Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24-26 April 2017)

COMPUTATIONAL ESTIMATES OF DATA-EDITING RELATED VARIANCE

Mark van der Loo, Jeroen Pannekoek, and Lisanne Rijnveld
Statistics Netherlands

I. INTRODUCTION

1. It is widely acknowledged that statistical data editing (data cleaning) is an essential step in the production of accurate and reliable official statistics. Indeed, the impact on estimated values of data editing and imputation procedures has often been demonstrated. Whitridge and Bernier (2006) for example, show that imputation of missing values can have a significant effect on estimated totals of economic variables and van der Loo and Pannekoek (2014) show that estimated means and naively estimated variances may vary significantly over a multi-step data editing process. Dasu and Loh (2012) propose metrics to evaluate the impact of data cleaning based on distributional properties pre- and post data cleaning in the context of very large databases.

2. Besides the changes of estimated means, variances, and other distributional parameters in edited data, the precision of these estimates is influenced by data cleaning as well. The specific case of variance caused by imputation has received considerable attention over the last decades. The methods based on multiple imputation and Bayesian bootstrapping by Rubin (1987, 1996), the resampling based methods of Rao and Shao (1992); Rao (1996); Shao and Sitter (1996); Shao and Wang (2002, 2008) and the model-assisted analytical methods of Särndal (1992); Särndal and Lundström (2005) are amongst the most well-known. However, a statistical data editing process usually consists of a sequence of steps where estimation of missing values is just one of the steps performed. Other steps, such as outlier detection, rule-based corrections, or error localization may well contain stochastic sub-steps or parameter estimations based on observed data that influence the variance of estimation.

3. In this paper, the variance caused by data cleaning is therefore studied from a general point of view, abstracting from the specific type of data cleaning step performed. Here, the process of data editing is considered to be a part of the estimator. Consequently the estimator includes the stochastic editing steps that contribute to the variance. It is our purpose to estimate the total variance of this estimator. In the next section we classify variance-contributing editing steps and formally decompose the variance into a sampling and a data-editing related term. In Section III we propose a simple bootstrap estimator for the variance. The procedure is applied to a data editing process on SBS data in a numerical experiment that is described in Section IV. We discuss conclusions and ways to improve the method further in Section V.

II. EDIT STEPS AND SOURCES OF VARIANCE

4. We consider a finite population from which a survey sample is drawn with the purpose of estimating some specified population parameters. In official statistics these parameters are usually simple functions of the population data, most often (group) means, totals, or percentages. The estimator of such a parameter is a function of the survey dataset. When all variables are measured without errors and the data are complete (no missing values), the estimator is usually a simple explicit function of the data, often a weighted mean or total. In some cases explicit estimators for the sampling variance of the estimator can be derived. The standard example is the estimator of the mean μ under simple random sampling without replacement. The estimator $\hat{\mu}$ equals the sample mean $\bar{x}$ of variable X and the estimated sampling variance of $\hat{\mu}$ is given by the well-known expression

$$\frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2, \tag{1}$$

where N is the size of the population and n the size of the survey sample. The form of this variance estimator depends explicitly on the definition of the estimator for μ and the sample design. It must therefore be determined anew for any combination of estimator and design.

5. In practice, survey data sets often suffer from missing values, outliers, and erroneous values. Applying a simple estimator (sum, mean) to the raw dataset is therefore out of the question. In stead, an elaborate data editing process first transforms the raw data set into a completed and consistent data set and only then are the estimates computed. This means that the whole data editing process is in principle part of the formulation of the estimator.

6. For variance estimation the current practice is to largely ignore this complication and to derive variance estimators that treat the edited dataset as if it was the observed data. This approach is valid only when the data editing procedure does not add to the sampling variance of an estimator—that is, when it consists of simple deterministic transformations transforming raw data into complete and consistent data. If this is not the case it is an approximation that should be validated.

