

**UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

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

Evaluating the quality of business survey data before and after automatic editing

Prepared by Sander Scholtus, Bart Bakker, and Sam Robinson, Statistics Netherlands

I. Introduction

1. Statistical results can be affected by measurement errors in the underlying data. National statistical institutes (NSIs) and other producers of official statistics therefore edit their data for errors as part of the process of generating statistical output (De Waal et al., 2011). Statistics Netherlands uses both automatic and manual editing in the production of economic statistics. Automatic editing methods are more efficient than manual editing – in terms of both costs and time – and yield results that are reproducible (Pannekoek et al., 2013). On the other hand, it is generally believed that the measurement quality of automatically edited data is lower than that of manually edited data (EDIMBUS, 2007).

2. In this paper, we propose to evaluate the measurement quality of automatically edited survey data in an objective way, by modelling the residual measurement errors in the data. We will compare the quality of an observed variable in the Netherlands' Structural Business Statistics (SBS) before and after automatic editing, in terms of validity (correlation of the observed variable to the true variable of interest) and bias (systematic deviation between the observed variable and the variable of interest). To identify the model, the survey data before and after automatic editing are linked to data from administrative sources.

3. We use a variant of the measurement model of Guarnera and Varriale (2015, 2016). In this model, the true variable of interest is represented by a latent (unobserved) variable and it is assumed that measurement errors in the observed variables occur according to a so-called "intermittent" mechanism. This means that each observation has a certain non-zero probability of being error-free (i.e., equal to the true value of the latent variable). This assumption seems appropriate for the data at hand and is also in line with existing methodology for data editing [see, e.g., Di Zio and Guarnera (2013) and Section II.A below]. The observations that do contain errors are modelled using linear regression techniques.

4. The remainder of this paper is organised as follows. We begin by briefly describing the data editing process for the Netherlands' SBS and the data that will be used here in Section II. An introduction to the measurement model is given in Section III. Section IV contains results of applying this model to our data. Possible implications and limitations of these results are discussed in Section V. Finally, some concluding remarks follow in Section VI.

II. Application

A. Automatic editing in the Netherlands' SBS

5. The SBS aim to provide an overview of employment and the financial structure (costs and revenues) of different sectors of the economy. Data are collected in a sample survey of businesses. The

sample is stratified by type of economic activity and size class. Businesses are classified by main economic activity according to the so-called NACE classification. We use the term “NACE group” to refer to a stratum of units with similar economic activities for which separate SBS estimates are published. The SBS questionnaire is tailored separately to each NACE group. On average, the SBS questionnaire produces a data set of about 100 different variables.

6. Figure 1 provides a schematic overview of the automatic editing process for the Netherlands’ SBS. Each box corresponds to a version of the data and each arrow between boxes corresponds to a process step during which changes can be made to the data.



Figure 1. Overview of the process for automatic editing in the Netherlands’ SBS.

7. The automatic process steps are, in chronological order:
- *Input processing (IP)*: Technical checks on the initial, unedited data and correction of uniform thousand-errors.
 - *Deductive processing 1 (DP1)*: Deterministic IF-THEN-rules to resolve other common errors with a known cause.
 - *Error localisation and imputation (EL&I)*: Automatic error localisation followed by imputation of missing and discarded values.
 - *Deductive processing 2 (DP2)*: Deterministic IF-THEN-rules to resolve inconsistencies not handled during EL&I (e.g., consistency between financial variables and stock variables).

8. Of these process steps, the EL&I step is the most complex from a methodological point of view. In this step, the data are made consistent with a given set of restrictions (so-called edit rules) by replacing observed values with new values if necessary. The selection of values to change is based on the paradigm of Fellegi and Holt (1976), which aims to minimise the number of changed values given that the resulting record has to satisfy all restrictions. This leads to a mathematical optimisation problem which can be solved automatically (De Waal et al., 2011).

9. The methodology of the two deductive processing steps DP1 and DP2 is less complex. These steps consist of applying a number of deterministic rules that can make changes to the data. An example of a rule that is used during process step DP1 is:

IF Depreciations < 0 THEN Depreciations := – Depreciations.

According to this rule, if any negative values are encountered for the variable Depreciations, these have to be replaced by their absolute values. There is in fact an edit rule (restriction) in the SBS which states that the value of Depreciations must be non-negative.

