DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 619–642, http://dx.doi.org/10.1515/JOS-2016-0032

Accuracy of Mixed-Source Statistics as Affected by Classification Errors

Arnout van Delden[1], *Sander Scholtus*[1], *and Joep Burger*[2]

Publications in official statistics are increasingly based on a combination of sources. Although combining data sources may result in nearly complete coverage of the target population, the outcomes are not error free. Estimating the effect of nonsampling errors on the accuracy of mixed-source statistics is crucial for decision making, but it is not straightforward. Here we simulate the effect of classification errors on the accuracy of turnover-level estimates in car-trade industries. We combine an audit sample, the dynamics in the business register, and expert knowledge to estimate a transition matrix of classification-error probabilities. Bias and variance of the turnover estimates caused by classification errors are estimated by a bootstrap resampling approach. In addition, we study the extent to which manual selective editing at micro level can improve the accuracy. Our analyses reveal which industries do not meet preset quality criteria. Surprisingly, more selective editing can result in less accurate estimates for specific industries, and a fixed allocation of editing effort over industries is more effective than an allocation in proportion with the accuracy and population size of each industry. We discuss how to develop a practical method that can be implemented in production to estimate the accuracy of register-based estimates.

Key words: Accuracy; editing; administrative data; short-term business statistics; bootstrap resampling.

1. Introduction

Publications in official statistics are increasingly based on a combination of sources, for instance, a sample survey combined with an administrative source. The combination of data sources sometimes results in a situation where observations are available for nearly the complete target population, but that does not imply that the outcomes are error free. In fact, numerous error types may occur, as exhibited by the total survey error framework for sample surveys (Biemer and Lyberg 2003; Groves et al. 2009), adapted for administrative data by Zhang (2012a). We believe that it is important for NSIs to quantify the implications of those errors for the accuracy of statistical outcomes based on mixed sources, because NSIs aim to publish information of sufficient quality for users.

[1] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Email: a.vandelden@cbs.nl and s.scholtus@cbs.nl.
[2] Statistics Netherlands, Department of Process Development and Methodology, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: j.burger@cbs.nl.

Acknowledgments: We thank Arjen de Boer and Danny van Elswijk for providing the raw data and Ton Bonné, Lei Dirrix, Willem Heijnen, Marian Immerzeel, John Spring in ‘t Veld, and Ivonne Valent for providing their help with the audit sample and Harm Jan Boonstra and Ton de Waal for useful comments on earlier drafts.

Knowledge on the effect of errors on the accuracy of mixed-source statistics is also useful for operational decisions, for instance in the editing process. Time, costs, and quality constraints all play a role in the decision how many units are edited manually in a statistical process to improve data quality. To this end, 'selective editing' methods have been developed (de Waal et al. 2011). These methods aim to limit manual editing by focussing on units with a high risk of influential errors, where an 'influential error' is defined as one "that has a considerable effect on the publication figures" (de Waal et al. 2011). In addition to the influence of records on the *values* of the publication figures, the effect on the *accuracy* of the figures is also important.

Estimating the effect of nonsampling errors on the accuracy of estimates in practical situations is not very straightforward as yet. Depending on the complexity of the combined data sources and the type of nonsampling error, sometimes analytical approaches are possible (Burger et al. 2015; Zhang 2012b). In cases with complicated error structures or when the effects of different processing and estimation steps are taken into account, this may no longer be possible. Bryant and Graham (2015) estimated the uncertainty caused by nonsampling errors using a Bayesian approach. Burger et al. (2015) treated a simplified situation where they did a sensitivity analysis on classification errors for which they used both an analytical and a parametric bootstrap approach. A bootstrap approach can also be applied in more complex situations where an analytical solution cannot be found. In the current article, we proceed with this work towards a more realistic modelling of the error structure.

