An innovative approach to improve the quality of the household and nuclei types reconstruction in Italy

Rosa Maria Lipsi[1], Anna Pezone[2]

[1]Istat, Rome, Italy, lipsi@istat.it

[2]Istat, Rome, Italy, pezone@istat.it

Abstract

Since 2018, the Italian National Institute of Statistics (ISTAT), as other European countries, moved from the traditional ten-year “door-to-door” census to a yearly “register-based” system (the Permanent Population and Housing Census) in order to produce annual detailed statistics, to enrich the supply & quality of statistical information, to reduce the statistical burden for respondents and the costs by the community. This transition represents a great innovation. However, every ten years, according to European regulations, EU Member States must send to Eurostat information on the main characteristics of their resident population and their social and economic conditions at national, regional and small areas levels, regardless of how they collected them.

A multisource approach, based on a combination of administrative data, registers (as RBI – Based Register of Individuals, RSBL – Statistical Base Register of Territorial Entities) and surveys data, has been used to provides information on Italian population and housing census for the 2021, as required by the EU regulation 2017/712.

The number of households and their characteristics is one of the mandatory information, but also one of the most complex aggregates to detect, validate and disseminate. The main problem to solve is the correct identification of households, as well as nuclei types. The reconstruction of the household in its internal composition is possible through the correction of individual variables as the relationship with the reference person, the age, the sex, the marital status, the year of marriage or civil union, analysed in relation to those of the other household members. In 2021, the most important set of the above variables becomes from ANPR (The National Register of Resident Population) that contribute to improve the quality of the RBI aiming to produce, at macro-micro level, official statistics on households. In order to obtain these statistics an efficient and efficacy strategy has been planned involving innovative generalized solution of E&I system and specific adaptations, for census demand, of the "Families Procedure" for the reconstruction of the household and nuclei types, usually used for social surveys.

Our goal is to describe the whole process to produce statistics on households and their characteristics by using RBI information, ANPR and survey data in order to highlight the main advantages of the innovative integration process and the quality of the data, by suggesting, for the future, how optimizing the process in terms of outcome, time and performance too.

Keywords: census, household and nuclei types, E&I process, data integration, data quality.

1. Introduction

Since October 2018, the Italian National Institute of Statistics (ISTAT) has been conducting the Permanent Population and Housing Census (PPHC), based on integration between registers and sample surveys, in line with European development policies and the ISTAT modernisation programme, in order to provide data representing the entire population, while reducing costs and response burden. The main register used for this purpose is the Base Statistical Register of Individuals (RBI) that gathered data from many administrative sources referring to people resident or not in Italy. The other register is the Statistical Base Register of Territorial Units and Addresses (RSBL). Furthermore, two census surveys, List-based survey (List) and Distribution range survey (Areal) were carried out, in 2018 and 2019, to produce data for variables not covered by registers and to estimate coverage errors of the RBI.

In 2020, during the pandemic emergency was not possible to carry out the field operations, so Istat (2022) produced the total amounts of municipal usual resident population, by gender, age and citizenship, using only the *signs of life* in the administrative sources (Integrated Archive of Usual Resident Population-AIDA), not included in the previous censuses.

RBI corrected with AIDA, for under and over coverage, originated RBI-CENS2020.

In 2021, Istat was again able to carry out census surveys (List and Areal), which integrated with RBI-CENS2021, by applying the methodologies used for RBI-CENS2020, provide the population count at 31st of December 2021 and the database for the production of census hypercubes, as required by the EU regulation 2017/712 (Eurostat, 2017).

Among the census hypercubes, the number of households and their characteristics is one of the mandatory information, but also a very complex aggregates to detect, validate and disseminate. The main problem is the correct identification of household and nuclei types, that requires the correction of individual and familial variables.

Our goal is to describe the whole process to produce statistics on the households and their characteristics, by using integrating data, in order to highlight the main advantages of the innovative integration process and the quality of the data, by suggesting future actions to improve the household reconstruction with the aim to minimize errors and optimize the procedure in terms of time and performance too.

