Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Variance estimation after mass imputation with an application to the Dutch population census

Prepared by Sander Scholtus and Jacco Daalmans (Statistics Netherlands)

I. Introduction

1. In this paper, we discuss methods for evaluating the design-based variance of estimated frequency tables based on mass imputation. The motivating application for this study is the Dutch decennial virtual population census. Since 1981, the Dutch census tables have been estimated by re-using data from existing sources rather than collecting data with a dedicated questionnaire. Nowadays, most variables needed for the census are available from administrative sources with (near-)complete population coverage. An exception occurs for *educational attainment*, which is observed partly in education registers and partly in the Labour Force Survey (LFS). For about 7 million Dutch persons (of a total of 17 million), educational attainment is not observed.

2. In the Dutch censuses of 2001 and 2011, the so-called repeated weighting method was used to estimate consistent sets of census tables using a combination of registers and survey data (Schulte Nordholt et al., 2004 and 2014). It is known that repeated weighting has practical limitations for the estimation of large numbers of high-dimensional frequency tables, and these limitations were indeed encountered during the 2011 census (Daalmans, 2018). As an alternative approach, it has been proposed for the next census to use mass imputation to predict all missing values of educational attainment, using both register data and LFS data (de Waal and Daalmans, 2018). This naturally leads to the question how to determine the precision of estimated entries in frequency tables based on mass-imputed data.

3. One approach is to try to use analytical approximations to estimate the variance. For the particular case of frequency tables, we will present a variance formula for mass-imputed estimates in Section II. This formula is of interest for a class of problems in which a categorical variable is mass-imputed, using one or more auxiliary variables. It is assumed here that a model is applied that predicts, for each record, the probabilities for each of the categories of the target variable; for instance, a logistic regression model. Imputations for missing values are then drawn based on these predicted probabilities.

4. In general, a drawback of analytical variance estimation is that a separate expression has to be derived for each estimator. In fact, for some estimators based on combined data, deriving an adequate variance expression may prove to be impossible. Therefore, a resampling method such as the bootstrap may be a more attractive option. In Section III, a generic bootstrap method will be described for this purpose. Unlike the analytical approach of Section II, this bootstrap approach does not depend on the specific context of mass imputation of a categorical variable and could therefore easily be applied to many other estimators based on combined administrative and survey data.

5. Both approaches have been tested on a small artificial population in a simulation study (Section V). In Section VI, we will describe an application of both approaches to real data of the Dutch virtual census 2011. Both applications in Sections V and VI make use of a specific imputation model that has

been proposed for the next Dutch census, the so-called continuation-ratio model; this model will be outlined in Section IV. Some conclusions follow in Section VII.

II. Analytical variance estimation for mass-imputed tables

6. We will now consider the estimation of frequency tables based on mass imputation in more detail. Let $\theta_{hc} = \sum_{i \in U} h_i y_{ci}$ denote the true count in a particular cell of a table involving our target variable. Here, U is the target population and y_c is an indicator variable such that $y_{ci} = 1$ if person i belongs to category c of the target variable and $y_{ci} = 0$ otherwise ($c = 1, \ldots, C$, where C denotes the total number of categories). Also, h is an indicator variable for the cross-classification of all other variables in the table, i.e. $h_i = 1$ if person i contributes to a cell according to these variables and $h_i = 0$ otherwise.

7. As is common in official statistics, we will consider the target population and the values of variables for units in this population as fixed. All variables other than the target variable represented by $y_1, \ldots, y_C$ are supposed to be completely observed for all units in the population. The target variable is partially observed. In general, it could be observed for some units in an administrative source and for other units in a sample survey. We suppose that missing values on the target variable are imputed throughout the population (i.e., mass imputation). After mass imputation, θ_{hc} is estimated as follows:

$$\hat{\theta}_{hc} = \theta_{hc1} + \hat{\theta}_{hc2} = \sum_{i \in U_1} h_i y_{ci} + \left(\sum_{i \in S_2} h_i y_{ci} + \sum_{i \in U_2 \backslash S_2} h_i \tilde{y}_{ci} \right). \quad (1)$$

Here, U_1 consists of all persons in U for which the target variable is observed in a register. From the remaining subpopulation, $U_2 = U \backslash U_1$, a probability sample S_2 is available with observed values of the target variable. Finally, for all $i \in U_2 \backslash S_2$, the indicator y_{ci} is unknown and replaced by an imputation $\tilde{y}_{ci}$.

8. We consider the register part of the population, U_1, as fixed. In general, S_2 may be a subsample of a sample S drawn from U. Note that the size of the register overlap $S_1 = S \cap U_1$ may then be random.

9. In the application to be discussed in Section VI, expression (1) is used to estimate a table in the Dutch virtual census, where $y_1, \ldots, y_C$ represent levels of educational attainment. In this case, the imputations are based on a model that is estimated only on data from S_2, as the register data are known to be selective. The results in this section can be applied to any frequency table involving a categorical variable that is mass-imputed. This includes the special case of only sample survey data ($U_1 = \emptyset$).

