istat
working
papers

N.2
2023

A study on bootstrap approaches for variance estimation of population counts with under- and over-coverage

Simona Toti, Marco Di Zio, Alessandra Ronconi

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.2
2023

A study on bootstrap approaches for variance estimation of population counts with under- and over-coverage

Simona Toti, Marco Di Zio, Alessandra Ronconi

Istat

Istat Working Papers

A study on bootstrap approaches for variance estimation of population counts with under- and over-coverage

N. 2/2023

ISBN 978-88-458-2103-5

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

A study on bootstrap approaches for variance estimation of population counts with under- and over-coverage

Simona Toti, Marco Di Zio, Alessandra Ronconi[1]

Sommario

Nel censimento della popolazione italiana del 2018, i conteggi della dimensione dei comuni sono calcolati utilizzando il Registro di base degli individui (RBI) opportunamente corretto per sovra e sotto copertura. Questo lavoro è finalizzato allo studio di un metodo per la valutazione dell'accuratezza delle stime di tali conteggi di popolazione. Il coefficiente di correzione è dato dal rapporto tra il complemento delle probabilità di under e over coverage. Tali probabilità sono stimate attraverso regressioni logistiche con effetti casuali applicate ai dati di indagine. Per calcolare la varianza di questo stimatore, vengono studiate alcune procedure basate sulla tecnica bootstrap. La procedura di ricampionamento bootstrap viene applicata considerando due contesti: nella prima si suppone che RBI sia affetto da errori di misurazione, nella seconda RBI viene considerato privo di tali errori. Il documento riporta i risultati di un'applicazione sperimentale ai dati di RBI e delle indagini campionarie condotte per il censimento permanente del 2018.

Parole chiave: Metodo bootstrap della pseudo-popolazione, stima della varianza, errore di misurazione, errore di copertura.

Abstract

In the 2018 Italian Population Census, counts are computed by using the Base Register of Individuals (BRI) corrected for over- and under-coverage. The aim of the paper is to propose a method for the evaluation of the accuracy of those count estimates. The correction coefficient is the ratio between the complement of the probabilities of under- and over-coverage. Those probabilities are estimated through logistic regressions with random effects applied to survey data. To compute the variance of the estimator, we study some procedures based on the bootstrap technique. The resampling procedure is applied considering two settings: in the first we suppose that BRI is affected by measurement errors, in the second BRI is free of measurement errors. The paper reports the results of an experimental application to the BRI and sample survey data for the 2018 permanent census.

Keywords: Pseudo-population bootstrap, variance estimation, register-based statistics, measurement error, coverage error.

1 Simona Toti (toti@istat.it); Marco Di Zio (dizio@istat.it); Alessandra Ronconi (alronconi@istat.it), Italian National Institute of Statistics/Istituto Nazionale di Statistica – Istat.

This Istat working papers stems from the efforts of the Istat Advisory Committee on Statistical Methods.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this Istat working papers N. 2/2023.

Contents

Pag.

1. Introduction

In the 2018 Italian Population Census, counts are computed by using the Base Register of Individuals (BRI) corrected for over- and under-coverage. An approach to achieve this goal involves the use of a coefficient that either expands or shrinks the register counts of population consistently with under-over-coverage (Pfefferman, 2015). The coefficient used for the correction is the ratio between the complement of the probabilities of under- and over-coverage at unit level. To estimate the coefficient, a sample from the register and an area sample are carried out. In Mancini *et al*. (2014), logistic regressions with random effects are applied to survey data to obtain the individual probabilities of over- and under-coverage conditional on individual characteristics such as gender, age, and citizenship.

The analytical form of the correction coefficients makes complex to achieve an explicit expression for the accuracy of the count estimates, for this reason, we propose to use resampling techniques for finite populations. As discussed in (Mashreghi *et al*., 2016), three bootstrap classes can be identified: pseudo-population, direct bootstrap, and survey weights methods. Under the pseudo-population bootstrap, the target population is estimated by creating a pseudo-population via the original sample, and then the bootstrap sample is drawn from the resulting pseudo-population. In direct bootstrap methods, bootstrap samples are directly drawn from the original data set, but some modifications have to be made so that the bootstrap variability reflects the sampling variability of the original sampling design, an example is the method proposed in Rao and Wu (1988). In the third group, instead of resampling observations from the original data set to create a bootstrap sample, the sample remains fixed, but a set of bootstrap survey weights is generated by making rescaling adjustments on the original survey weights (see Rao *et al*., 1992).

We opted for the pseudo-population bootstrap (Chen *et al*., 2019) also called "bootstrap without replacement" (Gross, 1980; Särndal, Swennson, and Wretman, 1992) because we need to deal with a complex situation given by an estimator based on the ratio of two logistic regression with random effects and data affected by measurement errors. In fact, the pseudo-population approach – once an error model is assumed – essentially consists in the repetition of the process and data generation mechanism, while the other methods require further computations not available in the literature.

In the proposed bootstrap procedure, we first generate a pseudo-register and then, by considering the under- and over-coverage probabilities, a pseudo-population is generated. Finally, samples are drawn from the pseudo-population and used to compute count estimates of municipalities. The overall procedure is repeated, and the estimated variance is given by the variance of those count estimates.

