carried out. The following sections present some descriptive results and some analyses of the accuracy of the experimental results.

Table 10
Distribution of imputations with the deterministic method by the different procedures

Deterministic procedures	Deterministic imputations	
	Frequency	%
FR	4,040	2.1
SI	41,011	21.2
AD11	39,217	20.2
REP	103,560	53.5
RER	5,885	3.0
Total	193,713	100.0

Source: the result of the application of the overall imputation procedure on the 63 municipalities selected for the experimentation. Istat. Reference date: December 31, 2019.

Table 11
Distribution of imputations with the probabilistic method by the different steps

Probabilistic stages	Probabilistic imputations	
	Frequency	%
I1	38,523	42.2
I2	27,289	29.9
I3	3,767	4.1
I4	21,201	23.2
I5	542	0.6
Total	91,322	100.0

Source: the result of the application of the overall imputation procedure on the 63 municipalities selected for the experimentation. Istat. Reference date: December 31, 2019.

8.3. Results of the experiment

8.3.1. Descriptive results

The application of the procedure described in Section 8.1 made it possible to impute the EA of all 285,035 non-geocoded individuals observed in the 63 test municipalities. Specifically, 68% (193,713 units) with deterministic methods and 32% (91,322 units) with probabilistic methods.

Since the procedure integrates methods with different levels of quality, it is useful to see the imputation by methods. Table 10 reports the distribution of cases imputed by the deterministic method for the different adopted procedures. Table 11 shows the distribution of cases imputed with the probabilistic method for the different imputation steps.

Among the deterministic methods, more than 53% of the imputations occur via the EA associated with the dwelling-owned address (REP method). On the other hand, the spatial method (SI) and the method based on residence address in the 2011 census (AD11) retrieve the missing EA in 21.2% and 20.2% of cases, respectively.

Using the probabilistic method, donor imputation based on the street in the first two steps allows the imputation of the missing EA in more than 72% of cases. Among the other steps, the fourth based on the EA observed in the 2011 census allows imputation in a further 23% of cases.

Table 12
Distribution of imputations with the deterministic method according to the different modes of application of the overall procedure

Imputation modes in the deterministic proceture	Deterministic imputations	
	Frequency	%
Family Reconstruction (FR)	4,040	2.1
Concordance of 4 methods	36	0.0
Concordance of 3 methods	8,745	4.5
Concordance of 2 methods	23,960	12.4
1 method	135,149	69.8
Hierarchy of methods	21,783	11.2
Total	193,713	100.0

Source: the result of the application of the overall imputation procedure on the 63 municipalities selected for the experimentation. Istat. Reference date: December 31, 2019.

In both cases, we notice that most of the imputations are obtained by the most reliable methods, ensuring a high level of quality of the imputed EA.

As explained in Section 8.1, in step 2 of the deterministic procedure, the four methods (SI, AD11, REP, and RER) are applied independently. The different methods can either return a value to be imputed or fail. If at least two values are equal, the prevailing value is imputed in step 3. If the values are different, in step 4 the value is imputed following a hierarchy between the methods. Table 12 shows the results of the different methods of applying the deterministic procedure carried out in the experiment (the first row of the table shows the number of imputations by the FR method that occurred in the stage before the application of the four deterministic methods).

The criterion of concordance between methods allows imputation in about 17% of the cases. On the other hand, in about 70% of the cases, the assignment of the missing EA occurs due to the implementation of only one method (3 out of 4 do not return any value). Finally, in 11.2% of the cases, since more than one method returned different values, the assignment of the EA occurs based on the hierarchy between the methods (see step 4 in Section 8.1).

Another analysis concerns the use of deterministic and probabilistic imputations for each EA (Table 13).

Considering only the EAs to which at least one new unit was assigned (14,715, i.e. 23.5% of the total EAs), in more than 75% of the cases, the EA is retrieved according to only one of the two approaches: 39.1% of the EAs only with the deterministic methods; 36.3% of the EAs with the probabilistic approach. The remaining 24.6% of EAs are assigned units with both methods.

Table 13
Distribution of imputed EAs according to the different possibilities of applying the methods

Application methods	EAs	%
EAs imputed only by deterministic method	5,750	39.1
EAs imputed only by probabilistic method	5,349	36.3
EAs imputed with both methods	3,616	24.6
Total	14,715	100.0

Source: the result of the application of the overall imputation procedure on the 63 municipalities selected for the experimentation. Istat. Reference date: December 31, 2019.