A. Sources of variance

7. In most surveys for official statistics a number of data editing steps are performed that are not of the simple deterministic kind and will in different ways influence the variance. We distinguish three different types of edit steps according to their influence on estimator variance.

Case 1: *Deterministic editing with fixed parameters*
These are editing operations that can be specified by deterministic rules or algorithms which may be parameterized by auxiliary data or control parameters that are not estimated from the data set to be edited. The outcomes of these operations do therefore not depend on the composition of a particular sample and don't add to the variance of estimation. Examples of deterministic editing operations include imputation with a historical value, deletion of incorrect minus-signs, and correction of a unit-of-measure error when the turnover by employee ratio is 300 times that of the previous year.

Case 2: *Deterministic editing with estimated parameters.*
These include edit operations depending on parameters that are estimated from the data set to be edited. This not only includes many common imputation methods (parametric models, hot deck) but also, for example, outlier detection using estimated cut-off values. A more subtle but similar case is when the reliability weights for error localization depend on the dataset to be edited. For example, one may derive them from the difference between observed and

predicted values. Editing steps with estimated parameters create cross-record dependencies. They therefore cause the estimator to vary depending on sample composition and contribute to the overall sampling variance.

Case 3: *Stochastic editing methods.*
Some editing or imputation operations involve methods that create modified values in an explicitly random manner. Well-known examples include random hot deck and k nearest neighbour imputation. However, error localization or certain roundoff correction methods can also include randomization, for example to choose from a set of equivalent solutions. Since these methods introduce new random processes they obviously contribute to the estimator variance.

The above cases do not classify data editing methods into mutually exclusive categories. Some methods combine elements of Case 2 and 3, such as imputation models that augment predicted values with randomized (normal) residuals.

B. Formal decomposition of estimation variance under data editing

8. Consider a general population parameter θ, estimated from a survey sample using a procedure which we label $\hat{\theta}$. We assume that this estimator includes some kind of data editing process. If the result of data editing depends on the realization of a stochastic variable δ, the total variance of $\hat{\theta}$ can be decomposed as follows (Eve's law[1]):

$$V(\hat{\theta}) = \underbrace{E_\delta V(\hat{\theta}|\delta)}_{\substack{\text{sampling}\\ \text{variance}}} + \underbrace{V_\delta E(\hat{\theta}|\delta)}_{\substack{\text{data editing}\\ \text{variance}}} . \tag{2}$$

The first term is interpreted as the sampling variance. If δ is fixed the second term vanishes as well as the expectation in the first term, and we have $V(\hat{\theta}) = V(\hat{\theta}|\delta)$. The second term is interpreted as the variance contribution added by data editing. Ignoring the effect of data editing on the variance of an estimator is equivalent to assuming that $V_\delta E(\hat{\theta}|\delta) \ll E_\delta V(\hat{\theta}|\delta)$.

9. As an example consider the outlier detection example of Case 2 mentioned above. Let us assume that a value is deemend erroneous when it exceeds some limit $\hat{\alpha}$ which is computed from the data set to be edited (e.g. the median plus 3/2 times the interquartile range). The value of $\hat{\theta}$ depends on $\hat{\alpha}$ so

$$V(\hat{\theta}) = E_{\hat{\alpha}} V(\hat{\theta}|\hat{\alpha}) + V_{\hat{\alpha}} E(\hat{\theta}|\hat{\alpha}), \tag{3}$$

where the second term can be interpreted as the 'outlier detection variance'.