10. In general, we suppose that the missing values of y_{ci} are imputed by drawing – independently for each person $i \in U_2 \backslash S_2$ – a vector $(\tilde{y}_{1i}, \ldots, \tilde{y}_{Ci})$ from a multinomial distribution with predicted probabilities $(\hat{p}_{1i}, \ldots, \hat{p}_{Ci})$, so that for each i exactly one of the values $\tilde{y}_{ci}$ is equal to 1 and the other values are equal to 0. The predicted probabilities $\hat{p}_{ci}$ are based on the observed distribution of y_c in the sample S_2, using some model. An example of an imputation model will be discussed in Section IV.

11. It is possible to evaluate the variance of $\hat{\theta}_{hc} - \theta_{hc}$ analytically. Here, we make the simplifying assumption that the predicted probabilities $\hat{p}_{ci}$ in S_2 add up to the observed totals for each category:

$$\sum_{i \in S_2} h_i (\hat{p}_{ci} - y_{ci}) = 0, \quad (c = 1, \ldots, C). \quad (2)$$

For many commonly used models, this assumption may be reasonable, at least as an approximation. For instance, it is satisfied exactly when imputation is based on a logistic regression model, fitted to a simple random sample S_2 by maximum likelihood, which includes the cross-classified variable h as a predictor (see, e.g., Agresti, 2013, pp. 192–193). Moreover, we note that assumption (2) can be verified in practice.

12. Under this assumption, the following formula is derived in Scholtus and Daalmans (2020):

$$\mathrm{var}(\hat{\theta}_{hc} - \theta_{hc}) = E\left\{ \sum_{i \in U_2 \backslash S_2} h_i \hat{p}_{ci} (1 - \hat{p}_{ci}) \right\} + \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \, \mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj}). \quad (3)$$

The first term reflects the additional variance due to imputing draws rather than predicted probabilities. To complete the specification of $\mathrm{var}(\hat{\theta}_{hc} - \theta_{hc})$, an expression is needed for $\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ in the second

term in (3). The precise form of this covariance depends on the imputation model and the sample design. For the proposed imputation model for the Dutch census, this is discussed further in Section IV.

13. Based on the observed data, the variance in (3) could be estimated as follows:

$$\widehat{\mathrm{var}}\left(\hat{\theta}_{hc} - \theta_{hc}\right) = \sum_{i \in U_2 \setminus S_2} h_i \hat{p}_{ci}(1 - \hat{p}_{ci}) + \sum_{i \in U_2} \sum_{j \in U_2} h_i h_j \, \widehat{\mathrm{cov}}\left(\hat{p}_{ci}, \hat{p}_{cj}\right). \quad (4)$$

Here, $\widehat{\mathrm{cov}}\left(\hat{p}_{ci}, \hat{p}_{cj}\right)$ denotes an estimator of $\mathrm{cov}\left(\hat{p}_{ci}, \hat{p}_{cj}\right)$. Again, the precise form of this term depends on the imputation model and the sample design.

III. A bootstrap method

14. The classical bootstrap (see, e.g., Efron and Tibshirani, 1993, for an introduction) uses resampling with replacement from an original sample to approximate the sampling distribution of a target estimator. This method cannot be used directly here, as it does not account for finite-population sampling. In fact, mass imputation is meaningful only in the context of a finite population.

15. Different extensions of the bootstrap to finite-population sampling have been developed; see Mashreghi et al. (2016) for a recent overview. For estimators that involve weighting or imputation, a particularly useful extension is based on generating pseudo-populations. This methodology was developed by, among others, Gross (1980), Booth et al. (1994), and Chauvet (2007). At Statistics Netherlands, Kuijvenhoven and Scholtus (2011) applied this type of bootstrap method to combined register and LFS data on educational attainment, for various estimators based on weighting. Here, we will describe a slight extension of their method that can accommodate more general estimators.

16. As a generalisation of (1), suppose that the estimator of interest is $\hat{\theta} = t(S, U_1)$, for some functional $t(.)$. The underlying finite-population parameter is θ. Let π_i denote the inclusion probabilities of sample S, and write the design weights $w_i = 1/\pi_i$ as $w_i = \lfloor w_i \rfloor + \varphi_i$, with $\lfloor w_i \rfloor \in \mathbb{N}$ and $\varphi_i \in [0,1)$. Here, $\lfloor z \rfloor$ denotes the integer part of $z \in \mathbb{R}$, i.e., the largest integer that is smaller than or equal to z. The bootstrap algorithm consists of the following steps.

For each $a = 1, \ldots, A$ do the following:

(a) Create a pseudo-population $\widehat{U}_a^*$ by taking ω_i copies of each unit $i \in S$, where the random inflation weight ω_i is drawn as $\omega_i = \lfloor w_i \rfloor$ with probability $1 - \varphi_i$ and $\omega_i = \lfloor w_i \rfloor + 1$ with probability φ_i.

(b) For each $b = 1, \ldots, B$ do the following:

- Draw a sample S_{ab}^* from $\widehat{U}_a^*$ according to the same design that was used to draw S from U. For $k \in \widehat{U}_a^*$ the inclusion probability is chosen to be $\pi_k^* \propto \pi_i$, with i the unit in the original sample S of which unit k is a copy. Here, the proportionality constant is chosen so that $\sum_{k \in \widehat{U}_a^*} \pi_k^* = |S|$ holds.
- Analogously to the original estimation procedure yielding $\hat{\theta} = t(S, U_1)$, construct the bootstrap replicate $\hat{\theta}_{ab}^* = t(S_{ab}^*, U_1)$.