2. Data and methods

2.1 Data

Data used to reconstruct the household and nuclei types are those of RBI-CENS2021, corrected for over (849,348 units) and under (149,059 units) coverage by integrating the RBI information with AIDA. For this reconstruction, the Italian Legal Population by household size, age, gender and citizenship was fixed. This population at 31st of December 2021, amounts to 58,678,795 people residing in 26,206,246 private households (Table 1).

2.2 Methods

For the household and nuclei types reconstruction, the variables used were those in the RBI_CENS2021 (individual code, household code, date of birth, age, sex, citizenship, household size, municipality of residence), enriched by those in the National Register of

Resident Population[1] (ANPR) (relationship with the reference person (RP), marital status and date of marriage or civil union).

Table 1: Distribution of the Italian Population and Households by household size by number of members. Absolute and Percentage values.

	Number of Individuals		Number of Households	
Household size by number of members	**A.V.**	**%**	**A.V.**	**%**
1	9,636,232	16.4	9,636,232	36.8
2	14,241,696	24.3	7,120,848	27.2
3	14,039,430	23.9	4,679,810	17.9
4	14,181,536	24.2	3,545,384	13.5
5	4,529,020	7.7	905,804	3.5
6 or more members	2,050,881	3.5	318,168	1.2
Total	**58,678,795**	**100**	**26,206,246**	**100**

Source: Our elaboration on Istat data

Furthermore, auxiliary variables, useful to the reconstruction process, were calculated. One of the difficulties encountered was the lack of information for undercover individuals for whom only household code, gender, age and citizenship were available.

For some variables it was necessary to carry out a set of initial Editing and Imputation (E&I) activities to verify the validity and correctness of the individual variables, such as the compatibility of the date of marriage with that of birth.

2.2.1 Reclassification of the variables: relationship with the RP and marital status

The relationship with RP of ANPR (30 categories) were reclassified to match the classification used in the census (23 categories). This reclassification was particularly complex especially when there is not a unique correspondence between two categories. For example, “Cohabiting with adoption or emotional ties” in ANPR corresponded to two categories of the census: “Cohabiting in consensual union with RP” and category “Other cohabiting person without being a member of a couple or a relative”.

The reclassification of marital status did not cause particular problems; however, the personal data had a lack of structural information due to the registry regulation especially for the foreign population; the municipality offices, in fact, do not record the marital status when individuals cannot produce adequate documentation of the country of origin, therefore, many individuals married abroad are registered in the registry office with "Unknown" marital status (about 1.4 million individuals equal to 2.4% of the total population and 30% of the total foreign population

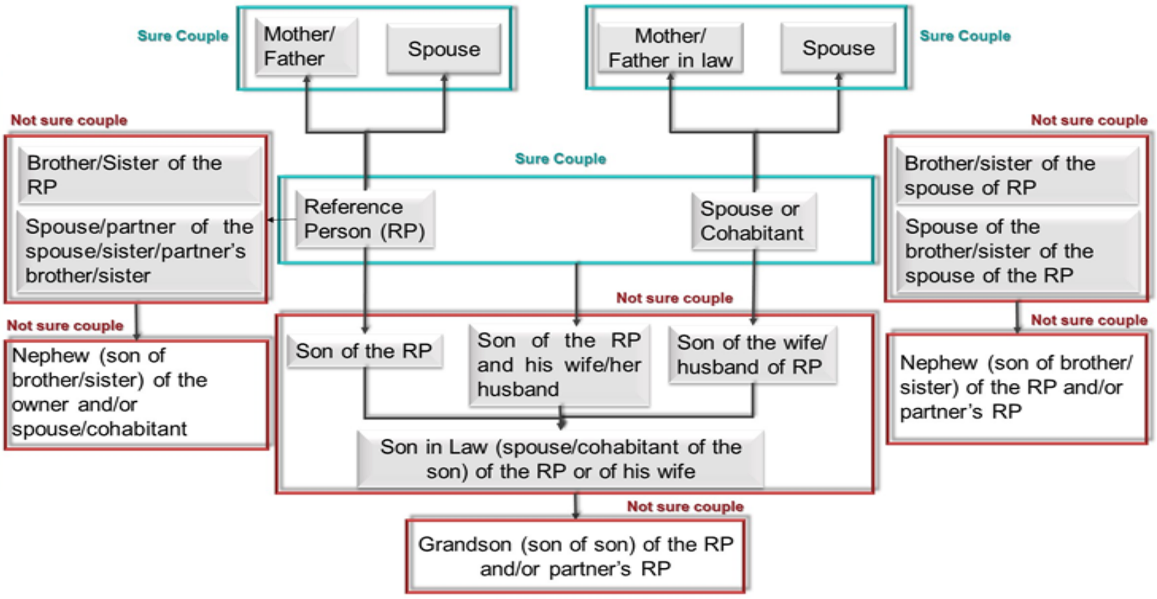
[1] Since 2018, Istat acquired, directly from the Ministry of the Interior (MI), data on the resident population at 1st January Year *t.* In 2022 MI ended the process of centralising data of all municipalites.