10. In a multi-step data editing procedure there may be multiple sources of randomness. In that case the total variance can be formally expanded by iterated application of Eve's law. As an example consider the case where imputations are generated as a prediction from a statistical model augmented with random disturbances. This is an example that mixes Case 2 and Case 3 mentioned earlier since an imputation is determined as

$$M(x, \hat{\beta}) + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \hat{\sigma}^2).$$

Here, $M(x, \alpha)$ denotes the model prediction at a value of predictor x, $\hat{\alpha}$ its parameters and $\hat{\sigma}^2$ variance of the model residuals. The parameters $\hat{\alpha}$ and $\hat{\sigma}$ have an associated sampling variance while ϵ is drawn from a distribution of which the parameters depend on the sample. The formal expansion of the variance conditional on $\hat{\sigma}$ can be written as

$$V(\hat{\theta}|\hat{\sigma}) = E_{\hat{\beta}} V(\hat{\theta}|\hat{\beta}, \hat{\sigma}) + V_{\hat{\beta}} E(\hat{\theta}|\hat{\beta}, \hat{\sigma}).$$

[1] This rule is sometimes called Eve's law because of the way the symbols E and V occur in the equation.

Inserting this expression into Eve's law for $V(\hat{\theta})$ we get

$$V(\hat{\theta}) = \underbrace{E_{\hat{\sigma}} E_{\hat{\beta}} V(\hat{\theta}|\hat{\beta}, \hat{\sigma}^2)}_{\text{sampling variance}} + \underbrace{E_{\hat{\sigma}} V_{\hat{\beta}} E(\hat{\theta}|\hat{\beta}, \hat{\sigma}^2)}_{\text{estimation variance}} + \underbrace{V_{\hat{\sigma}} E(\hat{\theta}|\hat{\sigma})}_{\substack{\text{method} \\ \text{variance}}}. \tag{4}$$

Again, the first term can be interpreted as the variance contribution caused by variation of the sample composition. The second term describes the variance caused by the sampling variation of the estimated parameters $\hat{\beta}$. The third term describes the variance caused by the added random term ϵ. Note that $V_{\hat{\sigma}}$ depends on both the variation of $\hat{\sigma}$ over the samples as well as on the variation of ϵ over the normal distribution.

11. In general, the imputation or correction methods as well as the mechanisms causing randomness are hard to analyze analytically. Deriving analytical expressions that estimate Equations such as (3) or (4) therefore becomes quickly intractable in practice. A computational approach to total variance estimation therefore becomes attractive.

III. A BOOTSTRAP ESTIMATOR OF THE TOTAL VARIANCE

12. As an alternative to deriving analytic expressions we propose to use a simple nonparametric bootstrapping scheme to evaluate the variance of an estimator that includes data editing. In this procedure, the raw data is resampled (with replacement) B times and the full estimation procedure is executed for each bootstrap sample. This then yields an ensemble of estimators $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B\}$ The bootstrap estimate $\hat{\theta}^*$ of θ is then given by the ensemble average

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b.$$

An estimate of the total variance $V(\hat{\theta}^*)$ is given by the bootstrap estimator

$$\hat{V}^*(\hat{\theta}^*) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}^*)^2.$$

Since the complete estimation procedure is executed for each bootstrap sample this estimator includes both terms on the right hand side of Equation (2), so

$$\hat{V}^*(\hat{\theta}^*) = \hat{E}^*_\delta \hat{V}(\hat{\theta}^*|\delta) + \hat{V}^*_\delta \hat{E}(\hat{\theta}^*|\delta).$$

13. Comparing the single estimate $\hat{\theta}$ with the bootstrap mean yields an estimate for the bias of $\hat{\theta}$:

$$\hat{\text{bias}}(\hat{\theta}) = \hat{\theta} - \hat{\theta}^*. \tag{5}$$

If we also have an estimator for the variance of $\hat{\theta}$ that ignores the data editing, say $\hat{V}'(\hat{\theta})$ then we can estimate the variance contributed by data editing as

$$\hat{V}_\delta(\hat{\theta}) \equiv \hat{V}_\delta \hat{E}(\hat{\theta}|\delta) = \hat{V}^*(\hat{\theta}^*) - \hat{V}'(\hat{\theta}). \tag{6}$$

Here, we introduced the short-hand notation $\hat{V}_\delta(\hat{\theta})$ to represent the data editing variance.