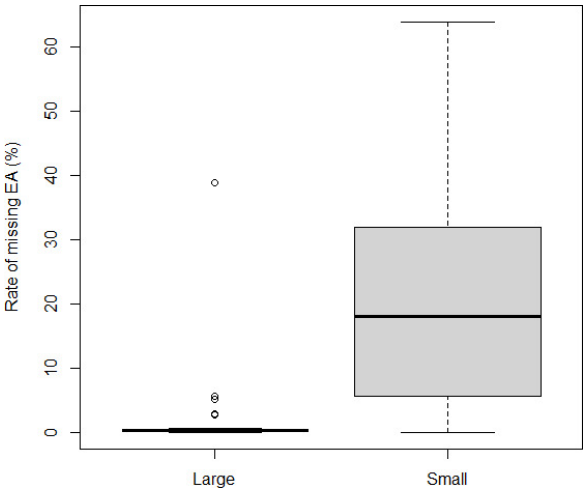

Fig. 1. Boxplot of the missing EAs expressed in percentage in large and small municipalities.

8.3.2. *Analytical results*

First, we notice that the rate of missing EAs is related to the size of municipalities: the average value of missing EAs in municipalities with more than 100,000 inhabitants is 2.4%, while it is 16% in the others. This is observed in the boxplot of the missing EA rate in percentages for small municipalities (up to 100,000 inhabitants) and large municipalities (over 100,000 inhabitants) in Fig. 1. Among large municipalities, only three have a missing EA rate above 5%, and among them, one has a particularly high value (38.4%). This very anomalous behaviour suggests a systematic problem with administrative data sources and deserves a clerical approach to understand the cause, instead of an automatic imputation.

Figure 2 depicts the boxplot of the median CV and the median width of 95% confidence intervals in large and small municipalities.

We notice a general high precision of the estimator and very narrow confidence intervals. Only two municipalities have an average error above 1%. They are affected by a high level of units with missing EAs, in fact,

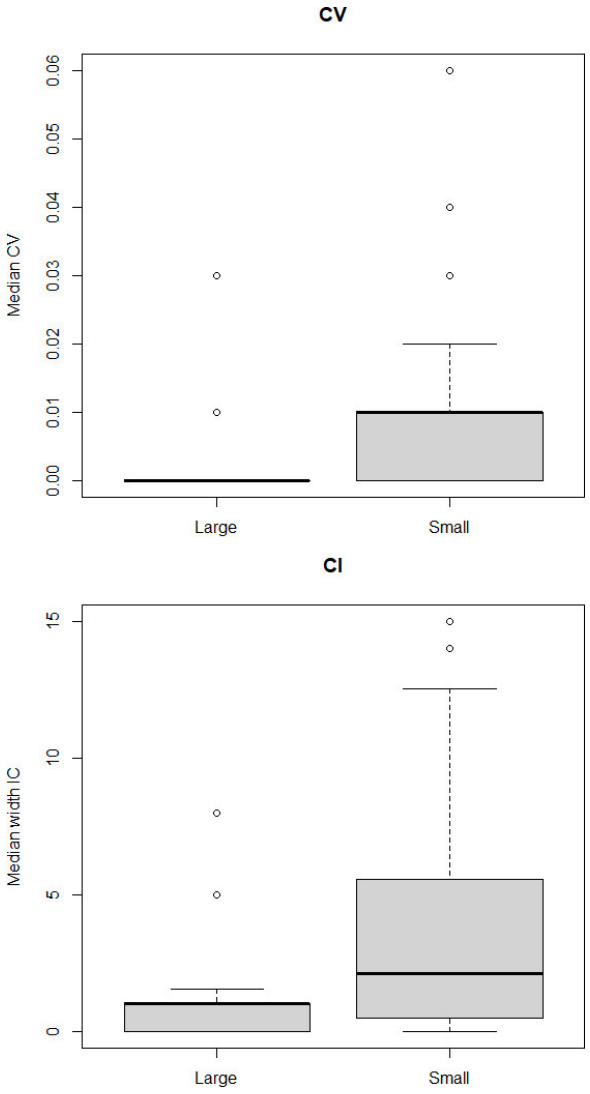

Fig. 2. Boxplot of the median CV and median width of 95% Confidence Intervals of population counts of EAs in large and small municipalities.

they have 5.2% and 38.8% missing EAs respectively, while the average of missing EAs in all municipalities considered is around 2%.

The rate of probabilistic imputation is higher in large municipalities (see Fig. 3).