To illustrate the method, we look at a case study on the estimation of the quarterly turnover of the 'car trade' based on a combination of survey and administrative data. The figures are classified by (groupings of) economic activity according to NACE rev 2, henceforth referred to as industry codes. Determining the correct activity code of economic units is often rather difficult and prone to errors (e.g., Christensen 2008). Reasons are that the surveyed units often have a mixture of economic activities, that activities change over time but those changes are often not reported to the relevant administrative organisations, and that the distinction between different codes is sometimes fuzzy. Previous work on the same case study by Burger et al. (2015) suggested that the publication figures are rather sensitive to classification errors.

The current article studies classification errors for two purposes: 1) to quantify their effect on the accuracy of statistical figures, and 2) to show if and how we can use this information to improve the accuracy of the estimates by selective manual editing. The current article provides key extensions to Burger et al. (2015) on both topics. Concerning the first topic, we estimate the accuracy (due to classification errors) of published figures under more realistic conditions, rather than providing a sensitivity analysis as was done in Burger et al. (2015). Concerning the second topic, we experiment with selective editing aided by the estimated classification-error model.

The remainder of the article is organised as follows. Section 2 presents a theory to estimate accuracy and model classification errors. Section 3 introduces the case study. Results on the estimated accuracy are given in Section 4. Next, Section 5 estimates the effect of supplementary editing on the estimated accuracy. Finally, Section 6 discusses the results and gives suggestions for further research. The Appendix describes a theory for correcting the bias in the bootstrap estimates of accuracy.

2. Theory to Estimate Accuracy and Model Classification Errors

2.1. Estimating Accuracy for Given Classification Errors

Consider a population of units ($i = 1, \ldots, N$) that is divided into industries based on economic activity as derived in a business register (BR). Denote the total set of industries by $\mathcal{H}_{\text{full}}$. Each unit (enterprise) i has an unknown true industry code $s_i = g$ and an observed industry code $\hat{s}_i = h$, where $g, h \in \mathcal{H}_{\text{full}}$. We suppose that for each unit random classification errors occur, independently across units, according to a known (or previously estimated) transition matrix $\mathbf{P}_i = (p_{ghi})$, with $p_{ghi} = P(\hat{s}_i = h | s_i = g)$. Note that – following, for example, Kuha and Skinner (1997) – we consider the true industry code as fixed and the observed industry code as stochastic.

In this article, we consider the relatively simple case where classification errors are the only errors that affect the publication figures. We are interested in the total turnover per industry: $Y_h = \sum_{i=1}^{N} a_{hi} y_i$, with

$$a_{hi} = I(s_i = h) = \begin{cases} 1 & \text{if } s_i = h, \\ 0 & \text{if } s_i \neq h. \end{cases}$$

In practice, Y_h is estimated by $\hat{Y}_h = \sum_{i=1}^{N} \hat{a}_{hi} y_i$, with $\hat{a}_{hi} = I(\hat{s}_i = h)$. Now we would like to assess the bias and variance of $\hat{Y}_h$ as an estimator for Y_h, that is,

$$B(\hat{Y}_h) = E(\hat{Y}_h - Y_h) = \sum_{i=1}^{N} \{E(\hat{a}_{hi}) - a_{hi}\} y_i, \tag{1}$$

$$V(\hat{Y}_h) = \sum_{i=1}^{N} V(\hat{a}_{hi}) y_i^2, \tag{2}$$

where in (2) we used the assumption of independent classification errors across units.

Given the transition matrix $\mathbf{P}_i$, it is not too difficult to derive analytical expressions for the bias and variance of $\hat{Y}_h$ in the situation considered here (Appendix and Burger et al. 2015). Here, we focus on an alternative approach to estimate the accuracy and use bootstrap resampling. In future applications we would like to assess the bias and variance of estimates due to other nonsampling errors besides classification errors, such as measurement, linkage, and coverage errors, as well as combinations thereof (van Delden et al. 2014). The bootstrap method can be generalised to handle these more complex situations.