14. As an example, consider the estimation of the mean μ of a variable X. The estimator $\hat{\mu}$ is given by $\bar{x}_e$: the sample mean after data editing. The naive estimator of variance (in the case of simple random sampling with replacement) is then given by Equation (1)

$$\hat{V}'(\hat{\mu}) = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{j=1}^{n}(x_{e,j} - \bar{x}_e)^2.$$

The bootstrap variance is given by

$$\hat{V}^*(\hat{\mu}^*) = \frac{1}{B}\sum_{b=1}^{B}(\bar{x}_{e,b} - \hat{\mu}^*)^2, \tag{7}$$

where $\hat{\mu}^*$ is the mean over the bootstrap ensemble $\bar{x}_{e,b}$, $b = 1, 2, \ldots, B$ (each bootstrap sample edited individually). The amount of data editing variance ignored by the naive estimator is thus estimated as

$$\hat{V}_\delta(\hat{\mu}) = \hat{V}^*(\hat{\mu}^*) - \hat{V}'(\hat{\mu}).$$

15. For many parameter estimates, the naive variance estimator ignoring δ is readily available. However, even in cases where variance estimators are not analytically known one can create a naive variance estimator by resampling from the edited dataset, so

$$\hat{V}'^*(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}'_{e,b} - \hat{\theta})^2.$$

The prime on the notation $\hat{\theta}'_{e,b}$ indicates that the ensemble of estimates is created by resampling from the data set that has been edited only once. The estimated data editing variance is now given by $\hat{V}^*(\theta) - \hat{V}'^*(\hat{\theta})$.

16. In practice the bootstrap procedure pointed out above may be simplified when a multi-step data editing process starts with one or more Case 1 data editing operations (simple deterministic steps that do not add to the variance). Such steps need not be part of the resampled estimation, so one simply performs resampling on the data set resulting from these initial steps. Complications occur when survey designs are more complex than simple random sampling without replacement. It is well known that for more elaborate schemes such as clustered or stratified sampling, bootstrapping schemes must be adapted accordingly. See Rao and Wu (1988); Rao (2006) for bootstrap approaches under complex survey designs.

IV. APPLICATION

17. To test the bootstrap estimator, we performed a numerical experiment using 304 records from the Structural Business Survey (SBS) on wholesalers. A set of 66 linear restrictions were imposed on 69 variables: 13 equality balance rules and 53 nonnegativity constraints. The raw dataset has about 38% of the cells missing and about one third of the records violated at least one rule (this is counting only the cases where rules could be checked). Using the procedure outlined below, the data set can be made complete and consistent with the restrictions in a fully automated way.

A. Data editing procedure

18. Figure 1 gives an overview of the automated data editing procedure applied in our numerical experiment, see e.g. Pannekoek et al. (2013) for more details. The first two steps consists of a

deterministic method for removing unit of measure errors (caused by reporting in unit currency rather than thousands of currency), followed by a deterministic method for finding and correcting typing errors in numerical data under linear constraints. These methods are fully deterministic and have no components that vary across sample compositions. They therefore do not add to the data editing variance and are left out of the bootstrap procedure.

19. The procedure for correcting rounding errors is based on the scapegoat algorithm of Scholtus (2011). This algorithm corrects violations of linear (in)equality restrictions on the order of a rounding error (e.g. 2 units of measurement) by randomly assigning a variable to adapt (it takes account of the restrictions in choosing what variables may be adapted). It therefore adds to the data editing variance through an intrinsic stochastic element.