9. Impact of imputation methods on the enumeration areas

This section analyses the consequence of the tested imputation methods applied to the cases of non-geocoding and thus the allocation of the correspond-

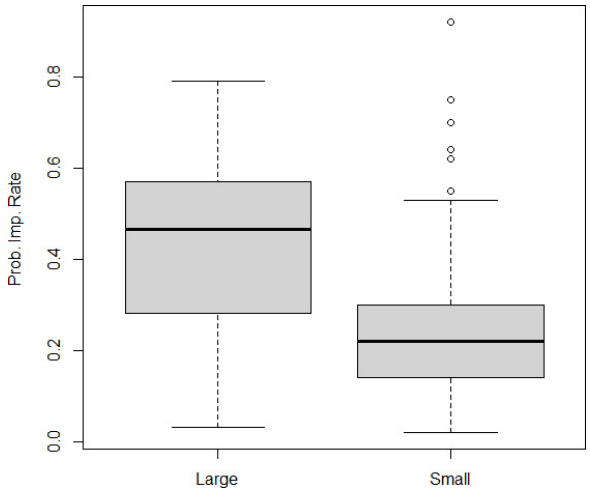

Fig. 3. Boxplot of the rate of probabilistic imputation in large and small municipalities.

ing individuals to EAs. Table 14 shows the distribution of the 62,588 EAs of the tested municipalities by the number of individuals assigned as a result of the EA imputation process.

It can be observed that 76.5% of the EAs (47,873) of the 63 municipalities tested remained unchanged as no new units were geocoded; in 12.8% of the EAs (7,990) a maximum of 2 new units were allocated; in 5.4% (3,386 EAs) between 3 and 10 units; in the remaining 5.3% (3,339 EAs) more than 10 new units. Some EAs with a very high number of allocations are also observed. Considering, for example, the 50 EAs with more than 500 allocations, in 68% of cases they belong to municipalities with a geocoding level $< 80\%$. These are, therefore, in most cases, EAs belonging to municipalities with low geocoding levels and which, therefore, required a strong quantitative commitment to the EA imputation procedure.

Finally, 506 EAs (less than 1%) that had no individuals before the application of the procedure were assigned at least one unit.

Another interesting analysis concerns the evaluation of the increase in units of EAs after the imputation process, trying to highlight cases (critical EA) where this increase is above a fixed percentage threshold.

Units without an EA may be allocated by the imputation process in a scattered manner on the municipality's EAs (a few units per EA) or in a more concentrated manner on a few EAs. In the first case, the impact on the number of individuals in the EAs will be very low, while in the second case it will be much more significant.

Table 14
Distribution of EAs by imputation number (in classes)

Imputations	EAs	%
0	47,873	76.5
1–2	7,990	12.8
3–5	2,115	3.3
6–10	1,271	2.0
11–20	999	1.6
21–50	1,083	1.7
51–100	561	0.9
101–300	540	0.9
301–500	106	0.2
501–1,000	46	0.1
$>$ 1,000	4	0.0
Total	62,588	100.0

Source: the result of the application of the overall imputation procedure on the 63 municipalities selected for the experimentation. Istat. Reference date: December 31, 2019.

Thus, after setting a threshold (e.g. 5%, 10%, or 20%), it is possible to calculate the number of critical EAs in which the percentage of allocated units is higher than this threshold. It is suggested to exclude EAs that, after the imputation phase, have a very small number of units (e.g. 10) and/or EAs where the number of new units is very low e.g. 5, due to the irrelevance of the numerical values.

Table 15 shows the number of critical EAs and the percentage related to the total number of EAs calculated according to different thresholds (5%, 10%, 20%) for some municipalities with different geocoding levels: Milan (99.9%), Rome (99.7%), Naples (97.1%), Florence (94.4%), Enna (81.9%) and Messina (61.1%).

From the values reported in Table 15, it can be verified that, for the three thresholds considered, the impact of the allocation in the EAs due to the geocoding imputation process for Milan and Rome is almost null, while for Naples it is quite acceptable. On the contrary, it is never acceptable for Messina. For Florence, there is a borderline situation only in the case of the 5% threshold. For Enna, even in the case of the highest threshold, there is an increase in the number of individuals of more than 20% in about 22% of the EAs.

This analysis could lead to the definition of a process indicator that refers to how individuals with the imputed EA are allocated between the EAs of the municipality.