In the bootstrap approach, following Burger et al. (2015), we apply the transition matrix $\mathbf{P}_i$ to the observed $\hat{s}_i$, which results in a new industry-assignment variable, denoted by $\hat{s}_i^*$. That is to say, we consider realisations of the alternative classification-error model given by

$$P(\hat{s}_i^* = h | \hat{s}_i = g) \equiv P(\hat{s}_i = h | s_i = g) = p_{ghi}. \tag{3}$$

We also define: $\hat{a}_{hi}^* = I(\hat{s}_i^* = h)$. By repeating this procedure R times (for some large R), we obtain a set of so-called bootstrap replications of the estimated total turnover in industry h: $\hat{Y}_{hr}^{*} = \sum_{i=1}^{N} \hat{a}_{hir}^{*} y_i$ $(r = 1, \ldots, R)$. The bootstrap bias and variance are then estimated as follows (Efron and Tibshirani 1993):

$$\hat{B}_R^{*}\left(\hat{Y}_h\right) = m_R\left(\hat{Y}_h^{*}\right) - \hat{Y}_h, \tag{4}$$

$$\hat{V}_R^{*}\left(\hat{Y}_h\right) = \frac{1}{R-1}\sum_{r=1}^{R}\left\{\hat{Y}_{hr}^{*} - m_R\left(\hat{Y}_h^{*}\right)\right\}^2. \tag{5}$$

with $m_R\left(\hat{Y}_h^{*}\right) = \frac{1}{R}\sum_{r=1}^{R} \hat{Y}_{hr}^{*}$. Details about the assumptions and computations can be found in Burger et al. (2015).

In practice, the total number of industries in $\mathcal{H}_{\text{full}}$ is large – about 300 in the Netherlands – and often one will be interested only in the accuracy of turnover estimates for a limited subset of target industries, rather than for all industries at once. In the remainder of this article we use $\mathcal{H}$ to denote the set of target industries, for which we want to compute (4) and (5), and $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$ to denote the other industries.

2.2. *Modelling Classification Errors*

2.2.1. Introduction to Modelling Classification Errors

To apply the above bootstrap method, we first need to estimate the matrix of classification-error probabilities. For simplicity, Burger et al. (2015) introduced three assumptions for this that we want to relax here. First, they assumed that the subset of target industries forms a ‘closed’ population, with only misclassifications among this subset. In terms of Burger et al.’s case study of the car trade, they assumed misclassifications only among the nine underlying industries within the car trade but no misclassifications between the car trade and other types of industry. Secondly, they assumed that the probabilities of misclassification are the same for all units in all industries; that is, $\mathbf{P}_i = \mathbf{P}$ and all diagonal elements of $\mathbf{P}$ are equal. Thirdly, they assumed that misclassified units are distributed uniformly over the remaining industries; that is, all off-diagonal elements of $\mathbf{P}$ are also equal. In the current article we use a more realistic approach. We still assume random classification errors, but we now estimate the transition probabilities p_{ghi} by means of an audit sample.

Suppose that each unit in the population has a transition matrix $\mathbf{P}_i$ with elements p_{ghi} as in Table 1, where $g, h \in \{1, \ldots, H\}$ stands for the target set of industries $\mathcal{H}$ for which

Table 1. Transition probabilities (subscript i omitted).

	Observed industry				
True industry	1	2		H	$H+1$
1	p_{11}	p_{12}	$\cdots$	p_{1H}	$p_{1,H+1}$
2	p_{21}	p_{22}	$\cdots$	p_{2H}	$p_{2,H+1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
H	p_{H1}	p_{H2}	$\cdots$	p_{HH}	$p_{H,H+1}$
$H+1$	$p_{H+1,1}$	$p_{H+1,2}$	$\cdots$	$p_{H+1,H}$	$p_{H+1,H+1}$

we want to estimate the accuracy of the totals $\hat{Y}_h$, and industry $H+1$ represents the union of all industries outside that target set, that is, the union of all industries in $\mathcal{H}_{\text{full}} \setminus \mathcal{H}$. In our case (see Section 3), we are interested in estimating totals of $H = 9$ industries in the car trade; the other industries outside the car trade but within the total set of possible NACE codes are summarised as a tenth 'industry'.