In the study, the resampling procedure is applied considering two settings: in the first, we suppose that BRI is affected by measurement errors, in the second we suppose that BRI is free of measurement errors. In the first case, it was necessary to introduce a measurement error model in the bootstrap procedure and, in absence of any information, errors were modelled through a multinomial distribution centred on the BRI observed frequencies.

We notice that the sampling rate of the Italian population census survey is relatively low, the bias of the estimator of the variance in finite population could be potentially negligible because of the small impact of the finite population correction factors (see the general discussion in Mashreghi *et al*., 2016). Hence, a comparison between pseudo-population and classical bootstrap (bootstrap with replacement) is performed.

The paper reports the results of an experimental application of the procedure to the 2018 BRI and sample survey census data. All the experiments were performed by using ad-hoc developed R codes (R Core Team, 2021).

2. The pseudo-population bootstrap procedure with under- and over-coverage

The analysis of the Italian municipalities proceeds separately for each Italian Region (NUTS2), and within Region for over and under 18,000 people. For each of those strata (large/small municipality for a certain Region), the BRI count of individuals N_x^{BRI} with the characteristics x determined by gender, one of the 5 age classes and citizenship (Italian or not Italian) are corrected for under- and over-coverage to return the estimated count $\hat{N}_x$ by

$$\widehat{N_x} = N_x^{BRI} . \frac{1 - p_{over,x}}{1 - p_{under,x}} \qquad (1)$$

where $p_{over,x}$ and $p_{under,x}$ are the probabilities for an individual with profile x to be over- and under-covered. The probabilities are estimated via logistic regressions with random effects using survey data as already mentioned. We propose the use of pseudo-population bootstrap for the evaluation of the variance of $\hat{N}_x$.

The application of a pseudo-population bootstrap generally consists in expanding the sample through the sampling weights with the aim of reproducing a pseudo-population, and then drawing samples from the pseudo-population according to the designed sampling design. In our application, when creating a pseudo-population, we need to consider the over- and under-coverage mechanisms affecting the register as well. Hence, a modification of the pseudo-population bootstrap is needed. We first generate a pseudo-register, and then we derive a pseudo-population according to the under- and over-coverage of the register.

The following steps detail a proposal for the generation of a pseudo-population taking into account those elements.

Algorithm A1 for building a pseudo-population:

Pseudo-register generation.

Simulate the pseudo BRI counts by using the vector f^{BRI} composed of the relative frequencies f_x^{BRI} of x in the original (counts without correction for coverage) BRI and the total BRI counts N^{RBI}, by means of a multinomial distribution Mn(N^{BRI}, prob=f^{BRI}):

ps.BRI~rmultinom(n=1, size=N^{BRI}, prob=f^{BRI})

Correction of pseudo-register for over-coverage.

Simulate for each profile x, the counts of subjects that correctly are in pseudo BRI using $ps.BRI_x$ and f_x^{Nover}, that is the relative frequency of not over-covered subjects in the original survey from the list. Those counts are randomly generated from a binomial distribution Bin($ps.BRI_x$, prob=f_x^{Nover}):

ps.NOVERx~rbinom(n=1, size=$ps.BRI_x$, prob=f_x^{Nover})

Correction of pseudo-register for under-coverage.

Simulate for each profile x, the count of subjects under covered present in pseudo BRI using $ps.NOVER_x$ (remark: the counts of not over and not under covered are the same in

the population) and f_x^{Nunder} that is the relative frequency of not under covered subject in the original survey on area, by means of a negative binomial distribution NB($ps.NOVER_x$, prob=f_x^{Nunder}):

$$\mathbf{ps.UNDER}x \sim \text{rnbinom}(n=1, size=ps.NOVER_x, prob=f_x^{Nunder})$$

Step 1 is used for creating a pseudo-register, and steps 2 and 3 are aimed at the generation of a complete pseudo-population.

We notice that, in the first step BRI is randomly generated from the observed frequency distribution in BRI. This step is introduced with the underlying idea that BRI is affected by variability due to some unknown measurement errors.

We have also performed experiments based on the assumption that BRI is not affected by measurement errors. In practice, the resulting algorithm (henceforth denoted with A2) is the same as the algorithm A1 but without the first step. We emphasise that, in this setting, initial counts of BRI are still random, but their randomness is induced by the coverage errors modelled by means of binomial and negative binomial probability distribution.

2.1 The sample selection procedure

In a bootstrap approach, we should extract a sample according to the sampling design from the pseudo-population. This task is aimed at considering the sampling variability in the evaluation of the precision of the estimator. In the Italian census, it requires drawing a list and area sample from each pseudo-population.

We remind that the original surveys were performed by a two-stage sampling design. In the survey for estimating the over-coverage, the primary sampling units (PSUs) are the Italian municipalities and the secondary sampling units (SSUs) are the households. For the under-coverage area sample, the PSUs are still the municipalities, and a random sample of addresses was selected at the second stage.

In the bootstrap algorithm, the under- and over-coverage samples from the pseudo-population are drawn from the municipalities considered in the original census round, that is, we decided to sample only from the municipalities selected in the Census sample (*i.e.* only SSUs). This is because for producing the coefficient of variation (*CV*) estimates by municipalities, we should resample at the 2nd-stage (households/addresses). However, for any other domain of interest (*i.e.* other than municipalities) or characteristics *x*, resampling should be done at the 1st-stage level.

