

Capture Recapture Approach to Correct for Under-coverage in the Register Supported 2011 Italian Population Census

Loredana Di Consiglio, Marco Fortini

diconsig@istat.it; fortini@istat.it

ISTAT

Abstract: 2011 Italian census round aims at simplifying the municipalities activity through a smaller field workload, which implies less enumerators, and a greater use of administrative data in supporting the data collection phase. In this context, coverage of administrative registers is the main quality issue to be considered. Whereas over-coverage can be found during data collection phase through detection of not eligible units, evaluation of under-coverage rate needs of a distinct survey, either by extensive search on the field for missing units or by a sample based dual-system estimation. In this paper the second approach is described, addressing to the post-enumeration survey design, with special emphasis to modelling issues such as small area estimation to provide counts at municipality level.

Keywords: Administrative registers, small area estimation, coverage assessment.

1. Introduction

As a consequence of social and institutional changes occurred in the past years, Italian census methodology is shifting from a complete enumeration of households and people based on a field search, to a *register-supported* census, implemented through questionnaires *mailed out* to all households enlisted into the municipal population registers. These innovations, together with a mixed mode data collection method which makes use of web in addition to *mail back* and municipal data collection centers, is expected to allow for a substantial reduction in the number of enumerators, at the same time increasing the importance of administrative data for the Italian statistical system. Even if census enumeration is always affected by coverage error, its accurate measure becomes particularly important in the context mentioned above. In fact, direct measure of register over-coverage is obtained by checking for the presence of registered people on the field, whereas under-coverage assessment is more complex, needing of either supplemental lists of people or a demanding work of search for people living on the field but not enlisted in population registers.

An estimate of the gross under-coverage of Italian population registers is given in Fortini and Gallo (2009) through the use of administrative data concerning post-censal changes of residence between Italian municipalities due to the enumeration of people in municipalities where they were not resident. Although the cited source is biased downward for reasons not discussed in this work, the whole amount of under-coverage affecting population registers is estimated by means of a mixture model approach in approximately 400,000 people. Being 800,000 the estimated number of people missed by the 2001 Italian census, the use of municipal population registers as a source for households enumeration can be considered reasonable.

As the state of the available supplemental lists does not seem to offer a good solution, the complete search on the field of people to be enlisted into population registers would represent even a worst solution since the accomplishment of a such approach would require large extra resources commitment and organization efforts. A different approach relying on sampling survey and dual-system estimation is described in the following. To this purpose, alternative sampling schemes and estimators are evaluated through simulation of sampling universe regarding a properly designed artificial population.

2. Capture-Recapture Approach

As already mentioned, the municipality register will be the basis for the population count. Different alternatives can be proposed to tackle the problem of adjusting the population counts for under-coverage of population registers, in particular here we focus on the *statistical adjustment* of enumeration undercount based on capture-recapture methods. To be more precise, let us assume that a second list of the population is obtained by a full field enumeration, being the first one set by the population register; we denote by x_{1+} , x_{+1} and x_{11} the amounts of those enumerated by the municipality register, by the second list and by both of them, respectively. Assuming independence in the *capture* probability (i.e. in the probability of being enumerated) of the two different lists, homogeneity of the *capture* probability for each list, and that no errors occurs during the linkage process, an estimate of the true population count N is given by

$$\tilde{N} = x_{1+} + x_{+1} \left((x_{+1} - x_{11}) / x_{11} \right) = \frac{x_{1+} x_{+1}}{x_{11}} \text{ (see Wolter, 1986 for more details).}$$

Since a full field enumeration of the Italian population would be impractical, a sample of census Enumeration Areas is selected so to estimate the whole count x_{+1} . We will return in section 2.2 on methods for estimating the true count in this context. Here it is important to underline that the desired measure of the true count is needed at municipality level, besides aggregation such as provinces and regions.

2.1 Sampling design for the 2011 Post- Enumeration Survey

Two alternative sampling designs have been proposed sharing the same feature of being two-stage sampling design, with municipalities as primary sampling units. Note that, since this choice is necessary so to reduce operational costs, most of the domains of interests (municipalities in this context) will have zero sample size, implying the need of small area estimation methods. The two designs differ in the level at which the sample size is planned: region by class of population size in 4 classes: less than 5000, 5000-20000, 20000-50000, over 50000 (design A); aggregation of provinces inside region by the 4 classes of population size above (design B). Stratification and selection of municipalities according to demographic size is also assumed.

2.2. Estimators

The direct estimation of census counts are available only at planned domain level, whereas it is not obtainable at municipality level apart from those included in the sample. For this reason the estimation of the municipality true counts needs of model assumptions on their relationship with available predictors such as population counts coming from the register, population distribution by sex-age groups and others.

At the very first stage of analysis reported in this work, we considered a synthetic estimator based on invariance assumption of the under-coverage rates for municipalities belonging to the same planned domain. Being design A domains larger than those of

design B a larger bias is expected in the former setting than in the latter, together with a better performance in terms of estimates variability .

