Evaluating estimation quality after automatic statistical data editing processes

Lisanne Rijnveld

Statistic Netherlands, Division of Methodology and Quality

Henri Faasdreef 312, 2492 JP The Hague, The Netherlands

Email l.rijnveld@gmail.com.com

May 24, 2016

Abstract

Automatic statistical data editing (ASDE) is a process consisting of a combination of edit steps that correct errors in the data in order to obtain more accurate estimates. The classic estimation method of the reliability of an estimate does not take the effects of ASDE into account. The classic method to estimate the standard error only takes uncertainty due to sampling variance into account. However, because of stochastic components of some of the edit steps, ASDE is likely to add to the reliability of an estimate as well. The aim of this research was to provide an adequate measure for the reliability for analyses involving ASDE. Results from the simulation study show that measuring the standard error with a bootstrap approach provides a better estimation of the estimates reliability compared to classical estimation, because it takes the uncertainty due to the stochastic components into account. In addition, by investigating the standard error over the different edit steps, the bootstrap approach can be used to obtain information about the effects and quality of the edit steps.

Keywords: data quality, deductive correction, ratio imputation, resampling

1 Introduction

One of the main task of Statistics Netherlands is to provide reliable nation-wide figures on economic performance and financial structure of businesses. To a large extend, it relies on survey data and statistical methodology to generate such estimates. In surveys used for this purpose, information on financial variables like turnover and costs of a company is asked. Data on these variables for one company form a record; multiple records together form the data set from which statistical inference can be drawn. In practice, records often contain errors due to mistakes made filling in the survey (de Waal et al., 2011a). Reasons for these mistakes are among others, misinterpretation of questions, and the occurrence of typing errors.

In standard survey methodology, sample data is assumed to be correct and consistent to begin with, and based on those assumptions one can derive estimators for population parameters, in particular the sample mean and its standard error. Analyses based on data containing errors like missing or incorrect values will lead to undefined statistics and biased results. Therefore, the data is checked for errors prior to statistical analyses. This process of detecting and correcting errors is called *statistical data editing* (de Waal et al., 2011a). In this process, detection of errors is based on predefined rules the data must meet, which are called edit rules (Van der Loo and de Jonge, 2012). An example of an edit rule is

$$P = T - C \tag{1}$$

where P is the profit of a company, T its turnover, and C its costs. This means that a company's turnover and costs must sum up to its profit. Another example of an edit rule is that $T > 0$, meaning that the turnover should be positive. When a record violates one or more edit rules, the record is considered to contain errors and is called inconsistent. After erroneous values are localized, they are corrected based on statistical modification methods in such a way that the resulting record is a better approximation of the true data. The goal of statistical data editing is

to correct inconsistencies while leaving the reported data intact as much as possible (Van der Loo and de Jonge, 2012). Traditionally, data editing staff with subject-specific knowledge edits the data manually. Nowadays, it is possible to perform at least part of it automatically. Automatic statistical data editing (ASDE) is to be preferred over manual statistical data editing because of its reproducibility and time and cost saving features.

ASDE is a process consisting of a combination of edit steps with the aim to deliver an error free data set (Pannekoek et al., 2013). The underlying process structure of the edit steps is in general comparable and consists of three generic editing functions: verification, localization and imputation. First, the occurrence of an error is detected by checking the data for inconsistencies with respect to the edit rules (verification). Inconsistency of the data values with the edit rules means that there is an error. But if a violated edit rule involves several variables, as in edit rule (1), it is not immediately clear which of the variables are incorrect. Therefore, incorrect values in an inconsistent record have to be localized. This is called *error localization*. Finally, the localized erroneous fields are corrected, this means that the values are replaced with other values to meet the edit rules. Replacement with new values is called *imputation*.