20. The Error localization step is based on the generalized principle of Fellegi and Holt (1976). This means that for each record a set of variables is selected that minimizes a sum of reliability weights under the restriction that they can be imputed consistent with the edit rules. For each record, initial weights w'_j $(j = 1, 2, \ldots, n = 67)$ were computed as

$$w'_j = \min\left\{\frac{x_j}{\text{median}(X_j)}, \frac{\text{median}(x_j)}{X_j}\right\}.$$

These weights are lower (indicating a lower reliability) when a value deviates further from the variable median. Since the median varies over samples, this introduces sampling variance to the error localization procedure. The weights were subsequently scaled so

$$w_j = 1 + \frac{w'_j - \min_j\{w'_j\}}{\max_j\{w'_j\} - \min_j\{w'_j\}} \times \frac{1}{n}.$$

This scaling ensures that the feasible solution of least weight is amongst the feasible solutions that point out the least number of variables necessary for error correction. See van der Loo (2015) for a proof.

21. Once erroneous fields have been deleted, it is possible that the restrictions allow a unique value for one or more of the variables. If this is the case, such values are derived in a process called deductive imputation. Methods for deductive imputation under linear (in)equality constraints are described by de Waal et al. (2011, Chapter 9). These methods are fully deterministic and do not add to the data editing variance.

22. Next, fields that are left empty are imputed using ratio imputation where the 'total number of staff' was used as predicting variable in all cases. For cases where the predictor is missing, the overall median is imputed. The computation of the estimated ratios and medians clearly contributes to the sampling variance. The rations explain a varying amount of variance in the imputed variables. The explained variance varies from near zero (e.g. for amounts invested) to more than 80-90% (specific staff numbers, staff costs). For this work we focused on a number of overall financial variables (See III). The variance explained by staff number in these variables is about 20-25%, except for the total operating expenses for which it is about 10%.

23. In the final step, the values that were imputed using a model-based approach are adjusted such that each record satisfies the set of restrictions. The adjustment method is based on the successive projection algorithm described by Pannekoek and Zhang (2015). This method projects an inconsistent record onto the convex polytope described by the set of linear inequality restrictions by adjusting only the imputed values. This method is deterministic and in itself does not add to the data editing (sampling) variance.

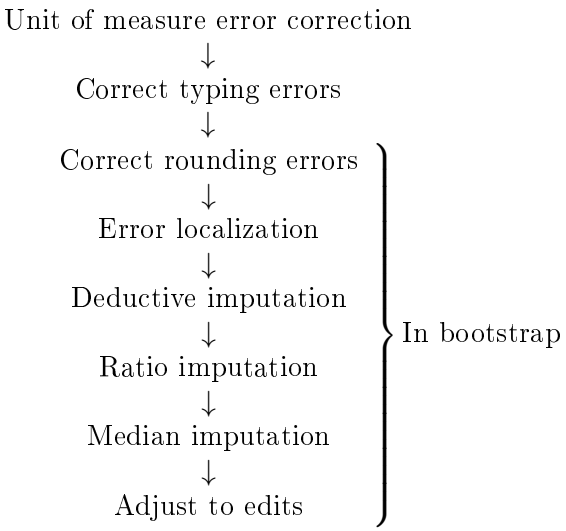

FIGURE 1. Overview of the data editing process.

B. Bootstrap procedure

24. A non-parametric bootstrap was used to replicate the relevant parts of the data editing procedure $B = 800$ times (outlined in Figure 1). For the current experiment we focused on six variables: 'total net turnover', 'other income', 'total income', 'operational expenses', 'operational income' and 'tax result'. These are the main variables in the overall account balance for an establishment.

25. The numerical experiment was executed using the R environment for statistical computing (R Development Core Team, 2011). The R packages `validate`, `errorlocate`, `deductive`, `simputation`, and `rspa` were used to perform rule management, error localisation, deductive correction, imputation and adjustment. (van der Loo and de Jonge, 2016, 2017; de Jonge and van der Loo, 2016; van der Loo, 2015, 2017).