3 Description of the Empirical Study

A simulation study was carried out in order to compare the two sampling strategies. Two pseudo-populations of municipalities, with different amount of character variability, have been generated on the basis of the study in Fortini and Gallo (2009). True N population counts, observed counts from population registers (X_{1+}) and from field enumeration (X_{+1}), both affected by undercount, were assigned to each municipality. N was determined starting from X_{1+} counts by adding the expected undercount resulting from the mixture model described in Fortini and Gallo (2009). X_{+1} were determined from a Poisson with parameter Np where p are the coverage rates coming from 2001 Census post enumeration survey. To account for the second stage sample of census areas, for each municipality *observed sampling values* x_{+1} and x_{11} were generated from Binomial and Hyper-geometric distribution respectively. The two pseudo populations differ in their registers' undercoverage variability being the former one defined by predicted values N coming from the mixture model cited above. Differently, for the latter population N values are defined through the observed counts of resident people who are not enlisted into the population register after their adjustment so to sum up to the national gross undercoverage estimated by the mixture model (equal to approximately 400.000). In this way the second pseudo-population represents a much more difficult framework for estimation due to its much larger variability.

A thousand samples of municipalities were selected according to sample design A and B from both populations. No design effect was applied at this stage. A direct Petersen estimate of the undercoverage rate was estimated at domain level, finally the rate was applied at municipality level to estimate of the true population size. The results are compared with the counts in population registers, used here as a benchmark even known they are downwards biased, since their zero cost of achievement.

4 Results and Comments

First of all, counts coming from registers were compared with those from sampling strategy. Since the comparison shows a similar pattern for design A and B, here we report bias and MSE of the synthetic estimator only for design A together with corresponding bias of the register counts. Even though variance of the synthetic estimator can be considered small enough, the improvement of the synthetic estimator in terms of bias may be questionable. On the contrary, at level of design planned domains, the use of the capture – recapture methodology (results not reported in the following) shows negligible values for both bias and MSE, with a great improvement with respect to registers. Enhancement in the small area estimator may leave room for further reduction of bias at municipality level, as well. With regard to the comparison between the performance of the two sampling strategies, the average bias of the synthetic estimator in Design B is, as expected, lower than average bias in Design A. Differences increase within larger population classes. With respect to MSE the two designs are equivalent for population 1. For population 2, Design A is better for classes 1 and 2, whereas for classes 3 and 4 the reduction in bias is such that design B improves

design A in MSE as well. However, evidences in favour of Design B are not irrefutable and further studies are necessary to achieve decisive results.

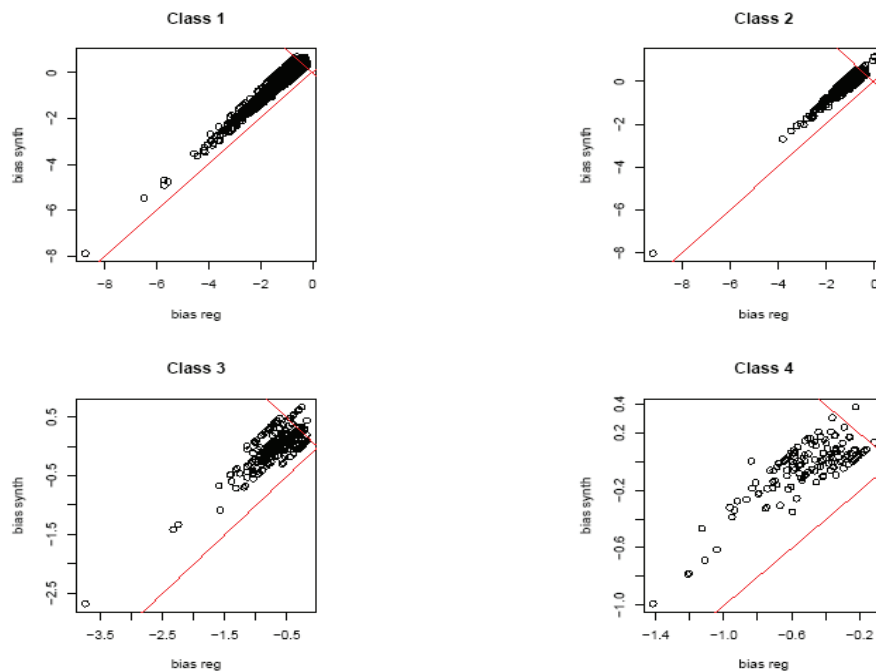
Table 1 Average and (Maximum Relative Bias) % of the synthetic estimator and register count for Population 1 and 2- design A by class of municipality size

	Class 1		Class 2		Class 3		Class 4	
	Register	Synthetic	Register	Synthetic	Register	Synthetic	Register	Synthetic
P1	0.952 (8.75)	0.309 (0.703)	1.064 (9.199)	0.20 (1.166)	0.643 (3.739)	0.186 (0.673)	0.515 (1.415)	0.129 (0.381)
P2	0.844 (14.118)	0.685 (2.038)	0.942 (14.92)	0.526 (1.319)	0.707 (6.301)	0.441 (1.30)	0.474 (2.426)	0.306 (0.844)

Table 2 Average Relative Root MSE % of the Synthetic estimator for Population 1 and 2- design A by different class of municipality size

	Class 1	Class 2	Class 3	Class 4
Pop1	0.339	0.232	0.208	0.146
Pop2	0.729	0.558	0.457	0.315

Figure 1. Bias of synthetic estimator vs register count for Pop 1 - design A by class of municipality size



References

- Fortini M., Gallo G (2009) Misure di sottocopertura anagrafica in base alla revisione post-censuaria del 2001, *Giornate di Studio sulla Popolazione*, Milano 2009
- Wolter K. M., (1986) Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-346.