To reduce the number of parameters to estimate, we split up the estimation of $\mathbf{P}_i$ into three parts: 1) the diagonal elements $\hat{p}_{ggi}$ with $g \in \{1, \ldots, H\}$, 2) the off-diagonal elements $\hat{p}_{ghi}$ ($g \neq h$ and $g, h \in \{1, \ldots, H\}$), and 3) the elements of row and column $H+1$. To begin with, we ignore the last row and column of the matrix and focus on the submatrix with $g, h \in \{1, \ldots, H\}$. We separate the estimation of the diagonal and nondiagonal elements as follows. Consider the contingency table of s_i and $\hat{s}_i$ in the population and let N_{gh} denote the stochastic number of units in cell (g, h). The corresponding expected value M_{gh} is given by

$$M_{gh} = \sum_{i=1}^{N} P(\hat{s}_i = h | s_i = g) \cdot I(s_i = g). \tag{6}$$

Denote the probability that unit i is classified correctly as $\pi_i = P(\hat{s}_i = g | s_i = g)$. The transition probabilities for $g \neq h$ are then given by:

$$\begin{aligned} P\left(\hat{s}_i = h | s_i = g\right) &= P\left(\hat{s}_i = h, \hat{s}_i \neq g | s_i = g\right) \\ &= P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right) \cdot P\left(\hat{s}_i \neq g | s_i = g\right) \\ &= P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right) \cdot (1 - \pi_i) \end{aligned} \tag{7}$$

where $P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right)$ is the conditional probability that unit i receives the code $\hat{s}_i = h$, given that this is a wrong code ($s_i = g \neq h$). From Equations (6) and (7) it follows that

$$\begin{aligned} M_{gg} &= \sum_{i=1}^{N} \pi_i I(s_i = g), \\ M_{gh} &= \sum_{i=1}^{N} (1 - \pi_i) P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right) I(s_i = g), \quad (g \neq h). \end{aligned} \tag{8}$$

We now introduce separate models for estimating the diagonal probabilities π_i and the conditional off-diagonal probabilities $P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right)$.

2.2.2. Modelling the Diagonal Probabilities

To estimate the diagonal elements of the $H \times H$ submatrix, we introduce the assumption that the probabilities π_i can be modelled by a logistic regression (McCullagh and Nelder 1989) on a number of independent variables. We estimate the parameters of the model by taking an audit sample of size $n \ll N$ from the population, for which both $\hat{s}_i$ and s_i are observed.

2.2.3. Modelling the Off-Diagonal Probabilities

Similarly to the diagonal probabilities, the off-diagonal probabilities might in reality also vary with i. However, the off-diagonal probabilities concern a large number of parameters and it would lead to a lack of degrees of freedom in the audit data if we also modelled those as a function of independent variables. To estimate the off-diagonal elements of the $H \times H$ submatrix, we therefore introduce the additional assumption that, given that a unit is misclassified, the conditional off-diagonal probabilities are independent of i:

$$P\left(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g\right) = \frac{P\left(\hat{s}_i = h | s_i = g\right)}{1 - \pi_i} \equiv \psi(g, h), \quad (g \neq h). \tag{9}$$

From (8) it now follows that

$$M_{gh} = \psi(g, h) \sum_{i=1}^{N} (1 - \pi_i) I(s_i = g) = \psi(g, h)\left(M_{g+} - M_{gg}\right), \quad (g \neq h) \tag{10}$$

where $M_{g+} = N_{g+} = \sum_{i=1}^{N} I(s_i = g)$ stands for a fixed but unknown row total. Hence we obtain:

$$\psi\,(g, h) = \frac{M_{gh}}{M_{g+} - M_{gg}}, \quad (g \neq h). \tag{11}$$

Note that, within each row, we have $\sum_{h \neq g} \psi(g, h) = 1$.