(c) Compute the variance estimate for $\hat{\theta} - \theta$ based on pseudo-population $\widehat{U}_a^*$ as $v_a\left(\hat{\theta} - \theta\right) = (B-1)^{-1} \sum_{b=1}^{B} \left(\hat{\theta}_{ab}^* - \overline{\hat{\theta}_a^*}\right)^2$, with $\overline{\hat{\theta}_a^*} = B^{-1} \sum_{b=1}^{B} \hat{\theta}_{ab}^*$.

Compute the final variance estimate for $\hat{\theta} - \theta$ by averaging over the pseudo-populations: $\widehat{\mathrm{var}}_{boot}\left(\hat{\theta} - \theta\right) = A^{-1} \sum_{a=1}^{A} v_a\left(\hat{\theta} - \theta\right)$.

17. The outer for loop of this algorithm is intended to reduce the noise due to the random assignment of integer-valued inflation weights to units with non-integer sampling weights in Step (a). Previous results in Chauvet (2007) and Kuijvenhoven and Scholtus (2011) suggest that this additional for loop may have little added value in practice (i.e., choosing $A = 1$ leads to variance estimates of a similar accuracy as choosing $A > 1$). For variance estimation, $B = 200$ replicates are often considered sufficient in the bootstrap literature (Efron and Tibshirani, 1993, Section 6.4). If the sample S is based on a multi-stage design, an extended version of the above pseudo-population approach may be used to account for clustering (Chauvet, 2007; Mashreghi et al., 2016).

18. The contents of the second point in Step (b) depend on the original estimation procedure. For $\hat{\theta}_{hc}$ in (1) based on mass imputation, in this step we basically re-estimate the imputation model and use this to impute the missing values in the pseudo-population. In general, the bootstrap sample S^*_{ab} may contain copies of units from $S \backslash S_2$, i.e., units that overlap with the register part of the population. In analogy with the original imputation procedure, only the subset of units in S^*_{ab} that originate from S_2, say S^*_{2ab}, is used to re-estimate the imputation model. Similarly, only the missing values for the subset of units in the pseudo-population $\widehat{U}^*_a$ that originate from S_2, say $\widehat{U}^*_{2a}$, are subject to imputation. The missing values in $\widehat{U}^*_{2a}$ occur for those units that are not contained in the bootstrap sample S^*_{2ab}, i.e., $\widehat{U}^*_{2a} \backslash S^*_{2ab}$. The register part of the pseudo-population is not imputed in the bootstrap procedure, as the contribution of the register part to $\hat{\theta}_{hc}$ is considered fixed. Thus, the bootstrap replicate $\hat{\theta}^*_{hc,ab}$ is computed in this step as $\hat{\theta}^*_{hc,ab} = \sum_{k \in U_1} h_k y_{ck} + \left(\sum_{k \in S^*_{2ab}} h_k y_{ck} + \sum_{k \in \widehat{U}^*_{2a} \backslash S^*_{2ab}} h_k \tilde{y}_{ck} \right)$, analogously to (1).

19. The bootstrap method is straightforward to implement and can in fact re-use most of the code that was created to compute the original estimates. It is a computationally intensive method. A potentially useful aspect is that the time-consuming parts of the above bootstrap algorithm have to be performed only once. For instance, with mass imputation, the mass-imputed pseudo-populations could be stored and used to compute a variance estimate for any estimator $\hat{\theta}_{hc}$ by generating the replicates $\hat{\theta}^*_{hc,ab}$ 'on the fly'.

IV. The continuation-ratio model for imputation

20. In Sections V and VI we will present applications of the variance estimation methods of Sections II and III to simulated and real data. In these applications, use is made of the imputation approach proposed for educational attainment in the next Dutch virtual census, which we will now discuss.

21. The imputation approach is based on logistic regression. Since educational attainment has $C > 2$ categories, the binomial logistic regression model cannot be applied directly. To account for the fact that educational attainment is an ordinal variable, de Waal and Daalmans (2018) proposed to use an extension of logistic regression known as the continuation-ratio model.

22. The continuation-ratio logistic regression model (Agresti, 2013, Section 8.3.6) consists of $C - 1$ ordinary binomial logistic regression models. Each of these binomial models refers to the conditional probability q_{ci} that person i does not attain a higher level than a particular level c, given that this person at least reached level c ($c = 1, \ldots, C-1$):

$$q_{1i} = P(y_{1i} = 1 | \mathbf{x} = \mathbf{x}_i),$$
$$q_{ci} = P\left(y_{ci} = 1 \middle| y_{1i} = \cdots = y_{(c-1)i} = 0, \mathbf{x} = \mathbf{x}_i\right), \quad (c = 2, \ldots, C-1).$$

Here, $\mathbf{x}_i$ denotes a vector of auxiliary variables used in the model. Note that each conditional probability q_{ci} refers to a binary choice ($y_{ci} = 1$ or $y_{ci} = 0$). The continuation-ratio logistic regression model thus consists of a sequence of models of the form:

$$\log\left(\frac{q_{ci}}{1 - q_{ci}}\right) = \boldsymbol{\beta}_c^T \mathbf{x}_i, \quad (c = 1, \ldots, C-1). \tag{5}$$