As already remarked, in principle we should reproduce the sampling design used in the list and area surveys. At this stage of the study, we are not able to exactly reproduce the sampling design because we cannot access all the information used for sample selection, hence we adopted a simplified sampling design. For each considered municipality, a simple random sample without replacement of individuals is drawn with sample size equal to the original survey data for this municipality.

2.2 The pseudo-population bootstrap procedure

The overall bootstrap procedure is the following:

a. Generate a pseudo-population according to A1 (or A2);
b. From the pseudo-population, K under- and over-coverage couple of samples are drawn from the municipalities considered in the original census round, with sample size equal to the original survey;
c. Estimate $\hat{N}_{b,x}$ is calculated on the under-over-coverage couple b, where b=1, ..., K;
d. (a)-(c) are repeated for L generated pseudo-populations;
e. At the end of the process, we have $B=L \times K$ bootstrap samples (replicates);
f. Bootstrap variance estimator of $\hat{N}_x$ is obtained using:

$$\widehat{Var}\,(\widehat{N}_x) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{N}_{b,x} - \widehat{N}_x\right)^2$$

where $\hat{N}_x$ is the average of $\hat{N}_{b,x}$ over the bootstrap samples.

3. Experimental study on the 2018 Italian Census and BRI data

A1 and A2 algorithms are applied to the 2018 Italian Census sample survey and BRI data restricted to the non-self-representative municipalities (municipalities with a population below 18,000 people).

In the experiment, for each pseudo-population, K = 100 under-coverage and K =100 over-coverage samples are drawn, and L =100 pseudo-populations are created.

At the end of the bootstrap iterations, we have 100 × 100 estimates $\hat{N}$. Those estimates are obtained by drawing 100 under- and over-coverage samples from the 100 generated pseudo-populations.

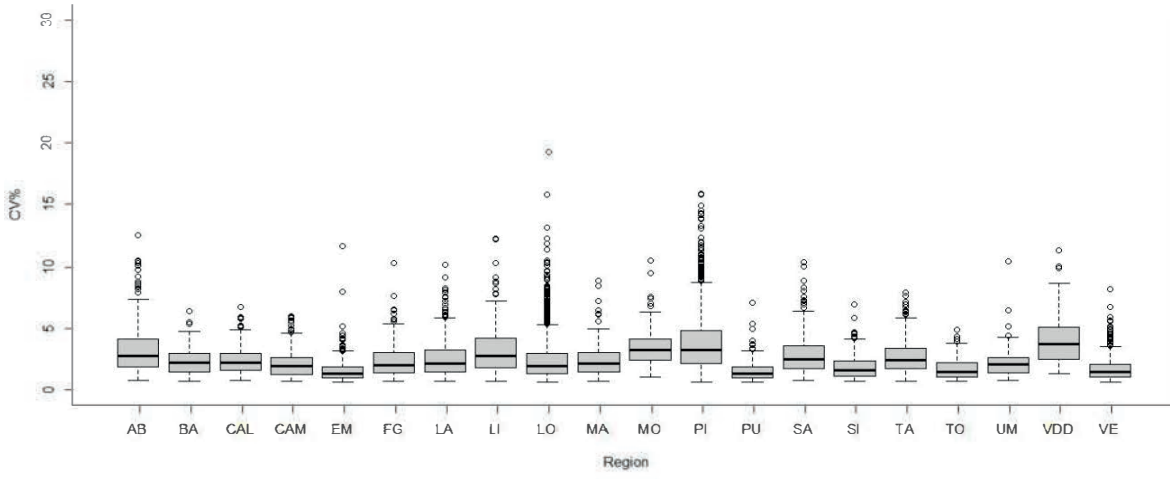
The variance calculated on the 10,000 $\hat{N}$ values is an estimate of the variance $V(\hat{N})$. This measure of dispersion is affected by the size of the municipality, then for comparisons, a standardised measure of dispersion is used:

$$CV = \frac{\sqrt{V(\hat{N})}}{E(\hat{N})}$$

The coefficient of variation CV is computed by dividing the standard deviation by the mean $E(\hat{N})$.

Figure 3.1 shows the coefficient of variation estimates ($\widehat{CV}$ %) in percentage for the 20 Italian Regions (Region names are abbreviated)[2] with algorithm A1.

Figure 3.1 – Boxplot of the Italian municipality's $\widehat{CV}$ % by Region. Algorithm A1



Source: Authors' processing on Istat Census data, 2018

2 AB as Abruzzo, BA as Basilicata, CAL as Calabria, CAM as Campania, EM as Emilia-Romagna, FG as Friuli-Venezia Giulia, LA as Lazio, LI as Liguria, LO as Lombardia, MA as Marche, MO as Molise, PI as Piemonte, PU as Puglia, SA as Sardegna, SI as Sicilia, TA as Trentino-Alto Adige/*Südtirol*, TO as Toscana, UM as Umbria, VDD as Valle D'Aosta/*Vallée d'Aoste*, VE as Veneto.

All the values are below 20%, with the Morterone municipality in the province of Lecco that is in the Lombardia (LO) region, assuming the maximum, 19.28 $\widehat{CV}$ % (Morterone $\hat{E}(\hat{N})$= 32.05 inhabitants). The municipalities with ($\widehat{CV}$ %>5 are 9% of the total, *i.e.* 657 on 7,362, with an expected mean size ranging between 32.05 to 539.00 inhabitants. On the contrary, for the 91% of municipalities with ($\widehat{CV}$ %<5, the expected mean size ranges between 269.6 to 17982.2, with a median size of 2463.90 inhabitants.