Now suppose that, in our audit sample, we count n_{gh} units in cell (g, h). In principle, we could estimate $\psi(g, h)$ by substituting these observed counts directly into Expression (11). However, this would yield unreliable estimates in practice, unless the audit sample was very large or H was very small. Therefore, we propose reducing the number of parameters further by using a log-linear model.

Denote: $m_{gh} = E(n_{gh})$. The information in the audit sample for the off-diagonal cells can be described completely by the following saturated log-linear model:

$$\log m_{gh} = u + u_{1(g)} + u_{2(h)} + u_{12(gh)}, \quad (g \neq h), \tag{12}$$

with the identifying restrictions $\sum_{g=1}^{H} u_{1(g)} = \sum_{h=1}^{H} u_{2(h)} = \sum_{g=1}^{H} u_{12(gh)} = \sum_{h=1}^{H} u_{12(gh)} = 0$. Log-linear models can be used to describe and test effects in contingency tables (Bishop et al. 1975).

Clerical reviewers know from their practical experience that some specific misclassifications of NACE codes occur more often than others. To reduce the number of parameters to estimate, we have asked experts to appoint each off-diagonal cell to a cluster $q \in \{1, \ldots, Q\}$, where cells within the same cluster are supposed to have a comparable probability of misclassification and Q is small compared to the total number of off-diagonal cells. Denote $\delta_q(g, h) \in \{0, 1\}$ as the variable indicating whether cell (g, h) is appointed to cluster q. Note that $\sum_{q=1}^{Q} \delta_q(g, h) = 1$ for all $g, h \in \{1, \ldots, H\}$ with $g \neq h$. Instead of the saturated model, we now use the following log-linear model:

$$\log m_{gh} = u + u_{2(h)} + \sum_{q=1}^{Q} \delta_q(g, h) u_{3(q)}, \quad (g \neq h), \tag{13}$$

using the identifying restrictions $\sum_{h=1}^{H} u_{2(h)} = \sum_{q=1}^{Q} u_{3(q)} = 0$. This model can be understood as follows. Firstly, the number of units may differ between industries, leading to different expected values m_{gh}. This is accounted for by the column effect $u_{2(h)}$ in the model. (We have a practical reason for taking the column effect rather than the row effect; see the end of this subsection.) In addition we account for the effect of the clusters $\delta_q(g,h)$. Similarly to the sparse classification-error model by Zhang (2005), the simplifying assumptions used to derive (9) and (13) aim to provide an adequate description of the effects of the classification errors, rather than the mechanisms by which these errors arise. Note that the diagonal probabilities are close to one in most cases (see Subsection 4.1), so the assumption is therefore adequate.

Model (13) has a slightly unusual form, but it can be rewritten as a standard log-linear model with only main effects by embedding the original contingency table in a three-dimensional table with cells (g,h,q), treating all cells for which $g=h$ or $\delta_q(g,h)=0$ as structural zeros. The parameters of Model (13) may then be estimated by maximum likelihood (Bishop et al. 1975), which gives the estimated values:

$$\hat{m}_{gh} = \exp\left\{\hat{u} + \hat{u}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h)\hat{u}_{3(q)}\right\}, \quad (g \neq h). \tag{14}$$

By substituting these values into (11), with $\hat{m}_{gg}=0$, we obtain estimates of the conditional probabilities $\psi(g,h) = \hat{m}_{gh}/\sum_{h=1}^{H}\hat{m}_{gh}(g \neq h)$.