10. We will not describe the process steps of Figure 1 in more detail here. A general overview of methodology for automatic data editing can be found elsewhere, e.g., in De Waal et al. (2011) or Pannekoek et al. (2013). A detailed description of the data editing process of the Netherlands’ SBS is provided by De Jong (2002) and Hoogland and Smit (2008).

11. A feature of the above automatic editing process is that, during each process step, the vast majority of observed values are not changed. Thus, most of the observed values in the unedited data (first box in Figure 1) are equal to the corresponding values in the edited data (final box in Figure 1). This happens because the editing methods used for the Netherlands’ SBS all assume, either explicitly or implicitly, that most of the observed values are correct to begin with. For instance, the Fellegi-Holt paradigm that is used during the EL&I step is based on the assumption that errors are rare, and that a

record should therefore be made consistent with the edit rules by changing the least possible number of values.

12. For this study, we want to compare the measurement quality of variables in the input data (second box in Figure 1) and edited data (right-most box). This will give an impression of the overall effect of automatic editing on data quality. We take the second box as a starting point rather than the first box, because some of the technical checks carried out during the IP step are required to know whether the data are accessible at all. In fact, for questionnaires that are submitted on paper – which is still done by a minority of responding units – the data are digitised as part of the IP step, so no unedited data are available in digital form for these units.

13. In the actual production process of the Netherlands’ SBS, only a subset of the data after input processing is treated by the remaining automatic process steps in Figure 1. The other records are edited manually instead. A selection procedure is applied to the input data to assign records either to automatic or manual editing (Hoogland, 2006). For the present study, we created a version of the data in which as many records as possible were edited automatically, regardless of the selection that was made during actual production. By focussing on this data set, we can evaluate the “pure” effect of automatic editing on the measurement quality of SBS data, rather than the combined effect of measurement and selection.

14. In practice, not all records can be edited automatically. During the IP step, records can be rejected (discarded from further automatic processing) if certain key variables such as total turnover are missing. During the EL&I step, a small number of records for which no solution to the error localisation problem can be found are also rejected. In the actual production process, these records would then be treated manually instead. For the purpose of this study, they are treated as non-response. Fortunately, this concerns only a handful of records.

B. Data

15. For this application, we used SBS data of reference year 2012 for four different NACE groups within the economic sector “Trade”. These NACE groups are listed in Table 1. The SBS data were linked to two different administrative data sets collected by the Netherlands’ tax authorities: Value-Added Tax declarations (VAT) on turnover and the Profit Declaration Register (PDR) which contains many administrative variables that are similar to SBS variables.

Table 1. Overview of NACE groups considered in this application.

NACE	Description
45112	Sale and repair of passenger cars and light motor vehicles
45190	Sale and repair of trucks, trailers, and caravans
45200	Specialised repair of motor vehicles
45400	Sale and repair of motorcycles and related parts

16. The same data have been analysed previously by Scholtus et al. (2015) using a different type of measurement error model, a so-called structural equation model (SEM). Some additional information about these data sources can be found in that paper. Under an SEM, errors are not supposed to be intermittent: an observed variable is either completely error-free or it is always affected by errors. In order to identify all parameters of their SEM, Scholtus et al. (2015) required the inclusion of error-free measurements for a small random subsample of the original data (an audit sample). They obtained these audit data by letting subject-matter experts re-edit the SBS data for 50 randomly selected units in each NACE group with the aim of recovering the true values for all variables, outside of regular production. Although such an audit sample is not needed for the model that will be used here (see Section III), we did include these records in the present analysis, since they were already available. [Note: One other difference is that Scholtus et al. (2015) used the original, production-edited SBS survey data in which not all records were edited automatically (cf. paragraph 13). Results of estimating an SEM for the exact same data that are analysed here can be found in Scholtus et al. (2017).]

Table 2. Number of units in each NACE group. All figures refer to 2012 and, apart from the first line, to the population with large and/or complex units excluded.

NACE group	45112	45190	45200	45400
population (total)	18,680	1,790	6,054	1,763
population (w/o large or complex units)	18,556	1,739	6,018	1,759
SBS net sample, edited	914	165	269	74
SBS net sample, edited and linked to admin. data	810	158	231	58
net audit sample	43	45	43	43

17. Table 2 lists the number of available records in each NACE group. The editing process for very large and/or complex units differs from that of the other units (in particular, they are never edited automatically), so these were not included in the present study (second line in Table 2).