26. For the interpretation of bootstrap bias (Equation) or the data editing variance, it is important to understand to what extent the bootstrapping procedure has converged. Figure 2 shows how the bootstrap means of the main variables converge as a function of the size of the bootstrap ensemble. We find that after about 600 bootstrap samples all means have converged to within 1% of their final value. In figure 3 the convergence of the standard deviations is shown. The plot was generated by bootstrapping from the edited original dataset and so does not include the data editing. From this test we find that it takes about 12500 bootstrap samples before all standard deviations have converged to within 1% of their final value. At $B = 800$, the number of resamples used in this work, the standard deviations have converged to within about 3%-4% of their final values. These means that from the current numerical experiment, a bias of less than 1% and an observed data editing variance of less than 4% can be a consequence of non-convergence and should as such not be interpreted as significant. Finally, it is important to observe that may both over- or underestimate parameters before a satisfactory convergence level is reached.

C. Results

27. Table 1 presents the main results of our bootstrap experiment. To make comparison with the means easier, Table 1 reports the square roots of the bootstrap variance estimation $\hat{\sigma}(\hat{\mu}^*) = \hat{V}^*(\hat{\mu}^*)^{1/2}$

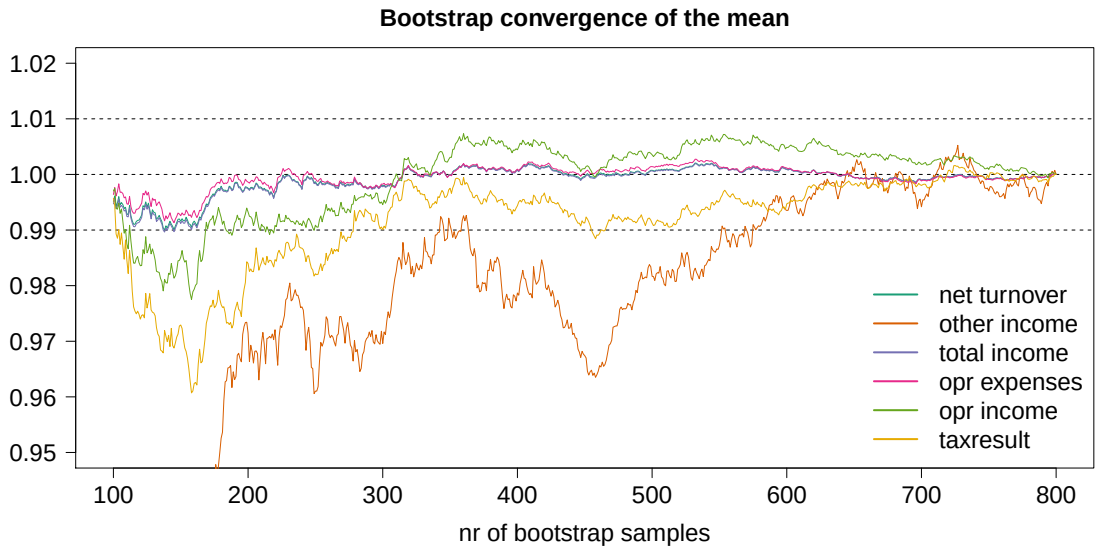

FIGURE 2. Convergence of the bootstrap means as a function of the number of bootstrap samples. Convergence is computed as the bootstrap mean at the number of bootstraps samples on the x-axis, relative to the final estimate at $B = 800$.