in Italy). To fill the lack of demographic information, another administrative source, the resident population by year of birth, gender and marital status at 31st of December 2021 (POSAS), was used to the imputation of this variable and was considered as a benchmark for the comparison between the RBI_CENS2021 data and the published demographic statistics on marital status.

2.2.2 Calculation of household auxiliary variables

The main auxiliary variable, already widely used in previous censuses, is the one that allows potential couples to be identified (Bianchi et al., 2020). Starting from the individuals, the couples are identified taking into account the relationship with RP, gender, age, marital status and year of marriage or civil union of the two partners by computing score on the basis of the individual information collected. The identification of potential couples, using optimization techniques (Bruni et al., 2001), allows identifying sure couples (Figure 1, green box) and not sure ones (Figure 1, red box), considering relationships, unique and otherwise, with RP.

Figure 1: Scheme of the identification of potential couples (sure and not sure) starting from the individuals and their relationship with the reference person (RP).


Sure Couple
Mother/ Father
Spouse
Mother/ Father in law
Spouse
Sure Couple
Not sure couple
Brother/Sister of the RP
Spouse/partner of the spouse/sister/partner's brother/sister
Sure Couple
Reference Person (RP)
Spouse or Cohabitant
Not sure couple
Brother/sister of the spouse of RP
Spouse of the brother/sister of the spouse of the RP
Not sure couple
Nephew (son of brother/sister) of the owner and/or spouse/cohabitant
Not sure couple
Son of the RP
Son of the RP and his wife/her husband
Son of the wife/ husband of RP
Not sure couple
Nephew (son of brother/ sister) of the RP and/or partner's RP
Son in Law (spouse/cohabitant of the son) of the RP or of his wife
Not sure couple
Grandson (son of son) of the RP and/or partner's RP

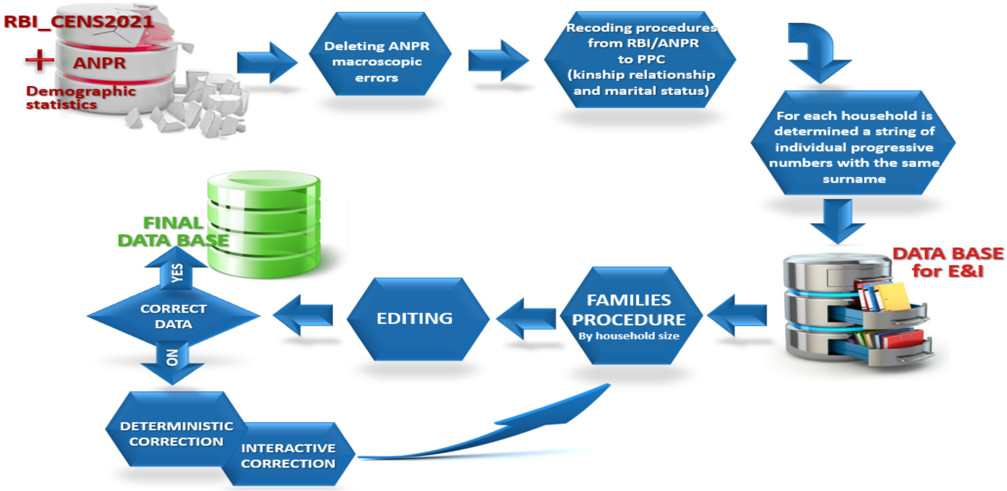
Another auxiliary variable calculated was the string of progressive numbers of individuals with the same surname within each household, respecting the anonymization process by guarantee the privacy of the individual data, according to the General Data Protection Regulation (GDPR Regulation 2016/679); this string was very useful for validating household members and correcting anomalies deterministically and interactively.

