UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
3-7 October 2022, (virtual)

Variance estimation for the mass imputation of the "Attained level of education" in the Italian Base Register of individuals: A comparison between analytical and MonteCarlo estimates

Prepared by Di Zio M., Filippini R., Toti S., (Istat, Italy)
dizio@istat.it, filippini@istat.it, toti@istat.it

I. Introduction

1. The Attained Level of Education (ALE) of the Permanent Italian Census relies on a high amount of administrative information. Nevertheless, it is necessary to resort to sample survey data to cope with delay of information and coverage gaps. Istat adopted a mass imputation approach integrating administrative and survey data for the ALE estimation of the Italian resident population, see Di Zio et al. (2019).

2. The procedure is based on a sequence of log-linear imputations. An evaluation of the variance of estimates is particularly relevant. Resampling methods are appealing for a complex imputation procedure like that used for the Italian ALE. In fact, roughly speaking, it is essentially only required to reproduce the procedure adopted by repeating the sampling and imputation phases. Nevertheless, given the high amount of data, it is not easily applicable in our context. Moreover, given the nature of register-based statistics that easily allows to compute estimates for many different domains of interest, it would be important to have an agile measure of uncertainty easily computable.

3. Scholtus et al. (2021) proposes an analytical formula for accuracy evaluation of a similar problem, but it is concerned with a logistic model and in a different and more simplified context than the one we are going to deal with. Those issues motivated our study towards an - as much as possible - simple analytical approach, even resorting to some approximation.

4. This paper details a proposal for variance estimation of imputed ALE for Italian residents and reports the experiments carried out to assess its validity and applicability.

5. The paper is structured as follows. Section II details the informative context and the mass-imputation procedure. Section III illustrates the proposal for the variance estimation. Section IV describes the experimental study carried out to evaluate the method proposed for the variance estimation and Section V reports and discusses the results.

II. Imputation of the Attained Level of Education (ALE)

A. Informative context

6. Core variables for each resident unit - such as place and date of birth, gender, and citizenship - are available from the Base Register of Individuals (BRI). The procedure for the ALE prediction is obtained by integrating different data: Administrative data, 2011 traditional Census data, and sample survey data.

(a) *Administrative data*. Administrative information on ALE is obtained by the Ministry of Education, University and Research (MIUR). MIUR provides information about ALE and course attendance for

people entering a study program after 2011 and covers the period from 2011 to *t-2,* scholar year (*t-2,t-1)*, where *t* is the reference year of the estimations.

(b) *2011 Italian Census data*. This is the last traditional Census conducted in Italy before the switch to the current 'Permanent Census' design. Those data are used for people who have not attended any courses since 2011 and, consequently, are not covered by the available administrative data so far introduced.

(c) *Sample survey data*. A sample survey is carried out to gather updated information, hence a direct measurement for ALE at time *t* for a subset of population (about 5%) is available. We refer to this sample survey as the census survey (CS).

7. The three sources of data are characterized by different patterns of information, i.e., a different set of variables and classifications of ALE. The structure of available information is summarized in Figure 1. Blue cells indicate that information is available for the specific subpopulation.

Figure 1. Structure of available information for mass-imputation of the attained level of education at time *t*

Source:	**BRI**	**MIUR**		**2011 Census**	**CS**	
Available information:	Core information *t*	ALE*t-2*	Course attendance *(t-2,t-1)*	ALE*t-2*	ALE*t*	**Sub population**
Coverage	■	■	■		■	A
		■		■	■	B
					■	C

8. The different information on ALE from 2011 to *t-2* determines the partition of the population of interest into three subgroups:

(a) Subgroup A is composed of all persons with administrative information on ALE from MIUR and is characterized by young people with longitudinal information on course attendance.

(b) Subgroup B is composed of persons not enrolled in any school course from 2011 to *t-2*, with information on ALE from MIUR at time *t-2* or from 2011 Census (this information can be considered approximately equal to ALE in time *t-2*).

(c) Subgroup C is composed of individuals neither in MIUR nor in 2011 Census. For this group, no direct information on ALE is available. Subgroup C is composed mainly of adults and is mainly characterized by a high percentage of Not Italian people.

9. The main difference is between group A and the others. Group A is composed by "Active" people, who are attending a course in academic year (*t-2, t-1)*, while groups B and C are "Inactive" people, not attending any course in the same period, which is the last available from administrative sources.