18. The survey and administrative data could be linked through business identification numbers in Statistics Netherlands' General Business Register (GBR). The administrative sources contain information for fiscal units rather than statistical units. The relationship between fiscal and statistical units is known in the GBR, but not all administrative units can be linked to a single statistical unit. Therefore, and also due to missing data in the administrative sources, it was not possible to link all units in the SBS data to administrative data (fourth line in Table 2). Scholtus et al. (2015) investigated whether the linked data might suffer from selection bias but found no indication that such a bias occurred.

III. The intermittent-error model

19. The following model is based on the model of Guarnera and Varriale (2015, 2016) for intermittent errors in measurements of a single variable of interest from multiple sources. Let η_i denote the true value of a variable of interest for unit i ($i = 1, \dots, n$). Suppose that we do not observe this variable directly, but we do have observations on $K \geq 2$ variables y_1, \dots, y_K that measure η . For each observed variable y_k , a 0-1-indicator z_k is introduced such that $y_{ki} = \eta_i$ if $z_{ki} = 0$. For units with $z_{ki} = 1$, y_{ki} contains a measurement error which is described by a linear regression model:

$$y_{ki} = \begin{cases} \eta_i & \text{if } z_{ki} = 0, \\ \tau_k + \lambda_k \eta_i + e_{ki} & \text{if } z_{ki} = 1, \end{cases} \quad (1)$$

where τ_k and λ_k are constants and e_{ki} follows a normal distribution with mean zero and variance σ_k^2 . In a single formula, the model for y_k can be expressed as follows:

$$y_{ki} = (1 - z_{ki})\eta_i + z_{ki}(\tau_k + \lambda_k \eta_i + e_{ki}). \quad (2)$$

The probability of observing an error on y_k is represented by the parameter $\pi_k = P(z_k = 1) = E(z_k)$. It is assumed that, for each unit, all z_k and all e_k are independent across different observed variables. This implies that measurement errors in different observed variables for the same unit are uncorrelated.

20. In addition to the above measurement model, we also use an ordinary linear regression model to describe the variation in the true values η_i across units as a function of covariates \mathbf{x} :

$$\eta_i = \boldsymbol{\beta}' \mathbf{x}_i + u_i, \quad (3)$$

where $\boldsymbol{\beta}$ denotes a vector of regression coefficients and it is assumed that u_i is normally distributed with mean zero and variance σ^2 .

21. The parameters of the model given by (2) and (3) provide several interesting indicators for the measurement quality of each observed variable. Firstly, we can look at the error probability π_k . A value of π_k closer to 1 indicates that more errors occur for variable y_k . Secondly, the intercept τ_k and slope λ_k in (2) describe the effect of systematic measurement errors in y_k . To the extent that τ_k deviates from 0 and λ_k deviates from 1, the observed variable y_k is biased with respect to the true value of η . Finally, to quantify the effect of the random measurement errors e_k on y_k , we can use the so-called indicator validity coefficient (IVC). For ordinary factor models, the IVC of y_k as a measure of η is defined as the absolute value of the correlation between y_{ki} and η_i (Sarıs and Andrews, 1991). Under model (2), we

distinguish between $z_{ki} = 1$ and $z_{ki} = 0$. For the observations on y_k that contain errors, the IVC can be computed from the standardised value of the slope parameter λ_k (say, λ_k^s):

$$\text{IVC}(y_k|z_k = 1) = |\text{cor}(y_k, \eta|z_k = 1)| = |\lambda_k^s| = |\lambda_k| \sqrt{\frac{\text{var}(\eta|z_k = 1)}{\text{var}(y_k|z_k = 1)}} = \sqrt{1 - \frac{\sigma_k^2}{\lambda_k^2 \sigma_\eta^2 + \sigma_k^2}}. \quad (4)$$