The edit steps differ from one another in the type of error they correct and the method they use to make these corrections. Two types of correction methods can be distinguished. Some edit steps correct errors in records always in the same way and independent of the other records in the data set. These corrections are called *deterministic corrections*. Edit step *typos*, that correct typing errors, is an example of an edit step that uses this deterministic method. Table 1 gives an example of a record that contains a typo (Hoogland et al., 2010). The first column shows that the record is inconsistent with respect to edit rule (1). This inconsistency can be solved by correcting one of the three variables. These three possibilities are shown in the last three columns, where the bold faced values are changed. Intuitively, correction two seems to be most appealing, because changing 283 into 238 is less intrusive

Table 1: Example of correction of an typo

		Correction		
	Record	1	2	3
Turnover	353	**398**	353	353
Costs	283	283	**238**	283
Profit	115	115	115	**70**

than correction one and three. It is more likely that the real value 238 is by mistake changed in 283, than 398 in 353 or 70 in 115. Edit step *typos* is configured such that, if a record does not meet edit rule (1), but it does when the numbers in one of the involved variables are interchanged, it changes the value this way (Scholtus, 2009).

Other edit steps in the ASDE process make use of a so called *stochastic method* in order to correct errors. This means that the resulting corrected value will not always be the same, but depends on a random mechanism. An example of an edit step that uses a stochastic method is *ratio imputation*. It is configured in a way that when an edit rule is violated, but no specific incorrect value can be localized, it randomly selects one of the involved variables and imputes that variable by multiplication of a record's auxiliary variable with the ratio:

$$mean(TargetVariable)/mean(AuxiliaryVariable)$$

This implies that edit step *ratio imputation* is also dependent on other records in the data, which forms a second stochastic component. Paragraph 3.3 will elaborate on different types of edit steps in more detail.

The stochastic components of some edit steps in ASDE should be taken into account when drawing inferences on the data, because they can influence the estimator, and are thereby adding to its uncertainty. The level of uncertainty gives a measure for the reliability of the estimator. Classical estimation takes into account that there is uncertainty due to possible variation of the estimates over different samples from the population, which is called *sampling variance*. Classical estimation does not include the effect of ASDE on the uncertainty of the estimate.

Uncertainty of an estimate is usually expressed as the *standard error*. The classical method to estimate the standard error is based on the sample variance. This a valid method to use in order to estimate sampling variance, but it might not be for analyses involving ASDE since it ignores the stochastic components of ASDE. Much research has gone into deriving estimators and analytical expressions for their variances for specific cases where classical estimation is inadequate. Examples are, among others, *small area estimation* (Rao, 2003), *weighting methods for non-response correction* (Bethlehem, 2009) and *calibrated estimation* (Kim and Park, 2010). However, estimation involving ASDE has not been subjected to similar research and no results are available on variance estimation for the specific case of ASDE.

Statistics Netherland aims to publish reliable statistics. This statistical information is often used for decisions made by, among others, the government. Therefore, it is important to have a valid way to estimate reliability. Eventually, an estimate is not better than its uncertainty, which makes it important to have an adequate reliability measure.

This paper focuses on the influence of ASDE on the reliability of an estimate. The aim is to provide a valid estimation method for the reliability when ASDE is involved. Chapter 2 gives an elaboration on the theoretical background. Chapter 3 describes the simulation study that is performed in order to answer the research question. Chapter 4 presents an overview of results from the simulation study. In chapter 5 findings will be discussed.

2 Theoretical background

Consider a simple random sample with the purpose of estimating the population mean of variable Y. The classic estimates for the population mean, population variance and standard error of the estimated mean are given by the following formula

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3}$$

$$SE_{\bar{y}} = \sqrt{\frac{s_y^2}{n}} \tag{4}$$

where $\bar{y}$ and s_y^2 are the sample mean and variance respectively, y_i the value on variable y in record i, n the number of records in the sample and $SE_{\bar{y}}$ the standard error of the estimated mean.

The underlying assumption for using $SE_{\bar{y}}$ as a measure of reliability is that the estimate $\bar{y}$ is obtained by the following process

$$\begin{gathered}[\text{simpel random sample without replacement}] \rightarrow \\ [\text{data collection}] \rightarrow [\text{table of data } y_i^{(0)}] \rightarrow [\text{estimate } \bar{y}^{(0)}]\end{gathered}$$

The only component here that contributes to the variance of the estimate is random sampling. In this case, the reliability of an estimate can be adequately estimated by classical estimation which takes sampling variance into account. But considering the poor quality of raw survey data, reality is closer described by the following process

$$\begin{gathered}[\text{simpel random sample without replacement}] \rightarrow \\ [\text{data collection}] \rightarrow [\text{ASDE}] \rightarrow [\text{table of data } y_i^{(1)}] \rightarrow [\text{estimate } \bar{y}^{(1)}]\end{gathered}$$

Another way of stating this is that ASDE is commonly interpreted as part of the data collection stage. However, since the discussed stochastic component and dependency on the survey outcome of editing steps, this interpretation is not justified. ASDE contributes to the estimator's variance and should be considered part of the estimate for the reliability of $\bar{y}$.

