doi.org/10.26398/asaproc.0025

Population census: Methodology for the list survey

Fabio Crescenzi[a], Emanuela Scavalli [b]
[a] Department for Statistical Production, Istat, Rome, Italy
[b] Directorate for Social Statistics and Population Census, Istat, Rome, Italy

1. Introduction

Starting from 2018, the Italian National Institute of Statistics (Istat) launched the permanent population and housing census (PPHC), based on the data integration of administrative sources with sample surveys. The first round of the PPHC (2018-2021) was designed starting from the register of the Italian population (RBI), a register obtained integrating multiple administrative sources, among which the local population registers of Italian municipalities (LAC).

In addition to the RBI data, address data from the statistical register of places (RSBL) and data from the education and employment registers were used and two sample surveys were designed and carried out annually (both area and list surveys) to assess coverage errors in the RBI and collect data for variables that are not fully available in the registers (Agafiţei et al., 2015).

In 2020, given that it was not possible to carry out the surveys due to the COVID-19 pandemic, the estimation method was changed using RBI data and "signals of life" (SoL) (Gallo and Zindato, 2021). [1] Taking advantage of this experience, the second round of the PPHC (2022-2026) integrates RBI data and SoL forming an extended population register (Solari et al., 2023).

SoLs are obtained from AIDA (Integrated Database of Usual Residents), which contains integrated data over time on individuals from more than forty administrative sources. SoL profiles are defined aggregating individuals in sub-populations having similar over-coverage attitude. An indicator function based on SoL profiles is defined (Bernardini et al. 2022), according to which any individual in a profile is classified as either included or excluded from the population count.

New techniques have been explored that use survey data to improve SoL profiles (latent class models and other data science methods) in relation to population groups that are not easily identifiable by deterministic criteria alone (Gallo and Zindato, 2021). At the same time, survey data can be used to estimate the bias in population counts and better define SoL profiles.

Two surveys are conducted to support the registers:

- An area survey, which consists of a sample of addresses that includes all households in an address: this survey aims to estimate the possible under-coverage of the population register.
- A list survey carried out on a sample of households from the LAC registers with the aims of estimating the distribution of the census variables of interest (occupation, education, dwelling, etc.) and defining SoL profiles.

The data from administrative sources was integrated with survey data (collected through list or areal sampling) through a statistical model. This paper describes the methodology employed in developing the list survey.

2. The list survey design

The list survey design is a two-stage sampling with stratification of the primary units (municipalities); the secondary units are households. The characteristics of all individuals

[1] Administrative 'signs of life' refer to activities of individuals that can be inferred from administrative records. Being self-employed or working for a company, being a civil servant, having a regular annual rent for a dwelling, attending school or university courses are examples of direct administrative life signs. An indirect life sign is instead an identifier of a status or condition, such as being an income or pension recipient.

belonging to the same family are recorded. In each municipality, the sample households are selected from the RBI list with equal probabilities.

The municipalities (updated until 2021/06/30) are divided into two groups:

i) self-representative (SR) municipalities annually participating in the survey (1,188 units);
ii) non self-representative (NSR) municipalities (6.716 units), participating in the survey only once in five years (approximately 1,343 per year).

The SR municipalities have been defined according four criteria:

1. municipalities with more than 17.800 inhabitants
2. municipalities provincial capital
3. municipalities that are SR in the labour force survey (not rotating)
4. critical municipalities according to parameters derived from the analysis of the surveys carried out in the previous census cycle 2018-2021.

Each SR municipality is considered as a separate stratum and is included in the sample each year.

Table 1. Number of sampled municipalities - period 2022-2026

	SR municipalities	NSR municipalities	TOTAL municipalities
Population ≥17.800	600		600
Provincial capital municipalities	109		109
SR municipalities in the LFS	916		916
Critical municipalities	149		149
TOTAL	**1,188**	**6,716**	**7,904**

Within each province, the NSR municipalities were grouped into strata of five units according to population size. So, we obtained 1,381 strata, 85 of which contained less than five municipalities because they were residual strata in the province. 61 strata have been constructed with border municipalities[2] (including a total of 305 out of 313 border municipalities –see Table 2). The other 8 border municipalities were included in the remaining 1,320 strata. To define the year in which the NSR municipalities had to be surveyed, we excluded the municipalities that participated in the 2021 survey.

Table 2. Number of border municipalities in the Italian provinces

Provinces	Border municipalities
12 - Varese	103
13 - Como	139
97 - Lecco	21
103 - Verbano-Cusio-Ossola	50
Total	313

3. Allocation of the households

The households to be interviewed were assigned to municipalities adopting a mixed strategy based on cost and organisational criteria as well as the sampling errors of the main estimates. The

[2] Municipalities with workers who are employed in a border zone of another State but who return daily, or at least once a week, to the neighbouring State in which they reside and of which they are nationals.

allocation strategy considered the following parameters:

- 100 is the minimum number of households to interview in a municipality;
- municipalities with less than 300 inhabitants are exhaustively surveyed (446 units);
- municipalities with a sampling fraction >90% and with difference between population and sample <40 households are exhaustively surveyed (8 units);
- households of border municipalities are oversampled in proportion to the border workers.