A useful decomposition of the squared CV, namely CV_2, derives from the law of total variance:

$$CV_2 = \frac{V(\hat{N})}{E^2(\hat{N})} = \frac{E\left(V\left(\hat{N}|p\right)\right)+V(E(\hat{N}|p))}{E^2(\hat{N})} \quad (2)$$

where *p* is the pseudo-population index. The first term on the right side synthetises the so-called within (pseudo-population) variance. The second term is interpretable as the between pseudo-population variance.

The estimates of the two terms are divided by the total $\hat{V}(\hat{N})$ obtaining for each $\hat{N}$ the fraction of total variance due to the within and the between components. The percentage of total variability attributable to the between pseudo-population variability prevails, ranging between 49.37% and 99.99% with 98% first quartile of the distribution.

In the pseudo-population bootstrap framework, it is possible to evaluate the pseudo-error occurred by the estimation. The comparison of the pseudo-population *p* values with the corresponding estimates allowing for a pseudo Mean Square Error (*MSE*), indicated as MSE_p. This quantity is estimated for each pseudo-population as follows:

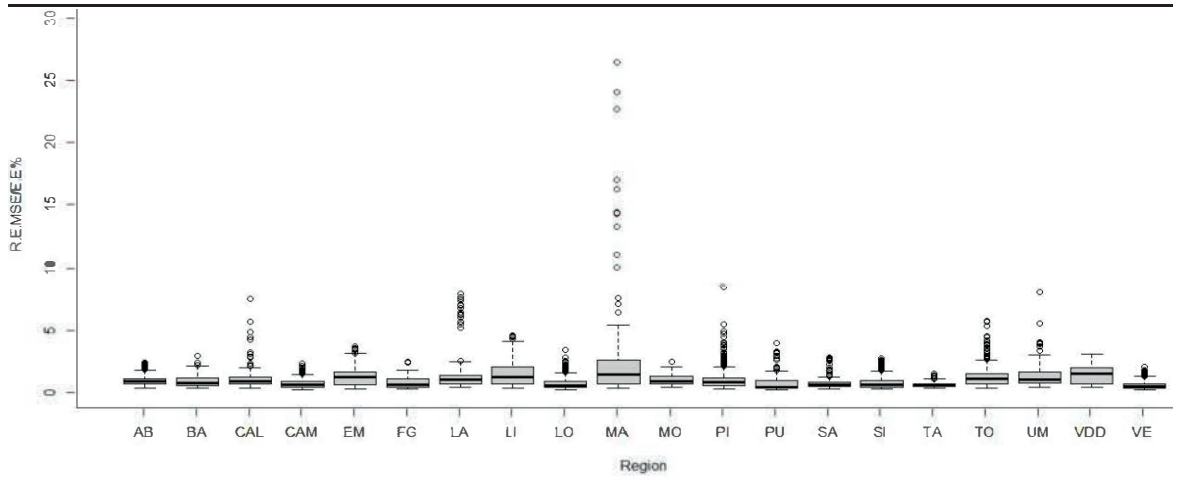
$$\widehat{MSE}_p = \frac{\sum_s^{100}(\widehat{N_{s,p}} - N_p)^2}{100}$$

where the sum index *s* is over the 100 sample estimates. Information on the fit of the model can be derived by the simulated $E(MSE_p)$ where the expectation is over pseudo-populations *p*. The square root of $E(MSE_p)$ is divided by $E(\hat{N})$ thus obtaining the root of the relative expected pseudo *MSE* (*REMSE*).

Figure 3.2 shows the values of $100*\frac{\sqrt{\sum_p^{100}\widehat{MSE}_p/100}}{m}$, where $m=\sum_{p,s}^{100}\frac{\widehat{N_{s,p}}}{10000} = \hat{E}(\hat{N})$. For 40 municipalities, corresponding to the 0.54% of 7362 analysed, the relative expected pseudo MSE percentage estimate is over 5%. Out of them, 10 municipalities belonging all to the Marche Region (MA) are over the 10%.