Here, σ_η^2 denotes the total variance of η which, under model (3), is given by $\sigma_\eta^2 = \boldsymbol{\beta}' \boldsymbol{\Sigma}_{xx} \boldsymbol{\beta} + \sigma^2$, where $\boldsymbol{\Sigma}_{xx}$ denotes the variance-covariance matrix of \mathbf{x} . (Note that, in the presence of covariates, σ^2 represents the unexplained variance in η .) Furthermore, the error-free y_{ki} can be seen as observations with $\text{IVC} = 1$. Hence, a natural definition of the unconditional indicator validity coefficient of y_k under model (2) is:

$$\text{IVC}(y_k) = \pi_k \times \text{IVC}(y_k|z_k = 1) + (1 - \pi_k) \times 1 = 1 - \pi_k \left(1 - \sqrt{1 - \frac{\sigma_k^2}{\lambda_k^2 \sigma_\eta^2 + \sigma_k^2}} \right). \quad (5)$$

Both $\text{IVC}(y_k|z_k = 1)$ and $\text{IVC}(y_k)$ lie between 0 and 1. Values close to 1 indicate a strong linear relationship between the observed value of y_k and the true value of η .

22. Under this model, there is a non-zero probability (namely $1 - \pi_k$) that an observed value y_{ki} is equal to the true value η_i and therefore error-free. The event that two different observed values y_{ki} and y_{li} for a given unit i are identical occurs with probability $(1 - \pi_k)(1 - \pi_l)$. Note that if we observe $y_{ki} = y_{li}$, then it must hold that $y_{ki} = y_{li} = \eta_i$. This follows from the above assumptions that the errors are normally distributed and independent across the observed variables, since the probability of drawing any specific value from a normal distribution equals zero. In fact, the same property holds for any random variable with a continuous distribution, so we do not need the assumption of normality here. Thus, under the assumptions of this model, if a record contains the same value for two (or more) observed variables, then that value must also be equal to the corresponding true value: it is possible to recognise some of the error-free values from the observed data themselves. Of course, all of this need not be true if the model does not hold for the data at hand. In particular, the assumption that errors in different variables are independent is a strong assumption that may not always be satisfied in practice.

23. Model (2)–(3) is an example of a so-called finite mixture model; see, e.g., McLachlan and Peel (2000). If, in addition to the observed variables y_1, \dots, y_K , we would also have observed all values of η (and thus indirectly observed all error patterns specified by z_1, \dots, z_K), maximum likelihood estimates of the model parameters could be obtained as follows:

$$\begin{aligned} \hat{\pi}_k &= \frac{n_k}{n}, \\ \hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i \eta_i \right), \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (\eta_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i)^2, \\ \hat{\lambda}_k &= \frac{n_k \sum_{i=1}^n z_{ki} \eta_i y_{ki} - (\sum_{i=1}^n z_{ki} \eta_i) (\sum_{i=1}^n z_{ki} y_{ki})}{n_k \sum_{i=1}^n z_{ki} \eta_i^2 - (\sum_{i=1}^n z_{ki} \eta_i)^2}, \\ \hat{\tau}_k &= \frac{1}{n_k} \sum_{i=1}^n z_{ki} y_{ki} - \frac{\hat{\lambda}_k}{n_k} \sum_{i=1}^n z_{ki} \eta_i, \\ \hat{\sigma}_k^2 &= \frac{1}{n_k} \sum_{i=1}^n z_{ki} (y_{ki} - \hat{\tau}_k - \hat{\lambda}_k \eta_i)^2, \end{aligned} \quad (6)$$

with $n_k = \sum_{i=1}^n z_{ki}$, the number of observations with an error on y_k .

24. In practice, not all η_i and z_{1i}, \dots, z_{Ki} are observed – but some of them are. As noted above, the true value η_i (and hence all z_{ki}) can be inferred from the observed data when $y_{ki} = y_{li}$ for (at least) two different observed values. To give an example: in the presence of $K = 3$ observed variables there are $2^3 = 8$ possible error patterns. We can derive η_i and the error pattern for all observations with

$$(z_{1i}, z_{2i}, z_{3i}) \in \{(0,0,0), (1,0,0), (0,1,0), (0,0,1)\}, \quad (7)$$

as for these observations at least two of the observed values are identical. For the remaining observations, we can infer that (z_{1i}, z_{2i}, z_{3i}) must have one of the four remaining patterns – that is, at least two of the three observed values must be erroneous. However, we cannot derive the exact error pattern and we cannot obtain the value of η_i for these observations.