Table 3. Allocation of household sample by year - period 2022-2026

Type of municipalities	Municipalities	Households sample
SR every year	1,188	604,062
NSR year 1	1,343	394,683
NSR year 2	1,343	394,522
NSR year 3	1,343	394,898
NSR year 4	1,345	395,272
NSR year 5	1,342	395,129

Approximately 2,531 municipalities (NSR+SR) and about 1 million households (see Table 4) were surveyed each year.

Interviews were conducted using a mixed mode technique (CAWI, CAPI, CATI), with a first phase of so-called "spontaneous response" and a second phase of field follow-up of non-respondents.

Table 4. Annual sample of Municipalities and Households. period 2022-2026

Year	Municipalities	Households sample
2022	2,531	998,745
2023	2,531	998,584
2024	2,531	998,960
2025	2,533	999,334
2026	2,530	999,191

4. Weighting system

The weights assigned to the sample units are obtained through a complex procedure enabling to correct for possible non-response bias from households that were unavailable or refused to be interviewed and ensure that the sample estimates are consistent with known totals of auxiliary variables.

The distributions of the population by gender, age class and citizenship (Italian *vs.* non-Italian) for each province and 12 main municipalities are known from the RBI. [3]

Indicating with ${}_kX$ (k=1, …, c) the known total of the k-th auxiliary variable for the generic province and with ${}_kX_{hij}$ the value assumed by the k-th auxiliary variable for the respondent unit hij, the above described condition is expressed by the following equation:

$$ {}_k X = {}_k \hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} W_{hij\,k} X_{hij} \quad (k=1, \ldots., c) $$

where H indicates the number of strata. In case a stratum is empty, it is aggregated to the neighbouring one.

[3] Torino, Genova, Milano, Verona, Venezia, Bologna, Firenze, Roma, Napoli, Bari, Palermo, Catania

The final weights for the respondent sampling units are determined as follows:

1. initial weights are obtained as the inverse of the inclusion probability of the units in the stratum;
2. correction factors for total non-responses are worked out as the reciprocal of the response ratio in the municipality to which each unit belongs;
3. basic weights corrected for non-response are then computed multiplying initial weights by these correction factors for total non-response;
4. provincial correction factors are obtained solving a minimisation problem under constraints, which make it possible to satisfy the condition of equality between the known totals of the auxiliary variables and the corresponding sample estimates;
5. final weights are computed multiplying the basis weights by the correction factors obtained at the previous step.

The correction factors of point 4 are obtained by solving a constrained minimum problem (Deville and Särndal, 1992), where the function to be minimised is a distance function (suitably chosen) between the basic weights and the final weights, the constraints regard the estimates of some auxiliary variables that have to be equal to the totals in the reference population derived from the RBI. The chosen distance function is the truncated logarithmic function; the adoption of this function guarantees that the final weights are positive and within a predetermined range of possible values, thus eliminating the extreme positive weights (too large or too small).

All estimation methods resulting from solving a constrained minimum problem of the type described above belong to a general class of calibration estimators.

5. Concluding remarks

The Italian permanent census involves only representative samples each year, unlike the traditional census, which involves all citizens and all households at the same time. In any case, the former returns estimates of the population at municipal level (Carbonetti et al. 2023).

The design of the list survey, as defined in the first round of surveys (2018-2021), was changed in the second round (2022-2026). The annual estimates concerning the municipalities are incomplete since the NSR units are only surveyed in one of the five-year cycle.

For this reason, the annual surveys are used to collect information that, when integrated with data from administrative registers, may help to define profiles that support the enumeration of the population and the estimation of census variables using alternative methods to the traditional ones.

The list survey was used to determine the RBI over-coverage error measure for the 2018 and 2019 population censuses. After 2022, these surveys will continue to play a crucial role in collecting data for variables that cannot (or can partially) be replaced by administrative data and in providing quality measures for the fully register-based population size estimation.

The implementation of more efficient registers is also fundamental because they are a starting point for the other sample surveys carried out by Istat, in order to integrate the information collected in a surveys with that stemming from administrative registers.

References

Agafiţei, M., Gras, F., Kloek, W., Reis, F. & Vâju, S. (2015). Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the IAOS,* **31**, pp. 203-211.

Bernardini, A., Chieppa, A., Cibella, N., Gallo, G., Solari, F., Zindato, D. (2022). Evolution of the Italian permanent population census. Lessons learnt from the first cycle and the design of the permanent census beyond 2021. *Conference of European Statisticians, Group of Experts on Population and Housing Censuses*, Geneva, *Switzerland, 21-23 September 2022*. Economic Commission for Europe.

Carbonetti, G., De Matteis, G., Di Zio, M., Fardelli, D., Ferrara, R., Lipizzi F. (2023).

Enumeration area imputation methods for producing sub-municipal data in the Italian permanent population and housing census. *Statistical Journal of the IAOS,* **39**(1), pp. 123-136.

Deville, J.C., Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**(418), pp. 376-382.

Gallo, G., Zindato, D. (2021). Italy: The combined use of survey and register data for the Italian permanent population census count in UNECE, *Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses.* https://unece.org/statistics/publications/CensusAdminQuality.

Solari, F., Bernardini, A., Cibella, N. (2023). Statistical framework for fully register based population counts, *Metron*, 10.1007/s40300-023-00244-5.

Zhang, L.C. (2022). Complementarities of survey and population registers, in N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, J.L. Teugels (eds.) *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat0835