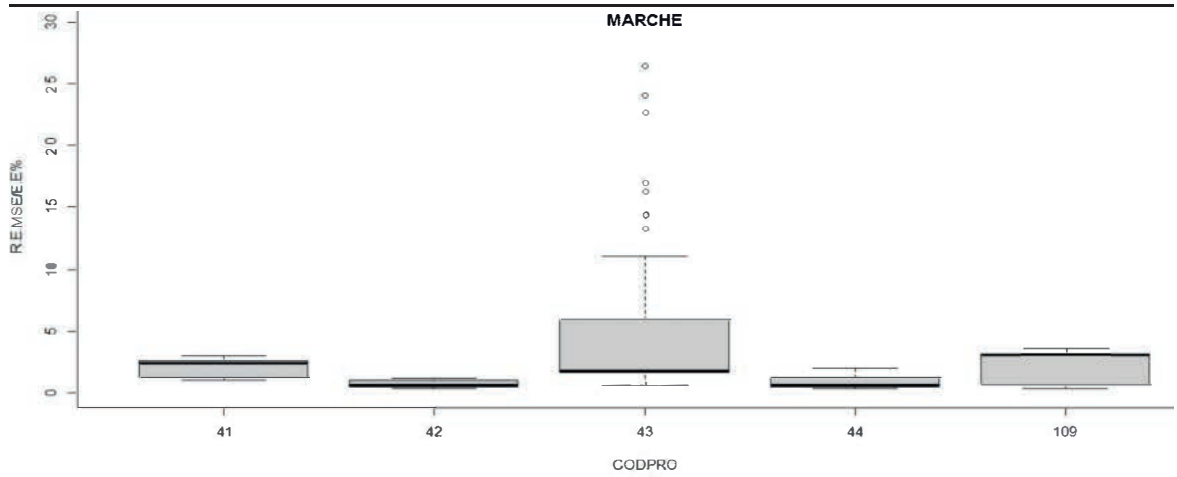
Figure 3.3 shows that the highest 10 values of $100*\frac{\sqrt{\sum_p^{100}\widehat{MSE}_p/100}}{m}$ are all referred to the Macerata Province (43) of Marche (MA) Region, 20% of the total 51 municipalities of this Province.

Figure 3.2 – Boxplot of relative expected pseudo MSE percent estimate for the Marche Region (MA), by Province



Source: Authors' processing on Istat Census data, 2018

Figure 3.3 – Boxplot of relative expected pseudo MSE percent estimate for the Marche Region (MA), by Province



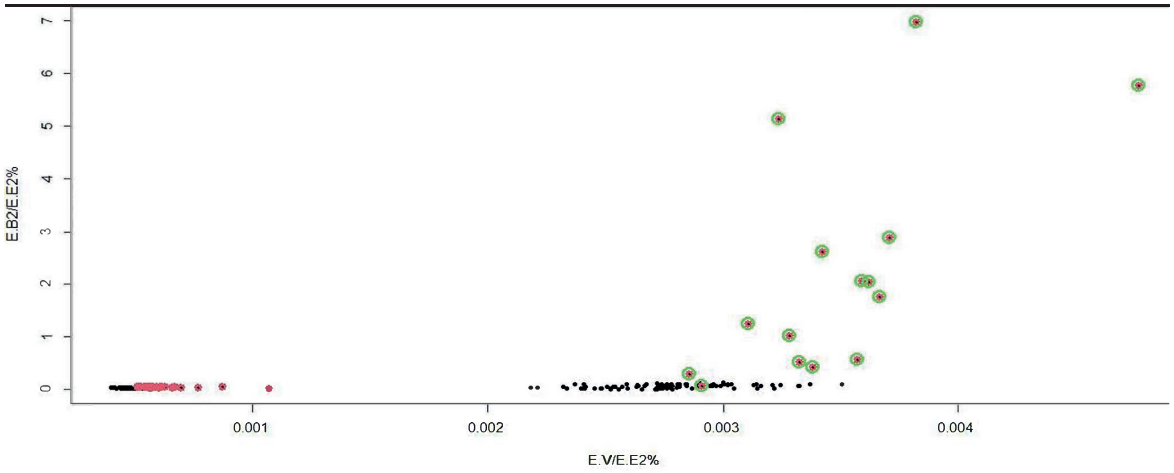
Source: Authors' processing on Istat Census data, 2018

To explain this result, we observe that again a decomposition is possible. Indeed, the expectation of classical MSE decomposition on variance and squared bias of estimator:

$$\frac{E(MSE_p)}{E^2(\widehat{N})} = \frac{E(E(\widehat{N} - N_p|p)^2)}{E^2(\widehat{N})} = \frac{E\left(V(\widehat{N}|p)\right) + E([N_p - E(\widehat{N}|p)]^2)}{E^2(\widehat{N})}$$

This decomposition can be a useful tool to detect and interpret critical cases. The scatter-plot of the estimates of the two components of the relative $E(MSE_p)$ for all municipalities of the Region are reported in Figure 3.4.

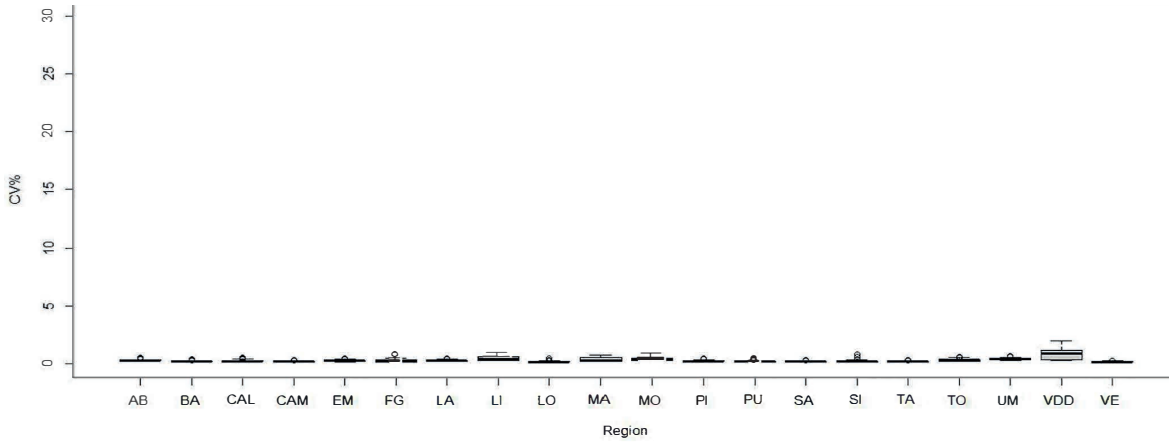
Figure 3.4 – Relative expected variance estimator percent (x-axis) vs relative expected bias percent (y-axis) estimates for Marche municipalities (a)



Source: Authors' processing on Istat Census data, 2018
(a) Red identifies the Macerata Province and green identifies rural municipalities (rural is not green).