23. Agresti (2013) noted that maximum likelihood estimates of all parameters in the continuation-ratio model can be obtained by estimating the $C - 1$ binomial logistic regression models in (5) separately, each of them being estimated on the subset of the sample that satisfies the relevant condition of the form $y_{1i} = \cdots = y_{(c-1)i} = 0$. From the estimated model parameters, the conditional probability that a person with characteristics $\mathbf{x}_i$ has education level c may then be predicted by

$$\hat{q}_{ci} = \frac{\exp\left(\widehat{\boldsymbol{\beta}}_c^T \mathbf{x}_i\right)}{1 + \exp\left(\widehat{\boldsymbol{\beta}}_c^T \mathbf{x}_i\right)}, \quad (c = 1, \ldots, C-1).$$

Subsequently, predictions for the marginal probabilities $p_{ci} = P(y_{ci} = 1 | \mathbf{x} = \mathbf{x}_i)$ as used in Section II can be derived by the following recursive relation: $\hat{p}_{1i} = \hat{q}_{1i}$,

$$\hat{p}_{ci} = \hat{q}_{ci} \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right), \quad (c = 2, \ldots, C-1),$$

and finally $\hat{p}_{Ci} = 1 - \sum_{c=1}^{C-1} \hat{p}_{ci}$.

24. To account for finite-population sampling – possibly with a complex survey design – pseudo maximum likelihood estimation can be used (Chambers and Skinner, 2003). Under this approach, a large-sample estimate of $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ for the continuation-ratio model is given in the appendix to this paper. The computation of these estimated covariances involves a recursive algorithm over $c = 1, \ldots, C$. Moreover, to evaluate expression (4) this algorithm has to be run for each pair $(i \in U_2, j \in U_2)$, or at least each pair with $h_i = h_j = 1$, which may be computationally challenging for real populations. Some efficiency can be gained by noting that $\text{cov}(\hat{p}_{ci}, \hat{p}_{cj}) = \text{cov}(\hat{p}_{ck}, \hat{p}_{cl})$ whenever $\mathbf{x}_i = \mathbf{x}_k$ and $\mathbf{x}_j = \mathbf{x}_l$.

25. In general, condition (2) need not hold exactly for the continuation-ratio model with $C > 2$. Sufficient conditions under which (2) holds exactly are discussed in Scholtus and Daalmans (2020). There it also argued that, for large samples, condition (2) should hold approximately in practice provided that the imputation model contains h, the variables that define the sampling design of S_2, and the interaction of h with these variables.

V. Simulation study

26. In this section, we describe a simulation study for an estimator based on mass imputation, comparing the analytical variance estimates from (4) and the bootstrap method from Section III. All computations were done in the R environment for statistical computing. The *survey* package (Lumley, 2018) was used for pseudo maximum likelihood estimation. A fast implementation of the analytical variance estimator was created using the *data.table* package (Dowle et al., 2019).

27. As a basis for this study, we used the data of the synthetic Samplonia population (see, e.g., Bethlehem, 2009). A target population of size $N = 5 \times 745 = 3725$ was created by concatenating five copies of all persons aged over 14 in Samplonia. In this simulation, there were no register data, so $U_1 = \emptyset$ and $U = U_2$. The sample $S = S_2$ was drawn according to a simple random sampling design without replacement, with sample size $n = N/5 = 745$.

28. Mass imputation of educational attainment for persons in $U_2 \backslash S_2$ was based on a simplified version of the approach proposed for the Dutch census. In this simulation study, *educational attainment* was classified into $C = 3$ categories, labelled as ‘low’, ‘medium’, and ‘high’. The continuation-ratio model was applied with auxiliary information of the form *gender* × (*age* + *income*). Here, *gender* consisted of two classes, *age* consisted of three levels, and *income* was used as a continuous variable. This model was based on the available variables in the Samplonia dataset; it should be noted that the imputation model proposed for the Dutch census uses different auxiliary variables (see Section VI).

29. The target frequency table in this study was *age* × *educational attainment* (both with three levels). Table 1 shows the true population counts (left panel) and approximate standard deviations of the estimated counts after mass imputation (right panel). The latter were obtained by drawing 20 000 samples from the population and for each of them estimating the model, applying mass imputation and tabulating.

Table 1. True counts and simulated true standard deviations for an artificial population.

	true counts			true standard deviations		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15–35)	330	795	400	34.5	42.2	36.8
middle (36–55)	115	560	480	22.3	36.8	36.1
old (56+)	120	525	400	22.8	35.6	34.5

30. We simulated 100 samples from the population and estimated the variances by two approaches:

- using the analytical variance estimator (4), with $\widehat{\text{cov}}(\hat{p}_{ci}, \hat{p}_{cj})$ from the appendix;
- using the bootstrap algorithm of Section III, with $A = 1$ and $B = 200$.

Table 2 shows the mean estimated standard deviations and (in brackets) their standard deviation across 100 simulations. It is seen that both approaches produced estimated standard deviations that were close to their true values on average. From the values in brackets it is seen that the estimated standard deviations from the analytical approach were more precise than those of the bootstrap approach.