10. The validation process of sub-municipal data

After the assignment of the non-geocoded units of the municipalities involved in the experiment, the census data for the enumeration area (EA) and administrative area (ADA) referring to the 2019 Census were

Table 15
Number and percentage of EAs in which a high proportion of individuals was allocated according to different thresholds (Th.), for some municipalities with different levels of geocoding (Population and Housing Censuses 2019)

Municipality	Geocoding level	Total EAs	EAs			Percentage		
			Th. 5%	Th. 10%	Th. 20%	Th. 5%	Th. 10%	Th. 20%
Milan	99.9%	5,707	0	0	0	0.0	0.0	0.0
Rome	99.7%	11,997	115	45	14	1.0	0.4	0.1
Naples	97.1%	3,859	264	170	117	6.8	4.4	3.0
Florence	94.4%	1,956	215	160	96	11.0	8.2	4.9
Enna	81.9%	85	57	22	19	67.1	25.9	22.4
Messina	61.1%	1,485	1,031	945	728	69.4	63.6	49.0

Source: processing by Census Department – Istat.

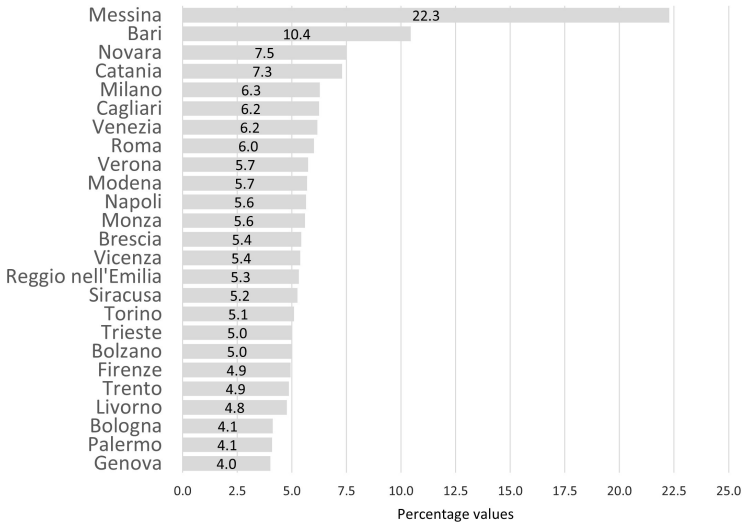

Fig. 4. Dissimilarity indices of the distribution of the census population by enumeration area for the largest Italian municipalities. Comparison between the 2019 Census and the 2011 Census.

determined for the 25 largest municipalities for which ADAs are available. The census variables produced at the sub-municipal level are:

- Population by gender and age group;
- Employed by gender and age group;
- Population by educational attainment;
- Foreign population by age group.

Although the above data have only an experimental character, they have been validated to have further elements to assess the goodness of the imputation procedures of non-geocoded units. The following validation checks were carried out for the data produced:

1. comparison of the distribution of the municipal population (and some of its components) for the ADAs of the municipalities in 2019 and at the 2011 population census;
2. comparison of the percentages of foreigners, the population aged 65 and over, the population with a university degree, and the employed population, in each administrative ADA in 2019 and at the 2011 census;
3. comparison of the distribution by EAs of the census population in 2019 and that in 2011 by calculating synthetic dissimilarity indices for each municipality;
4. identification of EAs with anomalous population values.

The analyses carried out shows that there are no significant differences between 2011 and 2019 in the distribution of the population between the different ADAs of the municipalities considered. The data produced are therefore characterized by robustness and a high degree of consistency compared to those of the last traditional census.[6]

[6]The 2011 population and housing census was the last census conducted in Italy in the traditional way, where individuals and households were allocated in enumeration areas at the same time as the census operations. The enumerators covered the entire municipal

Differences between the various components of the population (foreigners; population 65 years and over; graduates; employed) observed in some ADAs at the two reference dates do not seem significant. These results are consistent with the population dynamics found at the municipal level between 2011 and 2019.

The values of the dissimilarity index, which made it possible to compare the distribution of each municipality's population across its EAs in 2019 with that of 2011 (Fig. 4), are shown in percentage terms and express percentages of the population that are distributed differently across the territory compared to the initial reference year. Except for a few municipalities, and above all the municipality of Messina (whose territorial information is characterized by strong criticalities), the dissimilarity index appears very low and certainly plausible for the period considered.[7]

The last analysis carried out was aimed at identifying any EAs that might present significant anomalies in the population data. The identification of such EAs was carried out by combining several assessment elements: population changes since 2011 (considered in both absolute and relative terms), the percentage of people associated with addresses whose geo-coding to EAs is considered to be of high quality[8] within BRP, and the type[9] of EAs considered. This analysis also reported very few situations with possible serious anomalies in the demographic data (only 28 EAs out of almost 58,000 in all 25 municipalities considered), indicating the general reliability of the data produced.