We may notice that there are 10 observations clearly far from the bulk of data (outliers) and that it is mainly due to high values related to bias (y-axis). This information makes us return to check the model estimates, and we find that the critical Province (Macerata) is not able to explain the error (red points depict municipalities in this province), while if we consider the typology of municipality (in green rural, not green urban) we notice that errors are clustered with respect to this characteristic. Finally, we have observed an opposite behaviour on the parameter estimates concerning the municipality typology variable. This example is introduced to show how the bootstrap method proposed can have practical additional advantages in checking the model estimates. Figure 3.5 shows the CVs computed with A2. We notice that CVs are lower with respect to those computed with A1 (see Figure 3.1), as it was indeed expected. There is an average decrease of 2%.

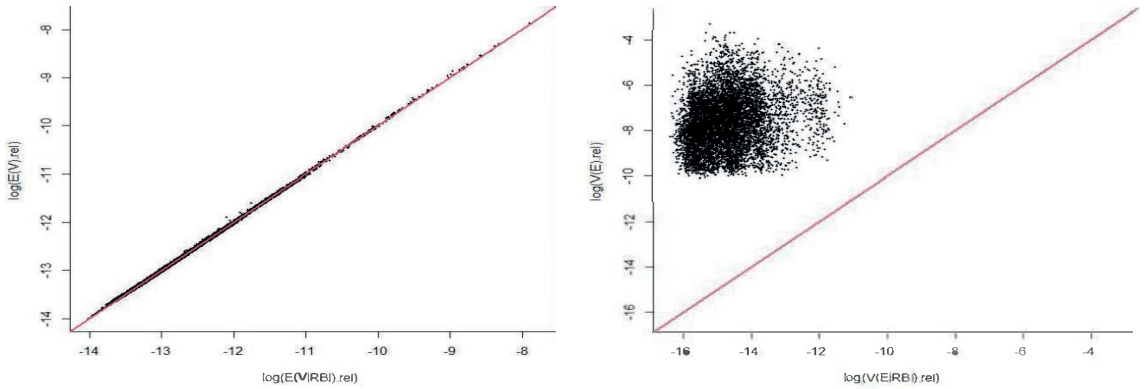
Figure 3.5 – Boxplot of $\widehat{CV}$ % for the Italian municipalities by Region. Algorithm A2



Source: Authors' processing on Istat Census data, 2018

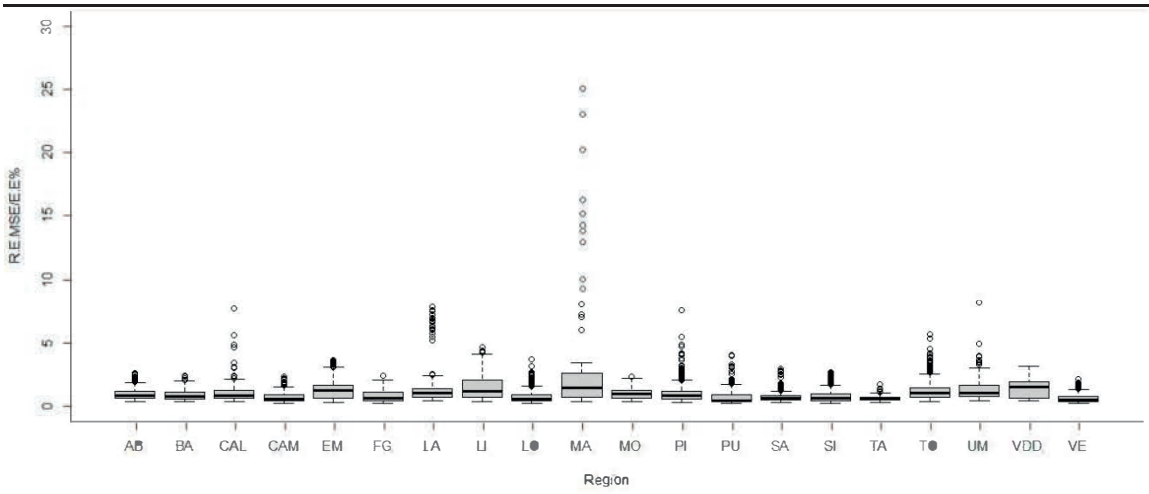
Figure 3.6 refers to the CV_2 decomposition, showing that the estimation of the first component, $E\left(V\left(\widehat{N}|p\right)\right)/E^2(\widehat{N})$, obtained by A1 and A2 are very similar (Figure on the left side). The Figure on the right compares the estimates of the second CV_2 component, *i.e.* between pseudo population variance, obtained by A1 and A2. The range of the A2 estimates (x-axis) are shifted far from that of A1 (y-axis). Then all the reduction in the CV_2 is clearly due to the stronger similarity between pseudo-population in A2.