10. In all the subgroups, data on ALE were reclassified according to the 8-item classification adopted by Istat for the purpose of disseminating Permanent Census data. The classification is: 1 – Illiterate, 2 - Literate but no formal educational attainment, 3 - Primary education, 4 - Lower secondary education, 5 - Upper secondary education, 6 - Bachelor's degree or equivalent level, 7 - Master's degree or equivalent level, 8 - PhD level.

11. ALE, at reference time *t*, is only known for people interviewed in the Census sample, which is a representative subset of the population of interest. For the 95% of population not in the Census sample, ALE should be estimated.

B. Mass imputation procedure

12. Groups A, B and C are characterized by different patterns of information which determine different model specification for the estimation of ALE in t.

13. In group A, administrative data provide longitudinal information on school enrolment. Thanks to the great informative capacity of these administrative data, we decided to not resort to ALE observed in CS_t.

Information on ALE in the year *t-2* and information on year attendance of educational courses in academic year (*t-2, t-1)* are available for all individual in group A. This allows estimating the probability of obtaining a new qualification based on schooling characteristics of each individual.

14. Among them, a subset of individuals with zero probability of changing the educational level, from *t-2* to *t*, is identified, for instance people attending year 1, 2 or 3 of primary school (Primary education is acquired at the end of year 5). Therefore, for this subset of "No-Change" people (N-CNG), the imputation of ALE is deterministic and it is not necessary to estimate a model, since ALE in *t* is equal to ALE in *t-2*.

15. People belonging to group A and not included in the "No-Change" data set have a non-zero probability of obtaining a higher qualification than that held in year *t-2*. For each individual of this "Change" subset (CNG), the estimate of the probability distribution of achieving a new qualification in time *t* is based on individual characteristics and school attendance in academic year (*t-2,t-1)*. The model is estimated using only administrative sources. The underlying hypothesis is that the probability of obtaining a higher qualification between the years *t-2* and *t* is equal to that between the years *t-4* and *t-2*.

16. Group B and C are composed by "Inactive" people, which are people not enrolled in any course covered by MIUR in academic year (*t-2,t-1)*. It is worth reminding that, due to some informative gaps in administrative sources, there is a non-zero probability that an individual belonging to these groups is either enrolled in academic year (*t-2,t-1)* or has been enrolled in previous academic years in a school course not covered by MIUR.

17. For people in group B, information on educational level is available from administrative sources or from data collected in the 2011 Census. For people interviewed in the 2011 Census, who was enrolled in a school course covered by MIUR between 2011 and *t-2*, the most updated information on ALE comes from MIUR. For people not enrolled in any school course after 2011, the only available information on ALE refers to 2011. In both cases, this information may not be error free due to coverage error (MIUR) or response error (2011 Census). For this reason, the model is estimated on units interviewed in CS_t considering the observed ALE as target variable.

18. For people in group C, neither MIUR nor 2011 Census report information on ALE, so it is necessary to estimate a probability distribution of ALE for each pattern of available information on individual characteristics. ALE observed in CS_t is considered as target variable.

19. Finally, as a last step of imputation, for all individuals observed in the CS_t, the observed ALE is directly used as prediction.

C. Imputation based on log-linear model

20. The idea underlying the prediction of ALE is that of estimating a model for the prediction of ALE at time *t* given the values of covariates X. The official procedure adopted by Istat (Di Zio et al, 2019) is based on log-linear imputation. As stated in Singh (1988), this method generalizes hot-deck imputation by choosing suitable predictors for forming "optimal" imputation classes. The approach is based on modelling the associations between variables. In particular, we estimate the conditional probabilities $h(ALE_t \mid X)$ and then impute ALE_t by randomly taking a value from this distribution. The conditional probabilities $h(ALE_t \mid X)$ are estimated by means of log-linear models as follows.

21. First, a log-linear model is applied to the contingency table obtained by cross-classifying the variables (ALE_t, X) to estimate their expected counts $\hat{\eta}_{ij}^{ALE_t\, X}$, from which we estimate the counts $\hat{\eta}_j^X$. The estimated conditional probability distribution $\hat{h}(ALE_t \mid X)$ is easily obtained by computing $\hat{\eta}_{ij}^{ALE_t\, X} / \hat{\eta}_j^X$. This approach includes as a special case the random hot-deck when all the interactions between variables are included in the model (saturated log-linear model), but it has the advantage of allowing the use of more parsimonious models by testing the associations among variables. This is an important characteristic especially when the number of variables and contingency table's cells increase.