Summarizing, the situation can be described as follows. The actual standard error of the estimated statistic based on a procedure including ASDE can be regarded as built up out of two components

$$SE_{total} = SE_{sampling} + SE_{ASDE} \tag{5}$$

where SE_{total} is the total standard error, $SE_{sampling}$ the standard error due to sampling and SE_{ASDE} the standard error as a result of ASDE. The value of SE_{total} is usually estimated by ignoring the ASDE variance component, that is implicitly setting $SE_{ASDE} = 0$, and using a classical method for deriving the estimate's variance, so

$$SE_{total}^{(0)} = SE_{sampling} \tag{6}$$

The central limit theorem states that if you have a population with mean μ and variance σ^2 and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. This means that the normal probability model can be used to quantify uncertainty when making inferences about the population mean based on the sample mean, resulting in the well-known confidence interval of $\pm 1.96SE$.

This paper proposes to compute an estimate $SE_{total}^{(1)}$ using a bootstrap procedure that is based on the following scheme

$$\begin{gathered}[\text{simpel random sample without replacement}] \rightarrow [\text{data collection}] \rightarrow \\ [\text{bootstrap}] \rightarrow [\text{bootstrap data sets}] \rightarrow [\text{ASDE}] \\ \rightarrow [\text{table of data } y_i] \rightarrow [\text{estimate } \bar{y}]\end{gathered}$$

With a bootstrap procedure it is possible to simulate multiple samples from the population by resampling with replacement from the original sample. Computing the variance of the obtained distribution of simulated sample means provide an alternative measure for the classical method of estimation of the estimate's reliability. This way, the classical reliability estimate $SE_{total}^{(0)}$ that does not take ASDE variance into account can be compared with the full bootstrap estimate $SE_{total}^{(1)}$ that does take ASDE variance into account. It will provide a measure for the extent to which the classical estimate functions as a reliable method for estimating the true standard error in analyses where ASDE is involved. Such comparisons will also be made between edit steps, in order to analyse the effect of each step on the standard error separately.

Furthermore, a comparison between the bootstrap distribution of the mean before and after ASDE will be made. A shift in expectation shows that ASDE possibly reduces bias but this may come at cost of an increase in variance.

Additionally, possible influence of the order of applying editing steps to the data is investigated by repeating the editing process for various data editing schemes in which the order of edit steps vary. This will provide valuable insight into the (relative) quality of data editing techniques.

3 Simulation

3.1 Input: Data and edit set

In order to conduct the research several bootstrap simulation studies are performed using R (version 3.2.3, R Core Team, 2015). A data set from Statistics Netherlands on wholesalers for the year 2007 is used, containing 323 records on 101 business variables. From this data set a subset of 17 main business variables is formed in order to keep the simulation output of reasonable size. 16 of the selected variables are target variables that contain information on financial amounts like *costs* and *turnover*. One variable is used as auxiliary variable which contains information about the number of staff in the company. Data on this auxiliary variable is required for the imputation of erroneous values in the last edit step *ratio imputation* of the ASDE process. For 12 records imputations cannot be realized because information on *number of staff* is missing, meaning that the data will still contain errors after the ASDE process is completed. In order to derive a complete and error free dataset these records are removed. Since the purpose of the research is to compare two standard error estimation methods (classic versus bootstrap) on the same data, the missingness can be ignored.

Besides data, the simulations also need an edit set as input. The 17 variables are selected so that a small edit set can be created, containing the following rules:

1. *Total turnover = net turnover + other turnover*

2. *Total costs = depreciation + cost of goods sold + staff costs + other costs*
3. *Operating profit before value adjustments = total turnover - total costs*
4. *Provision result = release of provisions - creations of provisions*
5. *Profit before tax = total income - total costs + capital gain/losses + financial result + provision result + exceptional gain/losses*

The selection of 17 out of 101 variables restricts the number of rules to five. When all variables are part of the model there are many rules that can be added to the edit set, which is assumed to enhance the data quality. Again, the purpose of this research is in the first place to compare two methods for reliability estimation, and not to improve data quality. Therefore, five edit rules are sufficient for this research.