25. Guarnera and Varriale (2016) noted that, to handle the cases with missing data on η , an EM algorithm (Expectation–Maximisation) could be used. They worked out how to obtain maximum likelihood estimates for a model with $K = 3$ observed variables, under the restriction that $\tau_k = 0$ and $\lambda_k = 1$. Robinson (2016) gives a detailed description of this EM algorithm, including an extension to estimate τ_k and λ_k . For the present study, an implementation of this algorithm was written in R. We refer to Little and Rubin (2002) for an introduction to EM algorithms in general. Below we also report asymptotic standard errors for the estimated parameters. These standard errors were obtained with the aid of a so-called Supplemented EM algorithm; see Little and Rubin (2002, pp. 191–196).

26. In principle, the above model could be extended to multiple target variables (Guarnera and Varriale, 2016) but an estimation procedure for such an extended model has not yet been developed. Note that it is possible to introduce covariates \mathbf{x}_i to predict the true value of the target variable, but potential errors in these covariates are not taken into account. Another important limitation of this model is that it relies heavily on the assumption that measurement errors in different observed variables are independent. Thus, correlated measurement errors cannot be taken into account under the model as formulated here.

27. In the application to be discussed below, the variable of interest was the total Turnover of a business. We used measurements of Turnover from four different sources: VAT, PDR, SBS-input and SBS-edited. However, measurement errors in SBS-input and SBS-edited are likely to be highly correlated, because any errors that occur in the input data and are not resolved during automatic editing will also be present in the edited data. We therefore estimated two separate models with either SBS-input or SBS-edited included besides VAT and PDR, so $K = 3$. In addition, as mentioned in paragraph 16, for a small subsample of units an audited version of Turnover was available which is supposed to measure η_i with certainty. In the context of the intermittent-error model, this simply means that η_i and (z_{1i}, z_{2i}, z_{3i}) can be inferred from the observed data for some additional units. All observed Turnover values were rounded to multiples of 1000 Euros, as this is the way they were reported in the SBS survey.

28. The above maximum likelihood estimation procedure assumes that (i) the residuals u_i in (3) are $N(0, \sigma^2)$ distributed and (ii) the measurement errors e_{ki} in (2) are $N(0, \sigma_k^2)$ distributed. For our data on Turnover, neither of these assumptions holds. By examining the cases with error patterns (7), for which η_i is known, it was found that the above assumptions were more reasonable after applying a logarithmic transformation to the data. We will therefore present results obtained with all variables measured on a log scale. To be precise, for error-prone observations ($z_{ki} = 1$), the measurement model (1) was replaced by

$$\log(y_{ki} + 0.5) = \tau_k + \lambda_k \log(\eta_i + 0.5) + e_{ki}, \quad (8)$$

where \log denotes the natural logarithm. We used $\log(y + 0.5)$ rather than $\log y$ to be able to handle cases with $y = 0$. A similar transformation was applied to (3). Note that model (8) is equivalent to

$$y_{ki} + 0.5 = \exp(\tau_k) (\eta_i + 0.5)^{\lambda_k} \exp(e_{ki}). \quad (9)$$

Thus, for the variables on the original scale the error structure is now multiplicative rather than additive.

IV. Results

29. Table 3 shows a selection of estimated parameters for the model with SBS input data; Table 4 shows the corresponding results for the model with SBS edited data. In all models, we used the observed (edited) Number of employees and Total operating costs in SBS and a constant as covariates \mathbf{x}_i to predict the true value of Turnover. In these tables, we have included both the conditional IVC (4) for error-prone observations (rows labelled “IVC-cond.”) and the unconditional IVC (5) for all observations (rows

labelled “IVC”). Note that the unconditional IVC is always closer to 1 than the conditional IVC, due to the contribution of observations that are error-free.

Table 3. Estimated indicator validity and measurement parameters (error probability, intercept and slope) for Turnover on a log scale (with standard errors); model with SBS input data for Turnover.