Table 2. Mean and standard deviation of estimated standard deviations for estimated counts after mass imputation, based on 100 simulations.

	estimated analytical st. dev.			estimated bootstrap st. dev.		
	educational attainment			educational attainment		
age (years)	low	medium	high	low	medium	high
young (15–35)	34.1	41.8	36.6	34.1	41.9	36.4
	(1.3)	(0.7)	(1.2)	(2.2)	(2.3)	(2.0)
middle (36–55)	22.5	36.8	36.1	22.7	36.6	36.0
	(1.7)	(0.8)	(1.0)	(2.4)	(2.0)	(2.1)
old (56+)	22.8	35.4	34.3	22.5	35.2	34.5
	(1.8)	(0.8)	(1.0)	(1.9)	(2.1)	(2.2)

31. We found that condition (2) was approximately satisfied in these samples; this is as expected, since the imputation model included the variable *age* as a predictor and all observations had the same design weight. Across 100 simulated samples the average computation time was about 2.4 minutes per sample for the bootstrap method and just under 5 seconds per sample for the analytical method.

VI. Application to real data

32. In this application we estimate variances for one table for the Dutch Population and Housing Census 2011. The table under consideration contains the Dutch population by *Geographic area* (12 categories), *Sex* (2 categories), *Age* (8 categories) and *Educational attainment* (6 categories), which makes up 12 × 2 × 8 × 6 = 1,152 cells in total. *Geographic area*, *Sex* and *Age* are available from central population registers that fully cover the target population of the census.

33. In this evaluation study *Educational attainment* is estimated from the Educational Attainment File (EAF), with reference day January 1, 2011. The EAF is a database that includes data from multiple registers and LFS data from multiple years. Data from the EAF can be matched to the population registers at the micro level. The EAF is planned to be used for the 2021 census. The EAF contains a 'register part' and a 'non-register part', which include 9,363,909 and 7,291,890 persons, respectively. These two parts refer to the people for which register information on educational attainment is and is not available (U_1 and U_2 in the notation of Section 2). The focus of this simulation study is entirely on the non-register part. For 340,472 out of 7,291,890 persons educational levels are available from an LFS. The missing observations are imputed at the micro level, using the continuation-ratio model.

34. It should be noted that for this study we had only limited information about the origin of the sample data S_2. First, we did not have any information about the larger sample S from which S_2 was obtained by removing the overlap with U_1. Second, the data in S_2 are an integrated sample of several LFS rounds and we only had the final recalibrated sampling weights but no information about the underlying sampling design and the way different years were combined. For the purpose of this study, we approximated the design of S_2 by that of a simple random sample without replacement from U_2. This should give reasonable results for comparing the analytical and bootstrap methods, as the same approximation was used for both approaches. For a future application to the real Dutch census, a better approximation will be made which accounts for the complex survey design of the LFS.

35. The EAF-based data set was enriched with information of other administrative data sources. The variable *Income* (6 categories: 5 quantiles and unknown/not available) was used as a stratification variable in the imputation model; that is to say, a separate continuation-ratio model was estimated for each income class. *Income* has been chosen because it has a relatively strong association with *Educational attainment* (Daalmans, 2017). The variables *Age*, *Geographic area* and *Sex* that are contained in the target table were also used as an auxiliary variable for most of the cases (see 41 below).

36. As before, the bootstrap method in Section III was implemented with $B = 200$ and $A = 1$ and all computations were done using R and the *survey* and *data.table* packages. A comparison of the results after 200 and 190 iterations suggested that 200 iterations were enough for convergence of the estimated standard deviations. The computation time was much longer for the bootstrap method than for the

analytical method: about 21.5 hours versus 48.5 minutes. Note that the bootstrap computations could easily be parallelised across multiple processors to save time. However, we did not do this here.

37. We first give an impression about the results of the bootstrap procedure. The so-called coefficient of variation (CV) has been computed for each cell, i.e. the ratio of the estimated standard deviation to the mean, i.e. the average cell count over the 200 bootstrap samples. As we consider the non-register part of the EAF only, the mean is derived from the non-register part of the EAF. In Figure 1, the CV is plotted against the sample size of the cell. As expected, cells with the least number of sample survey observations have the highest CVs.

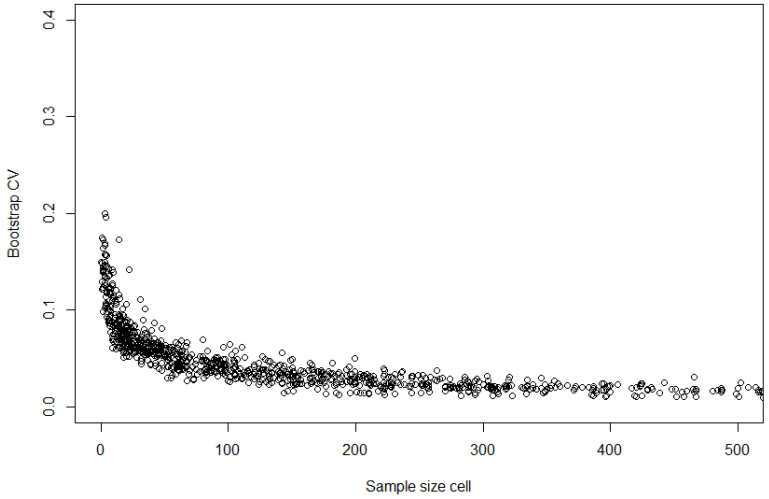

Figure 1. Coefficient of variation per cell for the bootstrap method.