3.2 Number of iterations

In order to produce reliable statistics the bootstrap simulation needs to be converged, meaning that repetition of the procedure yields similar results. Therefore, the number of bootstrap iterations (amount of simulated samples) are determined by repeating the bootstrap procedure for a sequence of increasing numbers. Results show convergence of the bootstrap at 8000 iterations.

3.3 Data editing steps

This paragraph discusses the 7 edit steps that are applied in the ASDE process of this research separately. These steps correspond to the editing techniques currently used for economic surveys (Pannekoek et al., 2013).

3.3.1 Thousand error correction

The first step in the edit process is the so called thousand error correction. In business surveys respondents are often informed to report the financial amounts in thousands of Euros, instead of Euros. When respondents ignore this instruction the reported amounts are a factor one thousand larger than they in fact are, and as a

consequence a thousand error occurs. This type of error can lead to substantial bias in aggregates, however, thousand errors can easily and reliably be corrected because the underlying error mechanism is clear de Waal et al. (2011a). It is not likely that this edit step will contribute to the estimate's variance. Bootstrap simulations need substantially more iterations to converge when thousand error correction is performed within the bootstrap procedure. Assuming that this edit step does not contribute to the estimator's variance, it is performed before the bootstrap simulation for time saving reasons.

Thousand error detection is done using a function that recognizes the value of target variable y in record i as a thousand error when

$$\left|\frac{y_i}{x_i}\right| > 100 \cdot median(\frac{y_1}{x_1}, \frac{y_2}{x_2} \ldots, \frac{y_n}{x_n}) \tag{7}$$

Where x_i is auxiliary variable *number of staff* for record i and n the number of records in the data. The absolute value of the ratio $\frac{y_i}{x_i}$ is taken because otherwise negative values are possibly not detected as thousand errors. When a value y_i is recognized as thousand error the value is replaced by $y_i/1000$.

3.3.2 Typing, sign and rounding error correction

The remaining steps of the edit process are executed within the bootstrap simulation, which uses the thousand error corrected data as input to simulate samples from.

The three steps after thousand error correction concern detection and correction of typing (discussed in paragraph 1), sign and rounding errors. Edit step *rounding* is configured in a way that it detects an error as rounding error when the violation of an edit rule can be solved by changing the involved values with no more than 2 units of measurement. Configuration of edit step *signs* is such that it detects a sign error when an inconsistency with respect to the edit rule can be solved by changing a sign symbol of the involved values. For a more comprehensive description of these edit steps refer to Scholtus (2008).

These three type of errors can deductively be corrected. Deductive correction

entails methods which use information available in inconsistent records to deduce and solve the probable cause of error. A number of algorithms for deductive correction have been proposed by (Scholtus, 2008, 2009) and are implemented in the `deducorrect` package in R. This package provides three functions which use available data in a record to detect and correct errors(Van der Loo and de Jonge, 2012):

1. `correctRounding` corrects rounding errors in numerical records that cause violations of linear equality rules. The method works by making small changes to a large enough set of randomly chosen variables.

2. `correctTypos` corrects typing errors in numerical records that cause violations of linear equality rules. The method works by computing correction suggestions and checking which suggestions correspond to correcting a typing error.

3. `correctSigns` corrects sign flips and value swaps in numerical records which violates linear equality rules. The method minimizes the number of value swaps and sign flips via a binary programming formulation.

These three edit steps will be applied in six orders, resulting in six different edit schemes, shown in table 2 for which bootstrap simulations are formed.

Table 2: Six different edit schemes in which the order of edit steps typo, rounding and signs vary

	Step						
Editing scheme	1	2	3	4	5	6	7
1	Thousand	Typo	Rounding	Signs	Deductive imputation	Ratio imputation	Adjust
2	Thousand	Typo	Signs	Rounding	Deductive imputation	Ratio imputation	Adjust
3	Thousand	Rounding	Typo	Signs	Deductive Imputation	Ratio imputation	Adjust
4	Thousand	Rounding	Signs	Typo	Deductive imputation	Ratio imputation	Adjust
5	Thousand	Signs	Rounding	Typo	Deductive imputation	Ratio imputation	Adjust
6	Thousand	Signs	Typo	Rounding	Deductive imputation	Ratio imputation	Adjust

The last two steps of the edit procedure, discussed in the next subsection, entail imputation of values which cannot be corrected by rounding, typo or sign corrections. Therefore these steps cannot be placed in an other position of the edit process.