In conclusion, the checks carried out did not reveal any particular problems in the data for the administrative sub-areas considered, as the data produced were robust and consistent with the 2011 population census. The data by EAs also appear to be reliable overall, although for some municipalities there is a clear need for further investigation.

territory, travelling through the enumeration areas and surveying all the population living within them.

[7]Taking the municipality of Rome as an example, the dissimilarity index value of 6 percent compared to 2011 is equivalent to about 150,000 people distributed differently across the territory compared to 2011. This quantity is roughly equivalent to the flow data on registry registrations and cancellations (for births, deaths and transfers of residence) observable in a single year on the municipal registry.

[8]The quality of geo-coding of addresses to enumeration areas is assessed by means of certain algorithms for matching information reported in different sources and administrative records.

[9]Enumeration areas are classified according to whether they are located in urban areas, small residential areas outside the municipality, rural areas or industrial zones.

11. Final remarks

The definition processes of the BRI and BRP are continuously evolving and, together with the improvement of the quality of the information entering these registers, a higher accuracy of the geo-coding operation of individuals and a reduction of non-geocoded units are expected.

Some of the problems come from the quality of geographical information in some of the administrative sources. Improving input data sources is essential to obtain high quality results. Some actions are engaged with the providers to improve the quality. An important initiative is that of providing a list of unrecognized and non-geocoded addresses to the municipalities. Municipalities are asked to correct errors, complete missing data, and update the information in their archives. In this task, municipalities may also make use of information collected through sample surveys in support of the census.

Nevertheless, it is useful to have a procedure for imputing missing enumeration areas since there will always remain a quota of non-geocoded units that will have to be processed for producing target estimates. In this paper, an EA imputation procedure based on the use of deterministic and probabilistic methods was illustrated, which enabled the complete assignment of all statistical units in the enumeration areas of the respective municipalities.

The success of the experimental phase and the outcome of the EA impact analyses convinced us to make the entire process of retrieving missing EAs structural in the process of producing sub-municipal census estimates. Subsequently, after the validation phase, the sub-municipal data can be disseminated as official results of the Permanent Population and Housing Census.

This work has therefore illustrated a procedure to be followed to overcome possible criticalities in statistical processes involving the integration of administrative and statistical sources. The production of official census data referring to small territorial areas requires a strong focus on the quality of the data used and the statistical and IT operations needed for the integration of sources and geocoding. This approach is fundamental for producing official data with a strong territorial value and with acceptable levels of accuracy for users who require them to conduct studies and research on the evolution of demographic, social, and economic phenomena referred to the territory.

References

[1] Falorsi S. Census and Social Surveys Integrated System. Note by the National Institute of Statistics of Italy. In: UN-ECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting. Geneva, Switzerland; 2017 October 4–6. Available from: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/WP23_ENG.pdf.

[2] Carbonetti G, Ciccarese A, Roncati R. Population and Housing Census. What are the users' needs. Review of Official Statistics. Istat; 2023 (To be published in 2023).

[3] Fardelli D, Orsini E, Pagano A. The address component of the Statistical Base Register of Territorial Entities. In: Book of Short Papers SIS. Pearson; 2021. pp. 1206-1211. Available from: https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/pearson-sis-book-2021-parte-2.pdf.

[4] Fortini M, Tuoto T. Probabilistic record linkage with less than three matching variables. In: Book of Short Papers SIS. Pearson; 2020. pp. 3-8. Available from: https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/Pearson-SIS-2020-atti-convegno.pdf.

[5] Crescenzi F, Lipizzi F. The integration of geographic and territorial data sources into the base register of territorial and geographical entities. Statistical Journal of the IAOS. 2020; 36(1): 143-149. doi: 10.3233/SJI-190586.

[6] Carbonetti G, Daddi S, De Matteis G, Di Zio M, Fardelli D, Ferrara R, Lipizzi F, Orsini E. New perspectives for the quality of sub-municipal data with the Italian permanent population and housing census. In: Book of Short Papers ASA Conference. 2022 (To be published in 2023).

[7] Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. International Journal of Health Geographics. 2008; 7(3): 1-10. doi: 10.1186/1476-072X-7-3.

[8] Curriero FC, Kulldorff M, Boscoe FP, Klassen AC. Using imputation to provide location information for nongeocoded addresses. PLoS One. 2010; 5(2). doi: 10.1371/journal.pone.0008998.

[9] Dilekli N, Janitz AE, Campbell JE, de Beurs KM. Evaluation of geoimputation strategies in a large case study. International Journal of Health Geographics. 2018; 17(1): 1-13. doi: 10.1186/s12942-018-0151-y.

[10] Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons, Hoboken; vol. 793. doi: 10.1002/9781119482260.