22. It is worthwhile noting that different log-linear models are used within groups A, B and C, mainly because of the different available information. As already remarked, in group A, a log-linear model is estimated

by using only administrative data, while for the other groups, log-linear models are estimated by using survey data as well.

23. For each subpopulation (CNG, B and C), variable selection is performed to detect the combination of covariates to be included in the model. The best log-linear model is chosen by means of cross-validation. More specifically, log-linear models for each subpopulation are built to estimate the following conditional probabilities:

(a) Subpopulation CNG: Pr (ALE_t| ALE_{t-2}, age, citizenship, school attendance)
(b) Subpopulation B: Pr (ALE_t| ALE_{t-2}, age, citizenship, province of residence, gender)
(c) Subpopulation C: Pr (ALE_t | age, citizenship, gender, apr, sirea).

24. Apr is an auxiliary information on ALE coming from an administrative source and it covers a particular subpopulation of individuals: those who changed their place of residence after 2014. It is a self-declared ALE and it comes with a more aggregate classification (4 levels[1]). Sirea refers to people who were targeted but not surveyed by the 2011 Census and were later detected by post-Census operations carried out in agreement with Italian Municipalities.

25. An in-depth analysis of the independent variables was necessary to appropriately reclassify the covariates in the model. In particular, suitable age levels were identified by taking into account the structure of the Italian school system and a classification in 14 levels was adopted[2]. Citizenship was aggregated into Italian/Not Italian to reduce the number of categories.

26. In order to impute all units, a sequence of log-linear imputation with a decreasing number of auxiliary variables are applied within each group. However, most units are imputed in the first step with the saturated model.

27. Finally, for the sub-set of units (referred to as CS_{t-1}) interviewed in the *t-1* Permanent Census Survey (and not in CS_t) to which the imputation process assigns a lower ALE than that observed in CS_{t-1}, to maintain the consistency between information, we decided to assign the ALE observed in CS_{t-1}.

III. Variance estimation of predicted ALE

28. The mass imputation procedure is complex and we would like having a very simple tool for the estimation of variance of ALE estimates even resorting to some approximation. The simplicity is a requirement useful for quickly computing the precision of estimates at different level of aggregation.

29. The estimation of ALE is obtained by combining the $\tilde{y}$ values predicted on the *N-n* units in BRI plus the *n* observed values y on the sample *s*. Expressing the frequency of an ALE modality as a mean:

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i\in s^c} \tilde{y}_i + \frac{1}{N}\sum_{i\in s} y_i = \frac{N-n}{N}\tilde{\bar{y}}_{s^c} + \frac{n}{N}\bar{y}_s \quad (1)$$

where $y_i = 1$ if the *i*-th unit takes the modality under evaluation and 0 otherwise (analogously for $\tilde{y}_i$).

30. $\hat{\bar{Y}}$ can be also written by considering a partition of the population in two subsets of units. The set *Det* (N-CNG) composed of units characterized by information that deterministically determine the ALE. For instance, people attending the first year of an elementary school will not be able to obtain in one year a higher level of

[1] Apr 4 levels of classification: 1- Up to primary education; 2 - Lower secondary education; 3 - Secondary and short cycle tertiary education; 4 - Tertiary and post tertiary education.
[2] Age levels: 0-8; 9-10; 11; 12-13; 14-17; 18; 19; 20-22; 23-25; 26-28; 29-39; 40-49; 50-69; 70-max

education. The other set Det^C (CNG ∪ B ∪ C) is composed of people that potentially can change their actual ALE. Hence

$$\hat{\bar{Y}} = \frac{N_{Det}}{N}\hat{\bar{Y}}_{Det} + \frac{N - N_{Det}}{N}\hat{\bar{Y}}_{Det^c}$$

31. In *Det,* there are deterministic predictions, and we notice also that a part of the sample will fall in this stratum. This set does not contribute to the variability of the estimator, and we may disregard it in the variance computation, that is:

$$V(\hat{\bar{Y}}) = \left(\frac{N-N_{Det}}{N}\right)^2 V\left(\hat{\bar{Y}}_{Det^c}\right) \qquad (2)$$