3.3.3 Error localization and imputations

The last two steps in the edit procedure concern the imputation of incorrect or missing values that could not be corrected by the previous edit steps *thousand*, *rounding*, *typos* and *signs* in the ASDE process. Before errors and missing values can be imputed, localization of errors is performed with the localizeErrors function from the `editrules` package in R (de Jonge and Van der Loo, 2015). This function localizes errors according to *Felligi and Holt's principle*. Under the Fellegi-Holt principle (Fellegi and Holt, 1976), the variables to be imputed are determined by making changes to the smallest possible number of variables so as to ensure that the record passes all of the edits. The input that is given to the localizeErrors function is a data set and an edit set. Values which do not pass all of the edit rules in the edit set are localized as errors and set to be missing. This results in a data set only containing correct values and missing values.

After error localization, where possible, the fifth step of the edit process is performed; localized values are imputed with deductive imputation. That is, if the value of a missing variable to be imputed is determined uni-vocally by the edit set, this value will be imputed. This is conducted by the deduImpute function from `deducorrect` package in R (Van der Loo et al., 2015).

In case values cannot be deductively imputed the sixth step of the edit process, *ratio imputation*, is executed. Ratio imputation can be applied for missing values on a quantitative target variable y, if a quantitative auxiliary variable x can be found which has an approximately constant ratio with target variable y (lsraëls et al., 2011). This research includes only quantitative variables and these, like *costs* or *turnover*, are all related to the size of the business. They are assumed to have an approximately constant ratio with the number of staff since that variable can be considered to be a measure of size (de Waal et al., 2011b) , therefore ratio imputation is an appropriate method to use here. In this paper R represents the ratio between the mean of target variable y and the mean of auxiliary variable *number of staff*

(x). The missing value on variable y in record i (y_i) is replaced by:

$$\tilde{y}_i = Rx_i \tag{8}$$

Where y can be any target variable in the data and x_i is the number of staff in record i.

The seventh edit step concerns adjustment of the ratio imputation and is called edit step *adjust*. After imputation it can happen that some of the edit rules are violated by the imputed values. In that case the imputed values are adjusted, as little as possible, such that the edit rules are satisfied. Weights are assigned to determine the amount of adjustment for each variable, variables that can be imputed more accurately can be adjusted less than variables that are imputed. Function `adjustRecords` from the R package `rspa` gives the option to include a weight vector to the model.

3.4 Simulation output

The output of the six bootstrap simulations are the means of 16 target variables after each step in the ASDE process. For each edit step and target variable, this results in the bootstrap distribution of 8000 simulated sample means per simulation. The standard deviation of the distribution of means is used to create a bootstrap standard error of the mean, which can be compared to the classical estimation of the standard error.

4 Results

A subset of the data including variables *total turnover*, *staff costs*, *creation of provisions*, *financial result* and *exceptional result* will be used as examples to discuss results in more detail. This subset provides a good representation of business data by means of the type of financial amounts and the range of values they can take on. In addition, they cover the effects that are shown by the 16 target variables.

4.1 Compare classic to bootstrap standard error estimation

Results of the bootstrap simulation (edit scheme 1) show that estimation of the standard error with the classic method and the bootstrap method yield overall similar results, except for variables *release of provisions*, *creation of provisions* and *financial result*. Some of the outcomes are visualized by the subset of representative variables in table 3. This table shows the classic standard error, the bootstrap standard error and the differences in % of the bootstrap standard error compared to classic standard error, relative to the classic standard error. The variables that have different

Table 3: Bootstrap standard error compared to classic standard error

	Standard error		
	Classic	Bootstrap	% difference
Total turnover	12920,914	12857,654	-0,490
Staff costs	279,500	277,905	-0,571
Creation of provisions	63,218	93,021	47,143
Financial result	8,795	14,054	59,796
Exceptional result	19,402	19,501	0,508

Note. The financial amounts concern Euros in thousands

results over the two estimation methods, show that the estimates obtained with the bootstrap approach are substantially higher, approximately 50%, than the estimates obtained by classical estimation.