38. We now turn to the difference in results for the standard deviations between the bootstrap method and the analytical approximation. First, we computed the ratio of the CV of the analytical method and the CV of the bootstrap method for each cell. Inspection of the distribution of these ratios shows that the difference between the analytical and bootstrap estimates is reasonably small. The median value of the ratio is 0.97, the 90th percentile is 1.79 and the 10th percentile is 0.59. This gives us some empirical evidence that the analytical procedure gives close approximations to the true variance.

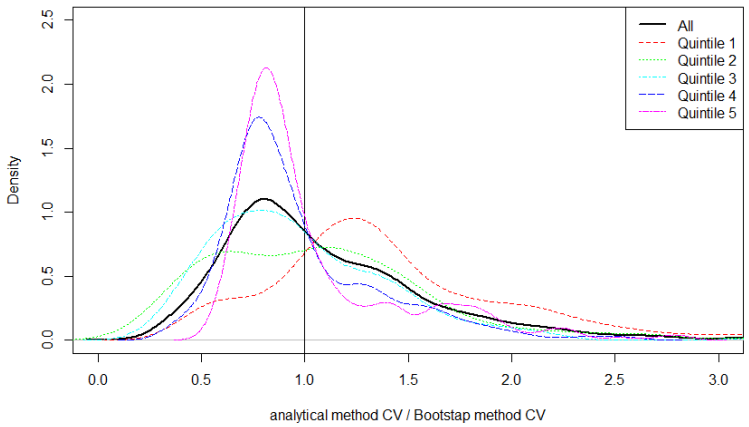

Figure 2. Ratio of coefficient of variation according to the analytical and bootstrap methods.

39. In Figure 2, density plots are shown of the ratio of the CVs of the analytical and bootstrap method. The solid black line represents the density of all cells in the table. The other lines represent the density for subsets of 20% of cells in the table, where the cells are ordered by sample size from smallest (quintile 1) to largest (quintile 5). As expected, the largest deviations occur for cells with relatively few

observations. Recall that the analytical variance estimates are based on a large-sample approximation. Interestingly, it appears that the analytical method tends to underestimate the CV slightly compared to the bootstrap method for all quintiles except the first one. A possible explanation for this underestimation is that condition (2) does not hold in this application, as the imputation model contains the main effects of *Sex*, *Geographic area* and *Age* but not their interactions. Thus, variance estimator (4) neglects a source of uncertainty. See Scholtus and Daalmans (2020) for a discussion of what happens when (2) does not hold.

40. To illustrate this further, we also computed the analytical CVs and bootstrap CVs for the table *Geographic area* × *Educational attainment*, i.e. one of the marginal tables of the original table. For this lower-dimensional table, condition (2) is expected to hold approximately in this application. In Figure 3, the ratio of the analytical and bootstrap CVs is plotted against the sample size per cell for the 72 cells in this table. It is seen that here the two approaches are in closer agreement and that the ratio of the CVs tends to 1 for cells with large sample sizes.

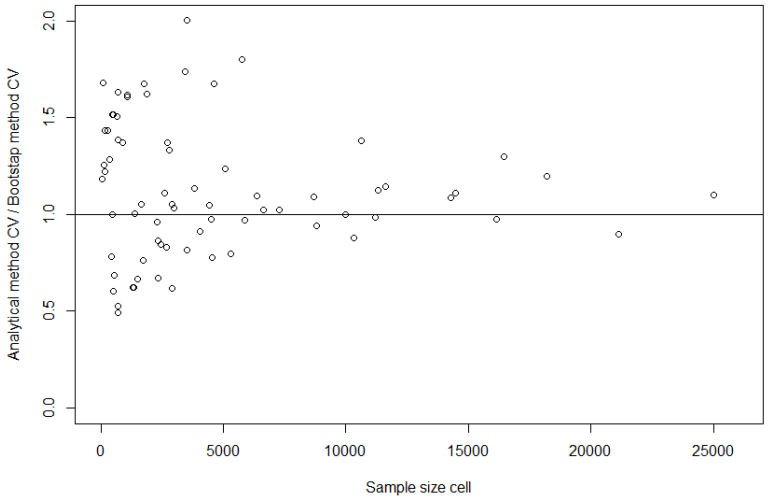

Figure 3. Ratio of coefficient of variation according to the analytical and bootstrap methods (marginal table: Geographic area × Educational attainment).

41. Finally, it should be noted that for some strata the number of observations was too small for a reliable estimation of all regression coefficients. Initially, *Sex*, *Geographic area* and *Age* were included in the model for each stratum. For some strata, this led to extremely high analytically derived standard deviations, when compared to the bootstrap. The problem was especially apparent for the highest income classes. In these strata low educational attainment barely occurs. As a consequence, the regression coefficients for estimating the probability of low educational levels could not be reliably estimated. Due to near-multicollinearity of the auxiliary variables, extremely high standard deviations were returned for some coefficients by formula (6) in the appendix. The impact of this diminished after reducing the regression model, by using only *Sex* as an auxiliary variable in certain strata. This shows that the analytical approach is sensitive to model selection. It also appears that the analytical variance estimates are more sensitive to model over-parametrisation than the mass-imputed estimates themselves.