32. The part characterised by the probabilistic imputation $\hat{\bar{Y}}_{Det^c}$ can be decomposed by considering the following subpopulations, that are in fact the populations used in the imputation procedure: CNG, B, C

$$\hat{\bar{Y}}_{Det^c} = \frac{N_{CNG}}{N-N_{Det}}\hat{\bar{Y}}_{CNG} + \frac{N_B}{N-N_{Det}}\hat{\bar{Y}}_B + \frac{N_C}{N-N_{Det}}\hat{\bar{Y}}_C$$

hence

$$V\left(\hat{\bar{Y}}_{Det^c}\right) = \left(\frac{N_{CNG}}{N-N_{Det}}\right)^2 V\left(\hat{\bar{Y}}_{CNG}\right) + \left(\frac{N_B}{N-N_{Det}}\right)^2 V\left(\hat{\bar{Y}}_B\right) + \left(\frac{N_C}{N-N_{Det}}\right)^2 V\left(\hat{\bar{Y}}_C\right) \qquad (3)$$

$$\hat{\bar{Y}}_{CNG} = \frac{N_{CNG}-n_{CNG}}{N_{CNG}}\tilde{\bar{y}}_{CNG} + \frac{n_{CNG}}{N_{CNG}}\bar{y}_{s_{CNG}}, \quad \hat{\bar{Y}}_B = \frac{N_B-n_B}{N_B}\tilde{\bar{y}}_B + \frac{n_B}{N_B}\bar{y}_{s_B}, \quad \hat{\bar{Y}}_C = \frac{N_C-n_C}{N_C}\tilde{\bar{y}}_C + \frac{n_C}{N_C}\bar{y}_{s_C}$$

33. As far as $V\left(\hat{\bar{Y}}_B\right)$ and $V\left(\hat{\bar{Y}}_C\right)$ are concerned, we note that most of the imputations are carried out by means of a saturated log-linear model (see Di Zio et al., 2019). It means that a classic random donor hot-deck within imputation classes defined by the auxiliary variables chosen for each segment of the population is performed. Hence, we may adapt the basic formula for the variance estimation in presence of donor imputation (see Wolter, 2007, appendix F2, Brick et al., 2004) obtaining

$$V\left(\hat{\bar{Y}}_B\right) = \frac{\left(1-\frac{n_B}{N_B}\right)\sigma^2_{y_B}}{n_B} + \frac{1}{(N_B)^2}\sum_k\left(N_{x_k} - n_{x_k}\right)\left(1 - \frac{1}{n_{x_k}}\right)\sigma^2_{y_B|x_k} \qquad (4)$$

where x_k for k=1,…, K are the imputation cells, N_B is the population size of B and n_B is the size of the sample s falling in B, $\sigma^2_{y_B}$ is the variance of Y (ALE) in the population B, and $\sigma^2_{y_B|x_k}$ is the variance of Y in B within stratum x_k. An analogous formula can be derived for subpopulation C. If the auxiliary variable X is strongly connected to Y, an estimate for $V\left(\hat{\bar{Y}}_B\right)$ can be obtained by using the sampling variance of y within stratum x_k, $\sigma^2_{y_B|x_k}$, for both the terms. In the first term the conditional variance should be obtained by a weighted sum of conditional variances with weights given by the square of the size of the strata.

34. For the CNG subpopulation, a slightly different formula should be derived. We remind that in this subset the predictions are obtained by estimating a log-linear model on previous data, and by applying the estimated model to the actual data

$$\hat{V}(\hat{\bar{Y}}_{CNG}) = \frac{\left(\frac{N_{CNG}-n_{CNG}}{N_{CNG}}\right)^2\left(1-\frac{N_{CNG}-n_{CNG}}{N_{CNG}}\right)\hat{p}_{CNG}\hat{q}_{CNG}}{N_{CNG}-n_{CNG}} + \frac{\left(\frac{n_{CNG}}{N_{CNG}}\right)^2\left(1-\frac{n_{CNG}}{N_{CNG}}\right)\hat{\sigma}^2_{s_{CNG}}}{n_{CNG}} \quad (5)$$

35. This formula is similar to the previous one, but $\hat{p}_{CNG}$ is estimated only on administrative data at time *t-2* without resorting to the sample *s*, while $\hat{\sigma}^2_{s_{CNG}} = \hat{p}_{s_{CNG}}\hat{q}_{s_{CNG}}$ are the frequencies estimated by using units of sample *s* that are in CNG.

36. The last identified subpopulation is CS_{t-1}. For simplicity, in this experimentation we just consider the sample variance for this CS_{t-1} subpopulation. Given the small number of individuals in this subpopulation, this will have a small impact on the final variance estimation.