parameter	45112		45190		45200		45400	
	estimate	std.err.	estimate	std.err.	estimate	std.err.	estimate	std.err.
π (VAT)	0.91	0.01	0.82	0.03	0.76	0.03	0.85	0.05
τ (VAT)	-0.26	0.06	0.67	0.22	0.07	0.10	-0.19	0.26
λ (VAT)	1.00	0.01	0.90	0.03	0.99	0.02	0.99	0.04
IVC-cond. (VAT)	0.98		0.96		0.98		0.97	
IVC (VAT)	0.98		0.96		0.98		0.98	
π (PDR)	0.64	0.02	0.34	0.05	0.55	0.03	0.68	0.06
τ (PDR)	0.04	0.11	0.85	0.55	0.26	0.13	-0.07	0.06
λ (PDR)	0.99	0.01	0.89	0.08	0.97	0.02	1.01	0.01
IVC-cond. (PDR)	0.96		0.84		0.98		1.00	
IVC (PDR)	0.97		0.95		0.99		1.00	
π (SBS, input)	0.10	0.01	0.40	0.05	0.13	0.02	0.08	0.04
τ (SBS, input)	0.68	0.68	-0.25	0.10	1.21	1.80	-1.87	0.89
λ (SBS, input)	0.80	0.11	1.04	0.01	0.40	0.31	1.20	0.17
IVC-cond. (SBS, input)	0.70		1.00		0.26		0.92	
IVC (SBS, input)	0.97		1.00		0.91		0.99	

Table 4. Estimated indicator validity and measurement parameters (error probability, intercept and slope) for Turnover on a log scale (with standard errors); model with SBS edited data for Turnover.

parameter	45112		45190		45200		45400	
	estimate	std.err.	estimate	std.err.	estimate	std.err.	estimate	std.err.
π (VAT)	0.91	0.01	0.82	0.03	0.76	0.03	0.85	0.05
τ (VAT)	-0.21	0.06	0.67	0.22	0.08	0.10	-0.20	0.26
λ (VAT)	0.99	0.01	0.90	0.03	0.99	0.02	0.99	0.04
IVC-cond. (VAT)	0.98		0.96		0.98		0.97	
IVC (VAT)	0.98		0.96		0.98		0.98	
π (PDR)	0.64	0.02	0.34	0.05	0.54	0.03	0.68	0.06
τ (PDR)	0.11	0.12	0.85	0.55	0.29	0.13	-0.08	0.06
λ (PDR)	0.98	0.01	0.89	0.08	0.96	0.02	1.01	0.01
IVC-cond. (PDR)	0.95		0.84		0.97		1.00	
IVC (PDR)	0.97		0.95		0.99		1.00	
π (SBS, edited)	0.11	0.02	0.40	0.05	0.14	0.03	0.10	0.04
τ (SBS, edited)	-0.30	0.49	-0.25	0.10	0.72	0.90	0.08	0.84
λ (SBS, edited)	0.98	0.07	1.04	0.01	0.70	0.16	0.94	0.17
IVC-cond. (SBS, edited)	0.83		1.00		0.68		0.88	
IVC (SBS, edited)	0.98		1.00		0.96		0.99	

30. The main results that can be seen in these tables are:

- The estimated parameters for the administrative sources do not differ significantly between both tables. In other words, the choice between using input or edited SBS data does not affect the parameter estimates for observed variables in the other sources. This is a positive result.
- According to the model, the VAT variable has relatively large error probabilities in all NACE groups ($\hat{\pi}$ lies between 0.76 and 0.91), whereas the corresponding probabilities for PDR and SBS are smaller. In all NACE groups except 45190, SBS has the smallest error probabilities, both before and after automatic editing.
- It appears that automatic editing does not have a large impact on the overall IVC of SBS Turnover. The error probabilities before and after editing also do not differ significantly, and in some NACE groups the probability of observing an error after editing is actually slightly larger. On the other hand, we can see that after editing the intercept and slope parameters $\hat{\tau}$ and $\hat{\lambda}$ for the SBS variable are closer to 0 and 1, respectively, in all NACE groups except 45190, where no change occurs. In NACE groups 45112 and 45200, it is also seen that the conditional IVC for the error-prone observations is improved by the editing procedure; in NACE group 45400 editing actually slightly reduces the IVC.

V. Discussion

31. The results in Section IV suggest that, according to our measurement quality indicators, the effect of automatic editing on SBS Turnover is very limited. Looking at the data, this is not unexpected, because in fact only a small number of Turnover values in our data set were changed during the automatic editing process of Figure 1. This naturally leads to the question whether automatic editing has any added value for the Netherlands' SBS. We believe that it does, but that this added value is mainly related to other quality criteria that are not considered by the above measurement error model.