Further individual auxiliary variables have been calculated because they were functional to the E&I process, such as age at marriage or civil union, duration of marriage or civil union, etc.

2.2.3 Editing and "Families Procedure"

The E&I and household reconstruction was a very complex process (Figure 2), especially considering the demographic and social changes observed within households over the years.

Figure 2: Flowchart of the E&I and household reconstruction process



RBI_CENS2021
+ ANPR
Demographic statistics
Deleting ANPR macroscopic errors
Recoding procedures from RBI/ANPR to PPC (kinship relationship and marital status)
For each household is determined a string of individual progressive numbers with the same surname
DATA BASE for E&I
FAMILIES PROCEDURE
By household size
EDITING
CORRECT DATA
YES
FINAL DATA BASE
NO
DETERMINISTIC CORRECTION
INTERACTIVE CORRECTION

After having reclassified the relationship with RP and marital status, and determined the auxiliary variables, it was possible to proceed with the editing using the edit rules between personal data and familial variables; subsequently incorrect or missing data were imputed to restore consistency among variables. Only at the end of these activities the "Families Procedure" (PF) carried out, which checked and corrected some variables and then calculate the household and nuclei types. At the end of PF, familial editing carried out to verify the coherence between the household members with respect to age, gender, marital status, relationship with RP, etc. in order to identify any anomalous households. This process allowed the first level validation to be carried out before the release of the data for the 2nd level validation by the thematic experts. The process described was cyclical and reiterated (Figure 2) and ended only when the optimal result was achieved.

The mentioned "Families Procedure" is a software package (Budano et al., 2010) used by Istat social surveys for the reconstruction of the household and nuclei types. This procedure defines steps for the correction of individual variables in relation to those of the other members of the household. The E&I process do not end in a single "step" but require a reiteration on the data, to restrict the errors in increasingly smaller subsets until they have zero numbers. After the correction phase the PF calculates for each individual, the household to which each component belongs, the nucleus type and their respective positions in it.

In order to improve the performance and reduce the time to execute the PF it was necessary to subset the number of households by the household size and by grouping some provinces.

3. Main Results

Before launching individual and familial editing, missing data were imputed, mainly for foreigners, using the Istat source of demographic statistics on marital status calculated with a different methodology. After imputations, the PF was launched in order to identify the household and nuclei types and then analysed to verify the correct determination of household characteristics. The main results of the preliminary analyses carried out, are described below. The number of failed edits, with at least one incorrect individual error, involved 4,826,145 households (18.4%). If we consider households with the same number of members, the highest percentage (38.2%) was observed for households with 6 or more members, highlighting the complexity of larger households (Table 2).

Table 2: Distribution of households and households with at least one individual error by household size. Absolute and percentage values.

Household size by number of members	Households with at least 1 individual error			N° of Households
	A.V.	**%**	**Row %**	**A.V.**
1	2,029,259	42.0	21.1	9,636,232
2	1,197,503	24.8	16.8	7,120,848
3	838,359	17.4	17.9	4,679,810
4	443,861	9.2	12.5	3,545,384
5	195,591	4.1	21.6	905,804
6 or more members	121,572	2.5	38.2	318,168
Total	**4,826,145**	**100**	**18.4**	**26,206,246**

Source: Our elaboration on Istat data

The main errors referred to missing data (Table 3) of the *year of marriage or civil union* (57.9%) and *marital status* (37.4%).