IV. Experimental study

37. In this experiment, the variance estimation procedure is applied to the BRI dataset, referred to people with age greater than or equal to 9 and resident in the Emilia Romagna region in year 2019. This dataset is composed by 4,141,737 individuals. For this population the official imputed ALE is available, and it is considered as it was the true value.

38. From this reference population, 200 simple random samples are extracted. The sample size is set at national level considering the same sampling frequency observed in the survey for the permanent census, equal to 4.7%.

39. For each of the 200 samples, the ALE mass imputation procedure is applied, and the resulting frequency distribution of ALE is calculated. For each of the 200 simulations, the ALE frequency, $\hat{\bar{Y}}_{sim}$, *sim*=1,…,200, is computed. The variance of $\hat{\bar{Y}}$ can be calculated by following two approaches:

(a) Monte Carlo approach (MC);
(b) Analytical approach, applying the formulas described in Section III.

The processing time for the simulation procedure takes about 2 days (each simulation on the Emilia Romagna region lasts about 14 minutes). The variance estimates on the simulated datasets are obtained using ad-hoc R code (R Core Team 2022).

40. The MC estimates of $V(\hat{\bar{Y}})$ are easily obtained by applying the classical variance formula

$$\hat{V}_{MC}\left(\hat{\bar{Y}}\right) = \frac{\sum_{sim=1}^{200}(\hat{\bar{Y}}_{sim}-\bar{\bar{Y}})^2}{199}.$$

where $\bar{\bar{Y}}$ is the average of $\hat{\bar{Y}}_{sim}$.

Calculations are carried out for the total population and for each specific subpopulations: N-CNG, CNG, B, C and CS_{t-1} (that is CS_{2018}). The results obtained by the MC approach are considered as the benchmark to evaluate the analytical results.

41. In the analytical approach, for each simulation, the specific subpopulations are identified and within each subpopulation, the analytical estimate of the variance is obtained.

(a) For B and C subpopulations, the values of N_{sim}, n_{sim} and $\sigma^2_{y|x_k\ sim}$, *sim*=1,…,200, are identified, thus $\hat{V}(\hat{\bar{Y}}_B)_{sim}$ and $\hat{V}(\hat{\bar{Y}}_C)_{sim}$ are computed.
(b) For CNG subpopulation, the values of N_{sim}, n_{sim}, $\sigma^2_{s\,\mathrm{sim}}$ and $\hat{p}_{\mathrm{sim}}$, *sim*=1,…,200, are identified and $\hat{V}(\hat{\bar{Y}}_{CNG})_{sim}$ is computed.
(c) For CS_{2018} subpopulation, we just consider the sample variance, so N_{sim}, n_{sim} and $\hat{p}_{\mathrm{sim}}$ are identified and $\hat{V}(\hat{\bar{Y}}_{CS2018})_{sim}$ is computed. Further analysis should be carried out to better understand the impact on the variance of the specific characteristics of the imputation process.

(d) We remind that for N-CNG subpopulation, given the deterministic imputation of ALE, we assume a variance equal to 0.

42. To obtain the total estimate of the variance of ALE within each simulation, $\hat{V}(\hat{\bar{Y}})_{sim}$, the variances obtained for each subpopulation are combined according to formula (2) and (3).

43. The final analytical estimate of $\hat{V}(\hat{\bar{Y}})$ - compared with the MC estimate $\hat{V}_{MC}\left(\hat{\bar{Y}}\right)$ - is the mean value of $\hat{V}(\hat{\bar{Y}})_{sim}$ computed on the 200 simulations:

$$\hat{V}_{an}\left(\hat{\bar{Y}}\right) = \frac{\sum_{sim=1}^{200} \hat{V}(\hat{Y})_{sim}}{200}.$$

V. Results of the experimental study

44. The variance estimates obtained by applying the analytical approach are compared to those obtained by the MC approach. The MC approach is considered as benchmark.

45. The estimated variance of the mass imputation of ALE is very low. Thus, the imputation procedure gives origin to very stable results for each of the 8 ALE modality. Moreover, the analytical approach shows a good approximation of the estimates if compared with the MC results (Table 1).

46. In some cases, the analytical approach slightly overestimates the MC variance (Illiterate, Literate but no formal educational attainment, Primary education and Master's degree or equivalent level). In two cases (Bachelor's degree or equivalent level and PhD level), the variance is underestimated. In any case, the differences are always low; the highest is observed for Master's degree or equivalent level. The best results are obtained for the ALE's modalities where the frequency of the population is high.