32. In particular, an important aim of automatic editing is to obtain a data set which is consistent with respect to a pre-defined set of edit rules. The edit rules define univariate and multivariate restrictions that would be expected to hold if the data were error-free. Two examples of edit rules for SBS are:

$$\begin{aligned} \text{Turnover} &\geq 0; \\ \text{Turnover} - \text{Total operating costs} &= \text{Profit}. \end{aligned}$$

It should be noted that it is possible that an SBS observation satisfies all edit rules while still containing one or more errors. On the other hand, an observation that does not satisfy all edit rules certainly contains errors, but these errors might be very small and therefore hardly affect the IVC or bias. For instance, it is not uncommon for SBS data to contain so-called rounding errors (De Waal et al., 2011, Chapter 2).

33. Regarding the importance for NSIs of obtaining data that are consistent with edit rules, Pannekoek and De Waal (2005) noted the following:

Statistically speaking there is indeed hardly any reason to let a data set satisfy all edits, other than the *hope* that enforcing internal consistency will result in data of higher statistical quality. NSIs, however, have the responsibility to supply data for many different academic and nonacademic users in society. For the majority of these users, inconsistent data are incomprehensible. They may reject the data as being an invalid source of information or make adjustments themselves. This hampers the unifying role of an NSI in providing data that are undisputed by different parties (...).

Thus, even when automatic editing does not significantly improve the measurement quality of a data set (in terms of validity and bias), it can still be useful as a relatively cheap way of obtaining consistent data.

34. On the other hand, the estimated error probabilities in Table 3 and Table 4 suggest that the automatic editing process has successfully corrected only a small subset of the errors in SBS Turnover that were actually present in these data. While the validity of the variables after editing is quite close to 1, the remaining errors do appear to cause a noticeable bias in some of the NACE groups. It may therefore be useful to apply a model during regular production to estimate the effects on statistical output of errors that remain in the data after automatic editing. If these effects are significant, improved output may be estimated by applying a correction for measurement errors. This correction requires a prediction of the true value η_i , given one or more of the observed values y_{ki} and the estimated model parameters. Predicted values for η based on all observed data are actually computed as part of the EM algorithm by which our model is estimated (Robinson, 2016), so the extension to a correction procedure is straightforward from a theoretical point of view.

35. Alternatively, a model could be used as part of a selective editing procedure during regular production to identify records that are likely to contain errors (either before or after automatic editing) for manual follow-up. Having estimated the model, one can compute the posterior probability that y_{ki} contains an error, taking into account all observed data and the estimated model parameters; see Guarnera and Varriale (2016). Note that for observations such as those in (7), these posterior probabilities are equal to 0 or 1. In combination with a measure of the expected error size (which could be derived from the predicted value of η_i under the model), these posterior probabilities can provide a basis for selecting the observations that are likely to contain the most important errors. See Di Zio and Guarnera (2013) for a detailed discussion of the use of intermittent-error models for selective editing.

36. As it stands, the model of Section III has some important limitations that were highlighted by the above application. Firstly, the assumption that the true values and measurement errors are normally

distributed may often be violated in practice. It is not known to what extent the maximum likelihood estimates for this model are robust to non-normality. In principle, other versions of the model could be developed for different distributions, but this has not been done yet. In fact, for many variables that occur in business statistics (such as Turnover) a log-normal distribution is reasonable, in which case the current model can be applied to the data after a logarithmic transformation. However, this does complicate the interpretation of the model parameters τ_k and λ_k , as can be seen in (8) and (9). An interesting alternative solution to handle non-normal data that is sometimes used in finite mixture models is to model these non-normal distributions themselves as mixtures of two or more normal distributions, which leads to a “mixture of mixture models” (McLachlan and Peel, 2000; Di Zio et al., 2007). It remains to be seen whether such an approach would work in our situation (e.g., identifiability might be a problem).

37. The model also assumes that measurement errors in different observed variables are independent. In principle, this assumption could be relaxed, but this would make the estimation procedure more complicated. Furthermore, for a given number of observed variables K , only a limited number of dependencies can be added before the model becomes under-identified.

38. In this application, we focussed on a single variable of interest (Turnover). Since automatic editing – in particular: error localisation based on the Fellegi-Holt paradigm – is a multivariate procedure, it would actually be more interesting to model several target variables simultaneously. A relatively straightforward extension of the model could be made if errors for different variables of interest are independent, but this assumption may often be unreasonable in practice. Without such an assumption, the model quickly becomes very complex as more latent variables are added (Guarnera and Varriale, 2016).