In practice, it may be useful to draw the audit sample as a stratified sample by observed NACE code (i.e., stratified by column in the above contingency table). In that case, we need to take the sampling fractions into account when estimating the classification probabilities. Suppose that column h has a sampling fraction of n_{+h}/N_{+h}, with $n_{+h} = \sum_{g=1}^{H} n_{gh}$ and $N_{+h} = \sum_{g=1}^{H} N_{gh}$. We can estimate the population count in the cell (g,h) by $\hat{N}_{gh,model} = \hat{m}_{gh}\left(N_{+h}/n_{+h}\right)$. Multiplying the left- and right-hand sides of (14) by N_{+h}/n_{+h} yields

$$\hat{N}_{gh,model} = \exp\left\{\hat{v} + \hat{v}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h)\hat{v}_{3(q)}\right\}, \quad (g \neq h), \tag{15}$$

with $\hat{v} = \hat{u}$, $\hat{v}_{3(q)} = \hat{u}_{3(q)}$ and $\hat{v}_{2(h)} = \hat{u}_{2(h)} + \log N_{+h} - \log n_{+h}$. The conditional probabilities $\psi(g,h)$ are now estimated by

$$\hat{\psi}_{model}(g,h) = \frac{\hat{N}_{gh,model}}{\hat{N}_{g+,model}}, \quad (g \neq h), \tag{16}$$

where $\hat{N}_{g+,model} = \sum_{h=1}^{H}\hat{N}_{gh,model}$ and $\hat{N}_{gg,model} = 0$. Under the assumption that the transition probabilities are comparable per cluster, this yields an efficient and robust estimation of $\psi(g,h)$. Note in particular that $\hat{m}_{gh}$ (and thus $\hat{N}_{gh,model}$) can be positive even when $n_{gh}=0$.

2.2.4. Modelling the Probabilities in Industry $H+1$

Recall that the set of target industries $\{1, \ldots, H\}$ is only a small subset of all possible industry types in the BR. Estimating transition probabilities among all possible industry combinations within the BR from an audit sample is not realistic, as this would require an extension of the sample to all (several hundred) industries in the NACE domain. Instead we looked into the yearly updates of the NACE codes within the BR. Denote the observed industry of unit i in year t as $\hat{s}_i^t$. Some of the units switch between industries in year $t+1$ compared to year t: $\hat{s}_i^t = h$ and $\hat{s}_i^{t+1} = g$. We believe that there is at least some association between the (unknown) classification-error probabilities p_{ghi} and the temporal transition probabilities in the BR. The latter reflect natural changes in economic activity, and we know that administrative delays in implementing these changes are an important cause of classification errors in the BR.

Data on yearly updates showed that the distribution of temporal transitions within the BR varies considerably among the $h \in \{1, \ldots, H\}$ industries. From these data we concluded that it is not realistic to use a two-level model whereby we estimate high granular (say one-digit) NACE code transitions within the whole BR as the first level and transitions within the underlying (more detailed) industries as the second level. Instead, we used an alternative two-level model. In the first level we estimate the overall probabilities $p_{g,H+1}$ and $p_{H+1,h}$ (the last column and row of Table 1), and in the second level we model the transitions to specific industries within industry $H+1$.

For the first level, consider the row in Table 1 with the transition probabilities of units with true industry $H+1$ (outside the target set of industries) that are observed in industry $h \neq H+1$ (inside the target set of industries). Some of these units are observed in the audit sample, so these probabilities can be estimated simply by extending the log-linear model from the previous subsection to the last row. (We assume here that the off-diagonal cells in the last row and column can be appointed to one of the clusters $q \in \{1, \ldots, Q\}$ just like the other off-diagonal cells.) Next, we consider the column in Table 1 with the transition probabilities of units with true industry $h \neq H+1$ that are observed in industry $H+1$. This type of classification error cannot be observed in our audit sample. To obtain a result, we assume here that the total number of "missed units" in the true industries $\{1, \ldots, H\}$ is equal to the number of "wrong units" in the observed industries $\{1, \ldots, H\}$, that is, that $\sum_{g=1}^{H} N_{g,H+1} = \sum_{h=1}^{H} N_{H+1,h}$. Note that if this assumption does not hold, the size of the observed population in the industries $\{1, \ldots, H\}$ is structurally too high or too low.