Table 1. Results of the estimates for the mass-imputation of ALE: MC and Analytical variance estimates, ratio between Analytical and MC variances, number of individuals for the reference population

ALE	$\hat{V}(\hat{\bar{Y}})$		Ratio (Analytical/MC)	N of reference population (.000)
	MC	Analytical		
1 Illiterate	1 .42E-08	1.53E-08	1.08	17
2 Literate but no formal educational attainment	6 .62E-08	7.34E-08	1.11	162
3 Primary education	1 .82E-07	1.92E-07	1.06	660
4 Lower secondary education	3 .78E-07	3.84E-07	1.01	1,152
5 Upper secondary education	3 .69E-07	3.66E-07	0.99	1,514
6 Bachelor's degree or equivalent level	7 .49E-08	7.21E-08	0.96	166
7 Master's degree or equivalent level	9 .81E-08	1.17E-07	1.19	451
8 PhD level	1 .16E-08	1.09E-08	0.94	20

47. The comparisons is performed at the subpopulation level as well (Table 2). Results show a good approximation of the analytical approach for subpopulations B, C and CS_{2018}, while in the CNG subpopulation the difference between MC and analytical estimates is higher for ALE's modalities 3 and 4.

48. We note that modality 1 (Illiterate) and 2 (Literate but no formal educational attainment) are ambiguous and difficult to attribute with certainty. Administrative sources do not allow to distinguish between them and the difference between MC and Analytical estimates are higher in the CNG subpopulation. Moreover in this subpopulation modality 1 and 2 are less frequent.

49. We observe that in general the variance computation is composed of a part measuring the sample variability and a part measuring the imputation variability. For the CNG, the part concerned with imputation variability is computed with parameters estimated on administrative sources and not from the sample.

Table 2. Results of the estimates for the mass-imputation of ALE by subpopulation: MC and analytical variances estimates

ALE	CNG		B		C		CS_{2018}	
	MC	Analytical	MC	Analytical	MC	Analytical	MC	Analytical
1	1 .39E-12	1 .36E-12	1 .16E-08	1 .68E-08	2 .08E-06	1 .61E-06	9 .83E-08	2 .40E-07
2	8 .26E-11	8 .00E-09	7 .14E-08	9 .25E-08	5 .70E-06	5 .14E-06	3 .85E-07	1 .21E-06
3	1 .14E-09	2.03E-08	2 .79E-07	2 .72E-07	5 .70E-06	7 .34E-06	2 .66E-06	2 .98E-06
4	4 .35E-09	2.71E-08	5 .03E-07	5 .21E-07	2 .01E-05	1 .90E-05	5 .49E-06	4 .52E-06
5	1 .37E-08	3.15E-08	5 .14E-07	4 .92E-07	1 .88E-05	1 .91E-05	4 .10E-06	4 .47E-06
6	9 .02E-09	1.46E-08	9 .72E-08	9 .76E-08	4 .56E-06	3 .58E-06	5 .46E-07	1 .11E-06
7	4 .22E-09	3.56E-09	1 .37E-07	1 .47E-07	6 .96E-06	8 .29E-06	2 .57E-06	1 .83E-06
8	1 .22E-10	9.05E-11	1 .40E-08	1 .50E-08	8 .32E-07	4 .93E-07	5 .83E-08	1 .67E-07

50. The results of the experiments are encouraging and suggest that the proposed method for the variance estimation of ALE for the permanent Census, can be taken into account for further studies. The analytical formula is particularly easy to implement and it is not time consuming as a bootstrap approach. This is a particularly important aspect for a complex case like that under evaluation. Further experiments should be performed to put the proposed method into the process, more studies should be referred to the variance estimation of ALE within domains composed of a moderate/small number of units as for instance Municipalities.

References

Brick, J. M., Kalton, G., & Kim, J. K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30(1), 57-66.

Di Zio M., Filippini R., Rocchetti G. (2019) An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. *Rivista di Statistica Ufficiale*. 2-3: pp.143-174.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scholtus S. and J. Pannekoek (2015). Mass-imputation of educational levels (In Dutch), Statistics Netherlands, Internal report, The Hague/Heerlen.

Scholtus, S., & Daalmans, J. (2021). Variance Estimation after Mass Imputation Based on Combined Administrative and Survey Data. Journal of Official Statistics (JOS), 37(2).

Singh A. C. Log-linear imputation. Methodology Branch Working Paper Statistics Canada. 1988; 88-29.

Wolter, K.M. (2007). *Introduction to variance estimation*. New York, Springer.