39. Finally, the maximum likelihood estimation procedure used here assumes that the data consist of independent, identically distributed observations. It would be good to extend the estimation procedure to take the effects of finite-population sampling into account, as survey observations are hardly ever independent in practice. This is likely to affect the standard errors of the estimated parameters.

VI. Conclusion

40. In this paper, we have used an intermittent-error model to estimate and compare the measurement quality of survey data on Turnover from the Netherlands’ SBS before and after automatic editing. The underlying assumption that some of the observed values are error-free appears to be suitable for SBS data. However, as indicated in Section V, the model has some important limitations. We maintain that it is useful to develop this model further to address these limitations.

41. In our application, we found that automatic editing methods had a minor effect on the validity and bias of business survey data. Overall, the measurement quality of the edited data was, at best, only marginally better than that of the input data. Of course, these results are based on a single data set for a small number of NACE groups. Also, the target variable Turnover is usually reported with relatively high accuracy in the SBS survey. Thus, these results may not extend to all applications of automatic editing in business statistics. Nevertheless, we can tentatively conclude that the main merit of automatic editing may be that it provides consistent data at low costs, but that it often does not significantly improve the measurement quality of individual variables in terms of validity or bias. This suggests that it would be good to develop measurement error models that can be used to estimate the effects of residual errors in edited data during regular production, which could then be used to correct statistics for measurement error, or to select observations for further manual editing. The intermittent-error model could be used as a starting point for the development of such a model for editing applications in business statistics.

VII. References

de Jong, A. (2002), Uni-Edit: Standardized Processing of Structural Business Statistics in the Netherlands. Working Paper No. 27, UN/ECE Work Session on Statistical Data Editing, Helsinki.

- de Waal, T., J. Pannekoek, and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, New Jersey.
- Di Zio, M. and U. Guarnera (2013), A Contamination Model for Selective Editing. *Journal of Official Statistics* **29**, 539–555.
- Di Zio, M., U. Guarnera, and R. Rocci (2007), A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. *Computational Statistics & Data Analysis* **51**, 2573–2585.
- EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Eurostat manual prepared by ISTAT, Statistics Netherlands, and SFSO.
- Fellegi, I.P. and D. Holt (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association* **71**, 17–35.
- Guarnera, U. and R. Varriale (2015), Estimation and Editing for Data from Different Sources. An Approach Based on Latent Class Model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing, Budapest.
- Guarnera, U. and R. Varriale (2016), Estimation from Contaminated Multi-Source Data Based on Latent Class Models. *Statistical Journal of the IAOS* **32**, 537–544.
- Hoogland, J. (2006), Selective Editing using Plausibility Indicators and SLICE. In: *Statistical Data Editing, Volume No. 3, Impact on Data Quality*, United Nations, New York and Geneva, pp. 106–130.
- Hoogland, J. and R. Smit (2008), Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics. Working Paper No. 2, UN/ECE Work Session on Statistical Data Editing, Vienna.
- Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data* (second edition). John Wiley & Sons, New York.
- McLachlan, G.J. and D. Peel (2000), *Finite Mixture Models*. John Wiley & Sons, New York.
- Pannekoek, J. and T. de Waal (2005), Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* **21**, 257–286.
- 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**, 511–537.
- Robinson, S.P. (2016), Modelling Measurement Errors in Linked Administrative and Survey Data. Master Thesis, Leiden University. Available at <http://www.math.leidenuniv.nl/en/theses/663>.
- Saris, W.E. and F.M. Andrews (1991), Evaluation of Measurement Instruments using a Structural Modeling Approach. In: Biemer, Groves, Lyberg, Mathiowetz, and Sudman (eds.), *Measurement Errors in Surveys*, John Wiley & Sons, New York, pp. 575–597.
- Scholtus, S., B.F.M. Bakker, and A. van Delden (2015), Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Discussion Paper 2015-17, Statistics Netherlands, The Hague. Available at <http://www.cbs.nl/nl-nl/achtergrond/2015/46/modelling-measurement-error-to-estimate-bias-in-administrative-and-survey-variables>.
- Scholtus, S., B.F.M. Bakker and S.P. Robinson (2017), Assessing the Quality of Business Survey Data before and after Automatic Editing. In: S. Scholtus, *Editing and Estimation of Measurement Errors in Statistical Data*, PhD Thesis, VU University, Amsterdam (forthcoming).