4.2 Influence of the edit steps on the differences in standard error

Table 4 present a stepwise overview of the mean and standard error after each edit step in the bootstrap simulation. Variables *total turnover* and *staff costs* are hardly influenced by any of the edit steps. Over all, for variables that are affected by the edit steps, edit steps involving imputation show substantial influence on the standard error.

To investigate the cause of the influence of ASDE on the the standard error, the ASDE procedure is run on the original data outside the bootstrap process. This gives insight in the performed corrections by the edit steps and the subsequent consequences of these corrections for the standard error.

Table 4: Bootstrap mean and standard error over different steps in the ASDE process

	Total turnover		Staff costs		Creation of provisions		Financial result		Exceptional result	
Step	Mean	Standard error	Mean	Standard error	Mean	Standard error	Mean	Standard error	Mean	Standard error
Thousand	46905,329	12857,670	2457,290	277,907	253,796	205,543	0,163	0,241	0,162	0,164
Typos	46905,329	12857,670	2457,290	277,907	253,796	205,543	-8,786	9,084	0,162	0,164
Rounding	46905,328	12857,670	2457,290	277,907	253,796	205,543	-8,788	9,084	0,158	0,164
Signs	46905,328	12857,670	2457,290	277,907	253,796	205,543	-8,788	9,084	0,158	0,164
Deductive imputation	46905,841	12857,654	2457,289	277,905	185,034	195,111	-3,289	14,355	-42,238	41,654
Ratio imputation	46905,841	12857,654	2457,289	277,905	211,549	181,275	-3,634	13,507	-19,627	19,418
Adjust	46905,841	12857,654	2457,289	277,905	140,067	93,021	-4,725	14,054	-20,001	19,501

It is beyond the scope of this paper to go into every detail of the performed corrections, but the ASDE process of variables *total turnover* and *creation of provision* will provide some examples.

Comparing the raw data to the data after ASDE revealed that data on variable *total turnover* is corrected for two records, both concerning the correction of thousand errors.

The influence of edit steps concerning imputation is illustrated by the ASDE process of *creation of provision*. For this variable 240 records are corrected during the ASDE process; 235 corrections concern the imputation of records for which data on *creation of provision* was missing in the raw data. Corrections performed by edit step *deductive imputation* concern the deductive imputation of 26 missing values. *Creation of provision* is involved in one edit rule

$$\textit{provision result} = \textit{release of provisions} - \textit{creations of provisions}$$

meaning that the imputation is based on the other two involved variables. Before edit step *deductive imputation*, data on record 164 for *release of provision* is 1070,000, *provision result* and *creation of provision* are missing. After *deductive imputation* these missing values are respectively imputed with 12868,270 and -11798,270. Deductive imputation is only possible when one value in the rule in unknown. Therefore, *provision result* is imputed based on another edit rule where it is involved. The imputation of variable *creation of provision* for record 165 with value -11798,270 changes the range from [-24,000, 14790,000] to [-11798,270, 14792,000], which has a strong influence on the mean.

Edit step *ratio imputation* is responsible for the imputation of the remaining

erroneous values. Imputations based on the ratio

$$mean(creation\ of\ provision)/mean(number\ of\ staff)$$

leads to a decreasing standard error. The covariation between these variables is find to be low.

In the last step of the ASDE process, *adjust* adjusts the imputations performed by *ratio imputation*, resulting in a substantial further decline of the standard error.

4.3 Compare bias before and after ASDE

The comparison of the bootstrap results before and after ASDE is shown 5. The bootstrap input data concerns the thousand error corrected data; the data before editing by edit steps with a stochastic component which can add to the uncertainty of an estimate. The input data is compared to the bootstrap output data resulting from the last step *adjust* of the ASDE process; the data that is edited by edit steps that can add uncertainty to the estimate. Presented results show that, although, the bias increases after ADSE for some of the variables, it remains smaller than the standard error and can therefore be ignored.