VII. Discussion and conclusion

42. In this paper we have developed two approaches for estimating the design-based variance of frequency tables after mass imputation of one of the variables. This problem is relevant for frequency tables in the Dutch census that involve educational attainment. More generally, the problem is relevant for other applications in which a categorical variable is imputed for an entire target population. In Section II we presented an analytical variance expression that allows for a broad class of imputation methods. Each imputation method gives rise to a specific completion of the expression. The special case of the continuation-ratio model was considered in Section IV. In Section III, we described a bootstrap method that can be applied more generally to estimators based on combined data sources.

43. The analytical variance estimator was derived under the simplifying assumption that condition (2) holds, at least approximately. This condition can be verified in practice. In situations where this condition does not hold, our analytical approach is likely to underestimate the true variance if the sampling fraction is not sufficiently small (Scholtus and Daalmans, 2020). The bootstrap method does not require this assumption and may therefore be used regardless of whether condition (2) holds.

44. In general, the computation time can be a limiting factor for the estimation of variances for large data sets. The analytical approximation requires the computation of a double sum. The number of terms in this sum can become quite large. However, the bootstrap method can be expected to have an even larger computational burden for many applications. The bootstrap method requires work that is equivalent to imputing the entire target population many times. In our case study, approximately 7 million records needed to be imputed 200 times. This led to a computation time that was about a factor 30 larger than for the analytical approximation. This illustrates that, although the bootstrap is more flexible, analytical approximations may still be useful for variance estimation in practice.

45. An alternative variance estimation approach that was not considered in this study is multiple imputation (Rubin, 1987). Similar to our bootstrap method, this would require imputing the missing values of educational attainment throughout the population several times. A practical advantage of multiple imputation compared to direct bootstrapping would be that it requires fewer replicates. A limitation of multiple imputation, compared to the approaches considered here, is that it cannot be used to estimate the variance of an arbitrary estimator, but only that of the associated multiple imputation estimator. Thus, in this approach the variance estimation method actually guides the choice of the estimator itself, which then has to be based on a multiply-imputed file. For practical reasons there are currently no plans to generate more than one imputation per person in the official microdata of the Dutch virtual census, so multiple imputation is not an option for variance estimation in this application. For this reason it was not considered here. However, it may be interesting to compare multiple imputation to our analytical and bootstrap method in a more extensive future study.

46. Future work may also focus on extending the variance estimation methods considered here. The analytical approach could be extended to the case of imputing multiple categorical variables, or a combination of categorical and numerical variables. An extension of the analytical approach is also needed to handle imputation methods that are not based on a parametric model, such as hot deck imputation. Finally, it may be interesting to develop variance estimation techniques that can account for uncertainty in the measurement of register-based variables, including the effects of micro-integration (Bakker, 2011) when overlapping data are available from a register and a survey.

References

A. Agresti (2013), *Categorical Data Analysis* (Third Edition). New York: John Wiley and Sons.

B.F.M. Bakker (2011), Micro-integration: State of the Art. In: *ESSnet on Data Integration, Report on WP1*, pp. 77–107.

J. Bethlehem (2009), *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: John Wiley and Sons.

J.G. Booth, R.W. Butler, and P. Hall (1994), Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association* **89**, 1282–1289.

R.L. Chambers and C.J. Skinner (eds.) (2003), *Analysis of Survey Data*. Chicester: John Wiley and Sons.

G. Chauvet (2007), *Méthodes de Bootstrap en Population Finie*. PhD Thesis (in French), L'Université de Rennes.

J. Daalmans (2017), Mass Imputation for Census Estimation. Discussion Paper, The Hague: Statistics Netherlands.

J. Daalmans (2018), Divide-and-Conquer Solutions for Estimating Large Consistent Table Sets. *Statistical Journal of the IAOS* **34**, 223–233.

T. de Waal and J. Daalmans (2018), Mass Imputation for Census Estimation: Methodology. Report, The Hague: Statistics Netherlands.

M. Dowle et al. (2019), *data.table: Extension of data.frame*. R package version 1.12.0, available at http://cran.R-project.org/package=data.table.

B. Efron and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

S. Gross (1980), Median Estimation in Sample Surveys. In: *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 181–184.
L. Kuijvenhoven and S. Scholtus (2011), Bootstrapping Combined Estimators based on Register and Sample Survey Data. Discussion Paper, The Hague: Statistics Netherlands.
T. Lumley (2018), *survey: Analysis of Complex Survey Samples*. R package version 3.35, available at http://cran.R-project.org/package=survey.
Z. Mashreghi, D. Haziza, and C. Léger (2016), A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys* **10**, 1–52.
D.B. Rubin (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
S. Scholtus and J. Daalmans (2020), Variance Estimation after Mass Imputation based on Combined Administrative and Survey Data. Submitted for publication.
E. Schulte Nordholt, M. Hartgers, and R. Gircour (2004), *The Dutch Virtual Census of 2001. Analysis and Methodology*. Voorburg/Heerlen: Statistics Netherlands.
E. Schulte Nordholt, J. van Zeijl, and L. Hoeksma (2014), *Dutch Census 2011. Analysis and Methodology*. The Hague/Heerlen: Statistics Netherlands.