Table 3: Distribution of missing data and failures by type of edits. Absolute and percentage values.

	Number of errors	
	A.V.	**%**
Missing data	**3,788,122**	**100**
Relationship with reference person	*174,585*	*4.6*
Marital status	*1,418,407*	*37.4*
Year of marriage or civil union	*2,195,130*	*57.9*
Individual Inconsistencies	**937,328**	**100**
Relationship with reference person	*42,174*	*4.5*
Marital status	*810,615*	*86.5*
Others	*84,539*	*9.0*
Familial Inconsistencies	**223,082**	**100**
Relationship with reference person and marital status	*114,301*	*51.2*
Age differences	*58,865*	*26.4*
Partners/children and marital status	*37,952*	*17.0*
Others	*11,964*	*5.4*

Source: Our elaboration on Istat data

After the imputation process, editing was launched, using 94 edit rules (13 individual and 81 familial edits) to identify any inconsistencies.

Individual editing identified 937,328 failed edits (1.6%) out of the total units. Most of the errors concerned marital status (86.5%). The inconsistencies found between two or more variables (Table 3) were mainly between *relationship with RP* and *marital status* (51.2%).

The number of errors of *marital status* (Table 4) were similar for women (50.88%) and men (49.12%); the 30-59 age group is the most affected by errors (62.44%). Finally, with respect to citizenship, the highest number of errors were observed among Italians (80.22%), especially women (40.62%), mainly due to changes observed among married people and de facto or legally separated people, categories present in the census, but not in ANPR.

Table 4: Distribution of the errors of marital status, by age groups, gender and citizenship (Italian (It) and Foreign (For)). Percentage values.

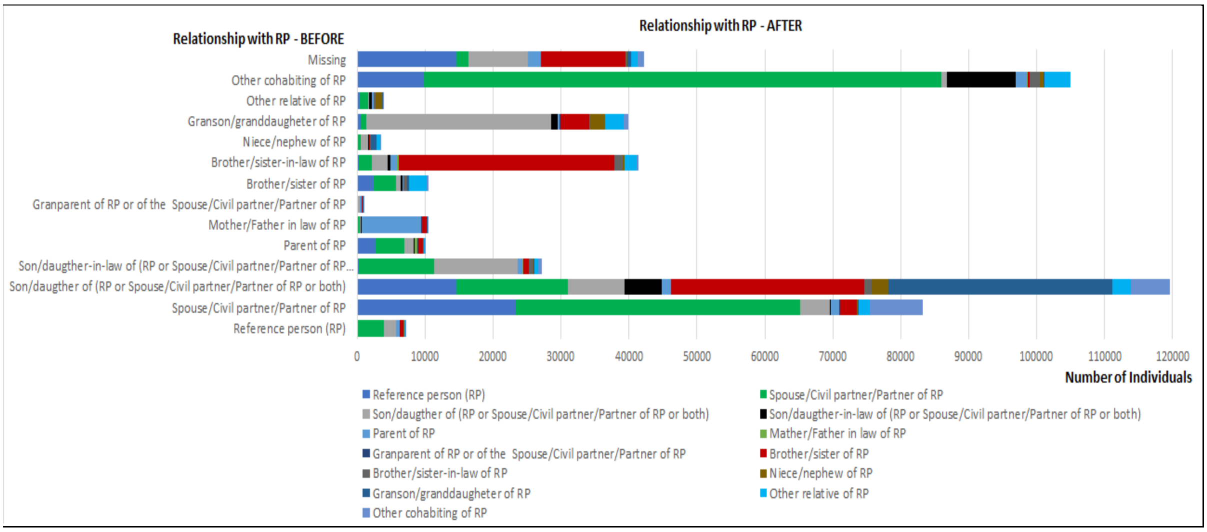
	Women			**Men**			
Age groups	**It**	**For**	**TotW**	**It**	**For**	**TotM**	**Total**
0-16	0.11%	0.13%	0.24%	0.13%	0.14%	0.27%	0.51%
17-29	0.61%	1.61%	2.22%	0.35%	2.04%	2.39%	4.61%
30-59	26.18%	6.56%	32.74%	23.41%	6.29%	29.71%	62.44%
60-84	12.93%	1.89%	14.83%	14.93%	1.02%	15.95%	30.78%
85 and over	0.79%	0.07%	0.85%	0.78%	0.03%	0.81%	1.66%
Total	**40.62%**	**10.26%**	**50.88%**	**39.59%**	**9.53%**	**49.12%**	**100%**