Table 5: Bootstrap statistics of the data before and after editing

	Bootstrap input data			Bootstrap output data		
	Original sample mean	Bias	Standard error	Original sample mean	Bias	Standard error
Total turnover	46938,059	-32,730	12857,670	46939,036	-33,195	12857,654
Staff costs	2450,242	7,047	277,907	2450,242	7,046	277,905
Creation of provisions	253,395	0,401	205,543	148,042	-7,974	93,021
Financial result	0,162	0,002	0,241	-7,161	2,436	14,054
Exceptional result	0,161	0,000	0,164	-18,137	-1,864	19,501

4.4 Influence of different editing schemes

The influence of the order of applying data editing steps is investigated by replication of the bootstrap procedure for the six different editing schemes (Table 2). For each variable, the standard error after the ASDE process is compared for the six editing schemes. Results show that the standard errors are similar.

5 Discussion

In the current paper, the effect of ASDE on the reliability of estimated means is investigated by a simulation study on survey data of wholesalers from Statistics Netherlands. The difference between classic estimated reliability and the actual reliability of the estimated means was investigated by comparing the standard error obtained by classic estimation after ASDE to the standard error obtained by the bootstrap approach that included ASDE as part of the estimation procedure. For this study, a subset of 16 financial variables was edited using a simple but realistic multi-step ASDE scenario. It is known that some of these steps include a stochastic or cross-record dependency which is not taken into account by classic estimation methods of the reliability.

Results of the simulation show that the classic method to estimate reliability can indeed underestimate the standard error of the mean when uncertainty is introduced by ASDE. A few variables suffered from a significantly higher standard error than others when estimated by the bootstrap approach. In the current case, these differences could be traced back to the imputation procedure, which was found to be unreliable regarding this variables. Due to a low covariation with the auxiliary variable *staff costs*, imputations based on the ratio lead to inadequate values. This result also shows that comparing classical estimation with estimation by a bootstrap approach provides an effective way of detecting issues with respect to applied edit steps in ASDE.

The example of edit step *deductive imputation* for variable *creation of provision* showed that this type of imputation strongly relies on the specified edit rules. Since this research did not specify the edit rule *creation of provision* > 0, this value is deductively imputed with a large negative value. This has serious consequences for the mean and standard error.

In current paper it was the aim to show that ASDE has an influence on reliability estimation. Therefore the small edit set and type of imputation method was appropriate to use. However, this might not be a good reflection of actual practice,

where much larger edit sets are used and the appropriateness of imputation methods is extensively examined beforehand. Future research with respect to the reliability of estimates in case ASDE is involved could focus on an elaboration of the edit set to represent a more realistic situation. Furthermore, different imputation methods could be used to investigate possible influences of improvement in the ASDE procedure on reliability estimation.

References

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*, Volume 558. New York: John Wiley & Sons.

de Jonge, E. and M. Van der Loo (2015). *editrules: Parsing, Applying, and Manipulating Data Cleaning Rules*. R package version 2.9.0.

de Waal, T., J. Pannekoek, and S. Scholtus (2011a). The editing of statistical data: Methods and techniques for the efficient detection and correction of errors and missing values. Technical report, Statistics Netherlands, The Hague.

de Waal, T., J. Pannekoek, and S. Scholtus (2011b). *Statistical Data Editing and Imputation*. New York: John Wiley & Sons.

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.

Hoogland, J., M. Van der Loo, J. Pannekoek, and S. Scholtus (2010). Methodenreeks: Thema controle en correctie. Technical report, Statistics Netherlands, The Hague.

Kim, J. K. and M. Park (2010). Calibration estimation in survey sampling. *International Statistical Review 78*(1), 21–39.

lsraëls, A., L. Kuyvenhoven, J. van der Laan, J. Pannekoek, and E. Schulte Nordholt (2011). Imputation. Technical report, Statistics Netherlands, The Hague.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rao, J. (2003). *Small Area Estimation*. New York: John Wiley & Sons.

Scholtus, S. (2008). Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Technical report, Statistics Netherlands, The Hague.

Scholtus, S. (2009). Automatic correction of simple typing error in numerical data with balance edits. Technical report, Statistics Netherland, The Hague.

Van der Loo, M. and E. de Jonge (2012). Manipulation of conditional restrictions and error localization with the editrules package. Technical report, Statistics Netherlands, The Hague.

Van der Loo, M., E. de Jonge, and S. Scholtus (2015). *deducorrect: Deductive Correction, Deductive Imputation, and Deterministic Correction*. R package version 1.3.7.