Figure 3.6 – Left side: on the log scale, A2 (x-axis) vs A1 (y-axis) estimates of the relative between pseudo population variance. Right side: on the log scale, A2 (x-axis) vs A1 (y-axis) estimates of the relative within pseudo population variance (a)



Source: Authors' processing on Istat Census data, 2018
(a) In red y=x line.

Figure 3.7 – Boxplot of the root of relative expected pseudo MSE percent estimate for municipalities by Region. Algorithm A2



Source: Authors' processing on Istat Census data, 2018

The simulated pseudo REMSE are similar in A1 and A2 (see Figures 3.2 and 3.7) and this is for both the components reported above. Since we have already noticed that there is a sensible decrease in the CVs computed with algorithm A2, we may conclude that the most relevant part of simulated pseudo REMSE is the bias, and that this component is not affected by the variability induced by the first step (the results on the Marche estimates remain the same).

4. Classic bootstrap considering BRI without errors

In the procedures A1 and A2, BRI data are considered affected by errors, and consequently they are designed as random variables. The variance computation includes this further source of variability. Since an error measurement model is not available, in A1 the general structure of a multinomial distribution with probabilities estimated from data is used. In A2, the pseudo-population is randomly generated by using the estimated coverage probabilities.

On the other hand, we can model the problem by considering data in BRI as not affected by errors and the variance of the estimator only depends on the sampling randomness, *i.e.* in $\hat{N}_x = N_x^{RBI}$.

$\frac{1-p_{over,x}}{1-p_{under,x}}$, N_x^{RBI} should be treated as a constant.

Since the sampling rate is approximately 5%, a classic bootstrap procedure (sampling with replacement) can be applied to the coverage survey samples as a valuable alternative to the proposed pseudo-population approach. The algorithm used for this last approach is the following.

Algorithm A3.

a. From the list coverage survey, draw with replacement a sample $s^*_{b,\,list}$ of PSUs (municipalities) with sample size equal to the original sample size of PSUs;
b. From the area coverage survey, draw with replacement a sample $s^*_{b,\,area}$ of PSUs (municipalities) with sample size equal to the original sample size of PSUs;
c. Using the original logistic regression models, estimate $1\text{-}p^*_{b,over,x}$ and $1\text{-}p^*_{b,under,x}$ using $s^*_{b,\,list}$ and $s^*_{b,\,area}$;
d. Compute $\hat{N}^*_{b,x} = N_x^{RBI} \cdot \frac{1-p^*_{b,over,x}}{1-p^*_{b,under,x}}$;
e. Redo (a)-(d) B=10,000 times;
f. The bootstrap estimated variances of $\hat{N}_x$ is obtained using:
g. $\widehat{Var}(\hat{N}_x) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{N}^*_{b,x} - \hat{N}_x\right)^2$.

The results concerning the CV of estimates after 10,000 runs of the bootstrap procedure A3 are reported in Table 4.1, together with the others obtained with A1 and A2.

Table 4.1 – Summary statistics of $CV = \frac{\sqrt{V(\hat{N})}}{E(\hat{N})}$ of the Italian municipalities' estimates obtained with A1, A2 and A3

CV	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
A1	0.006582	0.014433	0.021618	0.026554	0.032894	0.193754
A2	0.000951	0.001620	0.002018	0.002354	0.002661	0.019672
A3	0.000992	0.001613	0.001927	0.002291	0.002566	0.020051

Source: Authors' processing on Istat Census data, 2018

We observe that the variance in A2 and A3 is almost the same, while A1 is much higher. That was expected because A1 assumes that data are affected by a measurement error, modelled through a multinomial distribution (step 1 of A1) and the procedure correctly includes this further source of variability.

It is useful to remark that, the difference between A1 and A2 is due to two different issues:

1. the use of either a classic or a pseudo-population bootstrap;
2. the pseudo-population generated in A2 is random and not unique as it is in the usual pseudo-population bootstrap.

In Table 4.2, we report the part of the CV estimates referred to the sampling contribution to the variability, that is $\sqrt{E\left(V\left(\widehat{N}|p\right)\right)}/E\left(\widehat{N}\right)$, the first term of the CV_2 decomposition, see formula (2). In this case, all approaches show strong coherence in the results despite the difference of the estimation procedure and assumptions.

Table 4.2 – Summary statistics of $\frac{\sqrt{E\left(v\left(\widehat{N}|\boldsymbol{p}\right)\right)}}{E\left(\widehat{N}\right)}$ computed on the Italian municipality's estimates obtained with A1, A2 and A3

CV	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
A1	0.000913	0.001544	0.001925	0.002244	0.002536	0.019608
A2	0.000904	0.001548	0.001921	0.002240	0.002536	0.019262
A3	0.000992	0.001613	0.001927	0.002291	0.002566	0.020051

Source: Authors' processing on Istat Census data, 2018

5. Discussion

In this paper, we study resampling approaches to compute the variance of an estimator used for the population count estimates based on the Base Register of Individuals. The estimator is obtained as a weighted sum of the register counts, with weights aimed at correcting the produced figures with respect to over- and under-coverage errors.