Source: Our elaboration on Istat data

In addition, for *year of marriage or civil union*, there were more imputations of missing data (66.84%) and few inconsistencies with the year of marriage or union of the partner or incompatible with the year of birth.

The corrections of the *relationship with RP* were more complex (Figure 3).

Figure 3: Distribution of the relationship with RP before/after the E&I. Bars are % of each category.



Source: Our elaboration on Istat data

For this variable, there were few missing data (Table 3), only (4.6%). In the case of children (third bar from the bottom of Figure 3) the changes in relationship with RP occurred with grandchildren (27.6%, dark blue bar) and with siblings (23.7%, red bar).

4. Concluding remarks

In this work the process of the household and nuclei type reconstruction has been briefly described, highlighting the complexity linked both to the integrated use of data gathered from registers, administrative sources and surveys, and to the adaptation of PF to a huge amount of data. It is important to underline that PF was used for the first time on integrated data, without never having tested it on big dataset, relating to individuals and households belonging to the all resident Italian population. In addition, this process improved the quality of data released to Eurostat with reference to census hypercubes involving household and nuclei types. However, further studies, both on sources and methods, will be useful to reduce missing data and errors as much as possible. It will be interesting to apply Machine Learning methods or Artificial Intelligence to improve the household reconstruction minimizing errors, especially for households with numerous members which internal composition is difficult to detect.

Another hope would be to reengineer the PF aiming to optimize the speed of its execution and the performance by reducing some anomalous household. New generation programming languages can allow to better maintain the application, furthermore generalised solutions can allow to adapt the PF to the specific needs of other social survey.

Acknowledgment

The authors wish to thank Silvia Dardanelli and Salvatore Pagliara for helpful contribution and suggestions. The work is the result pursued jointly by the two Authors. Opinions expressed in this article are those of the Authors but do not involve the institutions they belong to.

References

ANPR (2024). *Anagrafe Nazionale Popolazione Residente*. https://www.anagrafenazionale.interno.it

Bianchi G, Filippini R, Lipsi RM, Pezone A, Scalfati F. (2020). *An overview of the editing and imputation process of the 2018 Italian Permanent census*. UNECE, online workshop on Statistical Data Editing.

Bruni R, A Reale and R Torelli, Optimization Techniques for Edit Validation and Data Imputation (2001). In Proceedings of Statistics Canada Symposium: Achieving *Data Quality in a Statistical Agency, Ottawa, Canada.*

Budano G. e P. Piergentili (2010), La Procedura Famiglie in G. Budano e S. Demofonti, La misurazione delle tipologie familiari nelle indagini di popolazione in *Metodi e Norme*, 2010, n. 46. Istat.

Eurostat, (2017). European Commission. Commission Regulation No 763/2008 of the European Parliament and of the Council, OJ L 105, 21.4.2017, p. 1–11.

GDPR (2016). *General Data Protection Regulation* (GDPR – Regulation 2016/679).

Istat (2022). Nota tecnica sulla produzione dei dati del Censimento Permanente: *la popolazione residente per genere, età, cittadinanza e grado di istruzione al 31.12.2021*. pp.14.