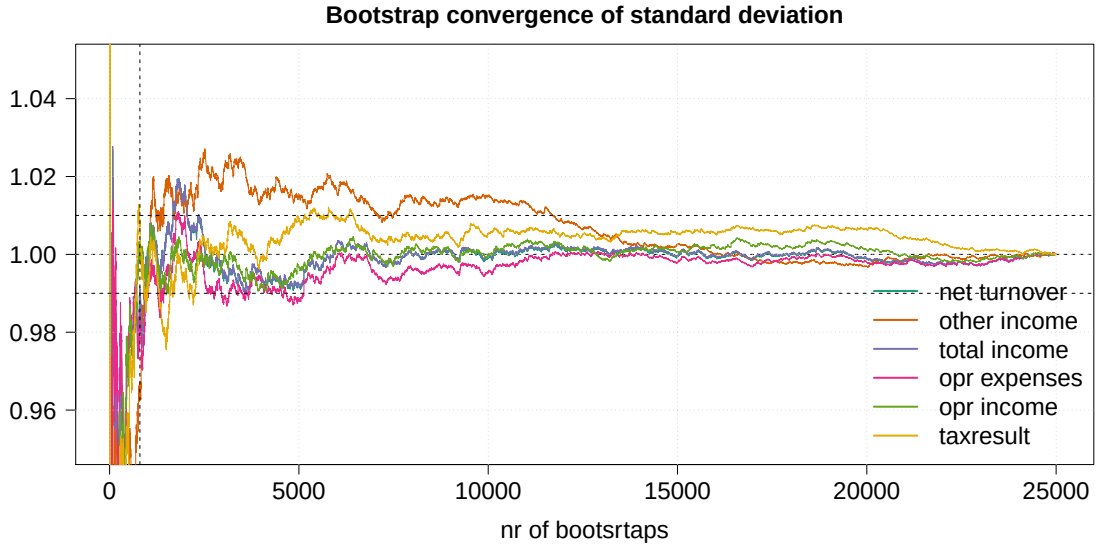

FIGURE 3. Convergence of the bootstrap standard deviations as a function of the number of bootstrap samples. Bootstrapping was perfomed on the original dataset after editing.

and similar for the naive variance $\hat{V}'(\hat{\mu})$. The final column represents the percentage of variance added by data editing. This is computed as $[\hat{V}^*(\hat{\mu}^*) - \hat{V}'(\hat{\mu})]/\hat{V}^*(\hat{\mu}^*) \times 100$, where the variance terms were defined in Equations (6) and (7).

28. For the main variables: 'net turnover', 'total income' and 'total operating expenses', the bootstrap estimated bias is less then a tenth of a percent. This does not exceed the fluctuations we would still expected at repeated $B = 800$ bootstrapping experiments so we do not consider this bias

TABLE 1. Bootstrap mean $\hat{\mu}^*$, standard mean $\hat{\mu}$, bootstrap estimated bias, bootstrap standard deviation $\hat{\sigma}^*(\hat{\mu}^*)$, naive standard deviation $\hat{\sigma}'(\hat{\mu})$ and percentage bootstrap variance for six SBS variables. Percentages in italics are considered insignificant when compared to the state of convergence after $B = 800$ bootstrap samples.

Variable	$\hat{\mu}^*$	$\hat{\mu}$	$\hat{\text{bias}}(\hat{\mu})$	$\hat{\text{bias}}(\hat{\mu})(\%)$	$\hat{\sigma}^*(\hat{\mu^*})$	$\hat{\sigma}'(\mu)$	$\hat{V}_\delta(\hat{\mu})$ (%)
net turnover	27181.0	27171.0	9.7	*0.03*	2371.5	2364.3	*0.6*
other income	130.8	127.1	3.7	2.83	75.1	75.4	*-0.7*
total income	27328.0	27314.0	14.0	*0.05*	2388.8	2378.1	*0.9*
opr expenses	21938.0	21958.0	-19.8	*-0.09*	2058.4	2096.7	*-3.8*
opr income	758.4	745.8	12.6	1.66	144.7	134.2	13.9
taxresult	725.8	708.7	17.1	2.35	176.2	162.5	14.9

significant. For 'other income', 'operational expenses' and 'operational income' we do find a significant bootstrap bias of 2.83, 1.66 adnd 2.35% respectively. These variables are sub-items on the balance sheet and have in the raw data a higher percentage of missings. Compared to the other variables, 'other income' is the variable with the largest relative effect of editing on the mean (-19% compared to the raw data) the largest percentage of missings in the raw data (8.6% with the others less than 3%). It is also the only variable with a mixed distribution: about 50% of the 'other turnover' values are reported zero.