Appendix: Large-sample covariances for the continuation-ratio model

47. Details on pseudo maximum likelihood estimation of the continuation-ratio model from Section IV based on the sample S_2, including a derivation of the large-sample covariances given below, can be found in Scholtus and Daalmans (2020). Due to space restrictions, here we only list two main results.

48. First, for the conditional probabilities q_{ci} in (5) it can be shown that, asymptotically, $\mathrm{cov}(\hat{q}_{ci}, \hat{q}_{dj}) \approx 0$ for $c \neq d$, while $\mathrm{cov}(\hat{q}_{ci}, \hat{q}_{cj})$ can be estimated consistently by:

$$\widehat{\mathrm{cov}}(\hat{q}_{ci}, \hat{q}_{cj}) = \hat{q}_{ci}(1-\hat{q}_{ci})\hat{q}_{cj}(1-\hat{q}_{cj})\mathbf{x}_i^T \left(X^T \hat{\Delta}_{\geq c,w_2} X\right)^{-1} \hat{\Gamma}_c \left(X^T \hat{\Delta}_{\geq c,w_2} X\right)^{-1} \mathbf{x}_j. \quad (6)$$

Here, the matrix X contains rows $\mathbf{x}_i^T$ for each unit in the sample and $\hat{\Delta}_{\geq c,w_2}$ is a diagonal matrix with elements $(\hat{\Delta}_{\geq 1,w_2})_{ii} = w_{2i}\hat{q}_{1i}(1-\hat{q}_{1i})$ and, for $c > 1$,

$$(\hat{\Delta}_{\geq c,w_2})_{ii} = \begin{cases} w_{2i}\hat{q}_{ci}(1-\hat{q}_{ci}) & \text{if } y_{1i} = \cdots = y_{(c-1)i} = 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $w_{2i} = 1/\pi_{2i}$ is the design weight of unit $i \in S_2$ based on its inclusion probability π_{2i}. In addition, $\hat{\Gamma}_c$ is given by

$$\hat{\Gamma}_c = \sum_{i \in S_{2,\geq c}} \sum_{j \in S_{2,\geq c}} \frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_{2ij}} \mathbf{x}_i (y_{ci} - \hat{q}_{ci})(y_{cj} - \hat{q}_{cj}) \mathbf{x}_j^T,$$

where π_{2ij} is a second-order inclusion probability of S_2 and $S_{2,\geq c}$ denotes the subsample of S_2 consisting of all units with $y_{1i} = \cdots = y_{(c-1)i} = 0$. For more details, see also Chambers and Skinner (2003).

49. Next, define the short-hand notation $T_{c,ij} = \sum_{k=1}^{c} \sum_{l=1}^{c} \mathrm{cov}(\hat{p}_{ki}, \hat{p}_{lj})$ for $c = 1, \ldots, C-1$. For large samples, the covariances $\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ can be estimated by the following recursive algorithm:

(a) Estimate $\mathrm{cov}(\hat{q}_{ci}, \hat{q}_{cj})$ by (6) for all $c = 1, \ldots, C-1$ and define $\widehat{\mathrm{cov}}(\hat{p}_{1i}, \hat{p}_{1j}) = \hat{T}_{1,ij} = \widehat{\mathrm{cov}}(\hat{q}_{1i}, \hat{q}_{1j})$ in accordance with the definition $\hat{p}_{1i} = \hat{q}_{1i}$.
(b) Repeat the following steps for $c = 2, \ldots, C-1$:
- Estimate $\mathrm{cov}(\hat{p}_{ci}, \hat{p}_{cj})$ by:

$$\widehat{\mathrm{cov}}(\hat{p}_{ci}, \hat{p}_{cj}) = \{\widehat{\mathrm{cov}}(\hat{q}_{ci}, \hat{q}_{cj}) + \hat{q}_{ci}\hat{q}_{cj}\}\hat{T}_{c-1,ij} + \left(1 - \sum_{k=1}^{c-1} \hat{p}_{ki}\right)\left(1 - \sum_{l=1}^{c-1} \hat{p}_{lj}\right) \widehat{\mathrm{cov}}(\hat{q}_{ci}, \hat{q}_{cj}).$$

- Estimate $T_{c,ij}$ by:

$$\hat{T}_{c,ij} = \sum_{k=1}^{c} \widehat{\mathrm{cov}}(\hat{p}_{ki}, \hat{p}_{kj}) - \sum_{k=2}^{c} (\hat{q}_{ki} + \hat{q}_{kj})\hat{T}_{k-1,ij}.$$

(c) Finally, define $\widehat{\mathrm{cov}}(\hat{p}_{Ci}, \hat{p}_{Cj}) = \hat{T}_{C-1,ij}$ in accordance with the definition $\hat{p}_{Ci} = 1 - \sum_{c=1}^{C-1} \hat{p}_{ci}$.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors would like to thank Jeroen Pannekoek for his useful comments.