The resampling procedure is applied in different settings:

1. BRI is affected by measurement errors;
2. BRI is not affected by measurement errors.

Moreover, since the sampling rate of the surveys used for the Italian permanent census is not high, a version of the algorithm based on classic bootstrap is implemented and compared with the pseudo-population bootstrap approach proposed in the paper.

The pseudo-population algorithm adopted in the paper is a slight modification of the one introduced in the literature since we need to consider the problem of over- and under-coverage of the register. Classic and pseudo-population approaches give very similar results. On the other hand, if we consider the BRI as affected by an error, the variance of the estimator is much higher. As far as this last issue is concerned, we need to remind that we have not introduced any explicit error measurement model, but only introduced the variability through a multinomial model centred on the observed frequencies with the relative variances. With this general model, the level of randomness introduced can be too large if compared to the real level of errors in the BRI. In fact, the percentage of total variability attributable to the between pseudo population variability prevails, ranging between 49.37% and 99.99% with 98% first quartile of the distribution. Further studies will be devoted to the introduction of less general and more plausible error measurement models in order to introduce some randomness in a pseudo-RBI.

An issue that is extremely important to mention is that of the sampling design used for the sample selection in the bootstrap algorithm that is not the official one adopted in the census but a stratified random sampling. This design tends to underestimate the variances because it ignores the intracluster correlation induced by the use of two-stage sampling. Further experiments are planned in order to use a sampling design closer to the official one.

In the paper, we focussed on producing CV estimates by municipalities, and thus we resampled at the 2nd-stage sample (households/addresses), however, for any other domain of interest (*i.e.* other than municipalities) or characteristics, resampling should be done at the 1st-stage level. Details for a pseudo-population bootstrap in the context of a two-stage design can be found in Chauvet, 2007.

The approach of this study is reinterpreted and implemented in a Bayesian inference setting (Ballerini *et al.*, 2021) as well. In such a context, the evaluation of uncertainty of count estimates is condensed in their posterior distribution from which single dispersion measures like standard errors or credibility intervals can be easily computed.

References

Ballerini, V., S. Toti, M. Di Zio, and B. Liseo. 2021. "Correcting Italian municipalities' size for under- and over-coverage using administrative data: A Bayesian approach". Poster presented at the *XIV Conferenza Nazionale di Statistica*, Rome, Italy, 30th November - 1st December 2021.

Chauvet, G. 2007. "Bootstrap pour un tirage à plusieurs degrés avec échantillonnage à forte entropie à chaque degré". *Working Papers*, N. 2007-39. Palaiseau, France: Center for Research in Economics and Statistics - CREST.

Chen, S., D. Haziza, C. Léger, and Z. Mashreghi. 2019. "Pseudo-population bootstrap methods for imputed survey data". *Biometrika*, Volume 106, Issue 2: 369-384.

Gross, S.T. 1980. "Median estimation in sample surveys". In *Proceedings of the Survey Research Methods Section*: 181-184. Alexandria, VA, U.S.: American Statistical Association.

Mancini, L., e S. Toti. 2014. "Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del Censimento 2011". *Istat working papers*, N. 8/2014. Roma, Italy: Istat. https://www.istat.it/it/archivio/139548.

Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A survey of bootstrap methods in finite population sampling". *Statistics Surveys*, Volume 10: 1–52.

Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture". *Journal of Survey Statistics and Methodology*, Volume 3, Issue 4: 425-483.

R Development Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, J.N.K., and C.F.J. Wu. 1988. "Resampling Inference with Complex Survey Data". *Journal of the American Statistical Association*, Volume 83, Issue 401: 231-241.

Rao, J.N.K., C.F.J. Wu, and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys". *Survey Methodology*, Volume 18, N. 2: 209-217.

Särndal, C-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York, NY, U.S.: Springer-Verlag, *Springer Series in Statistics*.

Informazioni per le autrici e per gli autori

La collana è aperta alle autrici e agli autori dell'Istat e del Sistema statistico nazionale e ad altri studiosi che abbiano partecipato ad attività promosse dall'Istat, dal Sistan, da altri Enti di ricerca e dalle Università (convegni, seminari, gruppi di lavoro, etc.).

Coloro che desiderano pubblicare su questa collana devono sottoporre il proprio contributo al Comitato di redazione degli Istat working papers, inviandolo per posta elettronica all'indirizzo: iwp@istat.it.

Il saggio deve essere redatto seguendo gli standard editoriali previsti (disponibili sul sito dell'Istat), corredato di un sommario in Italiano e in Inglese e accompagnato da una dichiarazione di paternità dell'opera.

Per le autrici e gli autori dell'Istat, la sottomissione dei lavori deve essere accompagnata da un'e-mail della/del propria/o referente (Direttora/e, Responsabile di Servizio, etc.), che ne assicura la presa visione.

Per le autrici e gli autori degli altri Enti del Sistan la trasmissione avviene attraverso la/il responsabile dell'Ufficio di statistica, che ne prende visione. Per tutte le altre autrici e gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione.

Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del Chicago Manual of Style.

Attraverso il Comitato di redazione, tutti i lavori saranno sottoposti a un processo di valutazione doppio e anonimo che determinerà la significatività del lavoro per il progresso dell'attività statistica istituzionale.

La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line gratuitamente.

Gli articoli pubblicati impegnano esclusivamente le autrici e gli autori e le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.