29. Regarding the data editing variance, we find that the variables 'operating income' and 'tax result' gain a detectable amount of data editing variance: respectively 13.9% and 14.9%. Percentages found with the other variables do not exceed the amount of fluctuation in variance if we were to repeat the $B = 800$ bootstrap experiment. Inspection of the data shows that of the six studied variables, these variables had the highest number of model-based (ratio or median) imputations. This is likely to explain the extra variance, especially given how crude the imputation methods in our experiment are.

V. CONCLUSION

30. We argue that besides the well-known imputation variance, other data editing steps can contribute to the overall variance of a population estimate based on a survey sample. The contribution of data editing variance can be studied by decomposing the total variance into components that are conditional on either sample composition or randomisation during an edit step. Analytical expressions for such (iterated) decompositions are hard to obtain but the total variance may be estimated using a bootstrap estimator. In this work we have applied a simple nonparametric bootstrap to estimate the total variance of a simple but realistic automated data editing procedure. By comparing this variance with a naive variance estimator that does not take data editing into account, an estimate of total data editing variance is derived for the first time.

31. In the current method we have made some simplifying assumptions which could be improved upon. In particular, we expect further improvements by taking better into account the sampling scheme applied to obtain the original data, and by using bootstrap methodology that is suited for samples that form a significant fraction of the population.

References

Dasu, T. and J. M. Loh (2012). Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment 5*(11), 1674–1683.

de Jonge, E. and M. van der Loo (2016). *errorlocate: Locate Errors with Validation Rules*. R package version 0.1.2.

de Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Wiley, Inc.

Fellegi, I. P. and D. Holt (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association 71*(353), 17–35.

Pannekoek, J., S. Scholtus, and M. Van der Loo (2013). Automated and manual data editing: A view om process desiogn and methodology. *Journal of Official Statistics 29*(4), 511–537.

Pannekoek, J. and L.-C. Zhang (2015). Optimal adjustments for inconsistency in imputed data. *Survey methodology 41*, 127–144.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, J. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association 91*(434), 499–506.

Rao, J. (2006). Bootstrap methods for analyzing complex sample survey data. In *Proceedings of Statistics Canada International Symposium Series. Symposium*.

Rao, J. N. and J. Shao (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika 79*(4), 811–822.

Rao, J. N. and C. Wu (1988). Resampling inference with complex survey data. *Journal of the american statistical association 83*(401), 231–241.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association 91*(434), 473–489.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology 18*, 241–252.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Scholtus, S. (2011). Algorithms for correcting sign errors and rounding errors in business survey data. *Journal of Official Statistics 27*, 467–490.

Shao, J. and R. R. Sitter (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association 91*(435), 1278–1288.

Shao, J. and H. Wang (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association 97*(458), 544–552.

Shao, J. and H. Wang (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Statistica Sinica 18*, 281–297.

van der Loo, M. (2015). Experiences with r based data editing systems. *Romanian Statistical Review 2/2015*, 141–152.

van der Loo, M. (2015). *rspa: Adapt Numerical Records to Fit (in)Equality Restrictions*. R package version 0.1.8.

van der Loo, M. (2017). *simputation: Simple Imputation*. R package version 0.2.2.

van der Loo, M. and E. de Jonge (2016). *validate: Data Validation Infrastructure*. R package v. 0.1.5.

van der Loo, M. and E. de Jonge (2017). *deductive: Data Correction and Imputation Using Deductive Methods*. R package version 0.1.2.

van der Loo, M. and J. Pannekoek (2014, September). Towards generic evaluation of data validation functions. In *Work session on statistical data editing (Paris)*. UNECE.

Whitridge, P. and J. Bernier (2006). The impact of editing on data quality. In *Work session on statistical data editing*. UNECE.