Under the above assumption it should hold that

$$\hat{N}_{+,H+1,model} \equiv \sum_{g=1}^{H} \hat{N}_{g,H+1,model} = \sum_{h=1}^{H} \hat{N}_{H+1,h,model} \equiv \hat{N}_{H+1,+,model}. \qquad (17)$$

Using this assumption, we can extend Expression (15) to $h = H+1$, where the cluster parameters $\hat{v}_{3(q)}$ are estimated on the cells (g,h) where $h \in \{1, \ldots, H\}$. In fact, we cannot estimate the effect $\hat{v}_{2(H+1)}$ in (15) directly from the audit sample. However, taking the sum of (15) with $h = H+1$ over all cells in this column we obtain:

$$\sum_{g=1}^{H} \hat{N}_{g,H+1,model} = \exp\left\{\hat{v}_{2(H+1)}\right\} \sum_{g=1}^{H} \exp\left\{\hat{v} + \sum_{q=1}^{Q} \delta_q(g, H+1)\hat{v}_{3(q)}\right\} \qquad (18)$$

According to (17), the left-hand sum should be equal to $\hat{N}_{H+1,+,model}$, which is known after the estimation of the log-linear model, including row $H+1$. In that case $\hat{v} = \hat{u}$ and the cluster effects $\hat{v}_{3(q)} = \hat{u}_{3(q)}$ are also known. Hence, $\hat{v}_{2(H+1)}$ can be solved from Expression (18). Next, the underlying estimates $\hat{N}_{g,H+1,model}$ can be obtained from (15).

Finally, we can use all estimated counts $\hat{N}_{gh,model}$ to obtain estimates of $\hat{\psi}_{model}(g,h)$ as in (16). This completes the first level of the model for industry $H+1$.

2.2.5. Subdividing Units in $H+1$ Into Underlying Industries

The model from the previous subsection allows us to estimate $P(\hat{s}_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H} | s_i = h)$ and $P(\hat{s}_i = h | s_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H})$, with $h \in \mathcal{H}$. During bootstrap simulation, these probabilities refer to the events of, respectively, a unit moving from a given target industry to an unspecified industry outside the target set ("outflow of turnover") and vice versa ("inflow of turnover"). For the purpose of quantifying the accuracy of turnover estimates for our target set of industries, it is not necessary to model the "outflow of turnover" in more detail. We do need a more detailed model for the "inflow of turnover". We applied a second-level model in which:

- the transition probabilities $P(\hat{s}_i = h | s_i = g)$ with $h \in \mathcal{H}$ and $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ are proportional to the corresponding yearly transitions in the BR, that is, the transitions from $h \in \mathcal{H}$ at $t-1$ to $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ at time t; and
- the turnover of those units are drawn from a log-normal distribution. For the log-normal distribution we made a distinction between units with size class 0–3 and other units.

The exact procedure for drawing from the second-level model and estimating its parameters is given in van Delden et al. (2015a).

2.3. *Bias Correction*

Burger et al. (2015) explain that $\hat{B}^*_R(\hat{Y}_h)$ in (4) is a biased estimator of $B(\hat{Y}_h)$ in (1). This can be understood, since the bootstrap replications start from the observed $\hat{s}_i = h$ rather than the true $s_i = g$ values. In the more simple situation described in Burger et al. (2015) this bias could be corrected easily. In our case it is also possible to compute an unbiased bootstrap estimator of $B(\hat{Y}_h)$; see the Appendix. In terms of the notation in the Appendix, we denote the original (biased) estimator $\hat{B}^*_R(\hat{Y}_h)$ as $\hat{B}^*_{0R}(\hat{Y}_h)$ and the corrected (unbiased) estimator by $\hat{B}^*_{1R}(\hat{Y}_h)$. A disadvantage of $\hat{B}^*_{1R}(\hat{Y}_h)$ is that it may have a large variance in practice. We therefore introduce a combined estimator, denoted by $\hat{B}^*_{\lambda R}(\hat{Y}_h)$:

$$\hat{B}^*_{\lambda R}(\hat{Y}_h) = \lambda \hat{B}^*_{1R}(\hat{Y}_h) + (1-\lambda)\hat{B}^*_{0R}(\hat{Y}_h) \tag{19}$$

where the relative weight λ is determined by minimising the mean squared error of $\hat{B}^*_{\lambda R}(\hat{Y}_h)$. The exact procedure actually involves optimal weights at a more detailed level than indicated in (19); see Expression (25) in the Appendix. More details are given in van Delden et al. (2015a). The results of our case study in Section 4 and Section 5 below were obtained using this combined bootstrap estimator for the bias.

The bootstrap variance $\hat{V}^*_R(\hat{Y}_h)$ in (5) is also a biased estimator of $V(\hat{Y}_h)$ in (2), but this bias is expected to be small in practice compared to that of $\hat{B}^*_R(\hat{Y}_h)$ (cf. van Delden et al. 2015a for more details). Therefore we did not attempt to correct this bias in our case study; the results below were obtained using Estimator (5) for the variance.

Note that our bias correction is specifically derived for classification errors affecting a level estimate, so the approach cannot be applied directly to more difficult problems (considering, for example, a combination of classification errors and measurement errors). A more general strategy for bias correction might be based on a ‘double’ bootstrap method (Efron and Tibshirani 1993; Hall and Maiti 2006).

3. Case Study: Data

The case study concerns estimates of quarterly turnover levels in the industry car trade (NACE rev. 2 code 45) for the first quarter (Q1) of 2012 until Q2 of 2014. The outcomes of the car trade are subdivided into nine industries. The quarterly turnover is estimated from a mixed-source production system (e.g., van Delden and de Wolf 2013).

Turnover in the small enterprises is derived from value-added tax (VAT) data. These enterprises are referred to as the complexity class *simple units*. On 1 January 2013 there were about one million simple units in the Netherlands, of which 28,605 were classified as car traders. The remaining units are observed in a census survey. There were 8,403 such enterprises within the whole domain of economic activities and 239 within the car trade (1 January 2013). For a subset of this group, there is a special business unit at Statistics Netherlands (CBS) with centralised data collection and data editing. This concerned 2,305 enterprises within the whole domain of economic activities and 49 within the car trade. This latter subset is referred to as the complexity class *most complex units*. The other units receiving survey data but not treated by this special business unit are referred to as the *complex units.*

The quarterly outcomes are published in different releases: 30 days (flash), 60 days (early), 90 days (late) and one year (final) after the end of the reference period. The computations in the current article concern the most recent releases available. For 2012 and 2013 this concerns the final release and for Q1 and Q2 of 2014 this concerns the late release. The available microdata covered nearly the complete target population. In late releases, quarterly nonrespondents are missing, as are units that report their VAT on a yearly basis. The latter group corresponds to 2–3% of the total turnover. Missing values are imputed. In the final release, the imputed quarterly turnover values of units that report VAT on a yearly basis are calibrated upon their reported yearly turnover values. We treat imputations here as if they are observed values, that is, we do not compute the effect of the imputation process on the accuracy.

The nine industries within the car trade vary considerably in the number of enterprises, total turnover and turnover per enterprise (van Delden et al. 2015a). In the first quarter of 2013, total turnover varies from 7,749 million euros (code 45112) to 51 million euros (code 45194). The division of total turnover in the different complexity classes also varies considerably across the nine industry codes (see Figure 1; the more detailed probability classes will be explained shortly). Note that throughout the article the industry classes are ordered from the largest to the smallest total turnover per industry.

The parameters of the classification-error model were estimated using three sources:

- We took an audit sample from the population of the *simple* enterprises within the car trade that existed on 1 July 2014 according to our BR. We randomly sampled 25 enterprises from each of the nine industries. Next, the true NACE codes were determined by two experts, examining the Chamber of Commerce information and Internet data and contacting the enterprise in case of doubt.

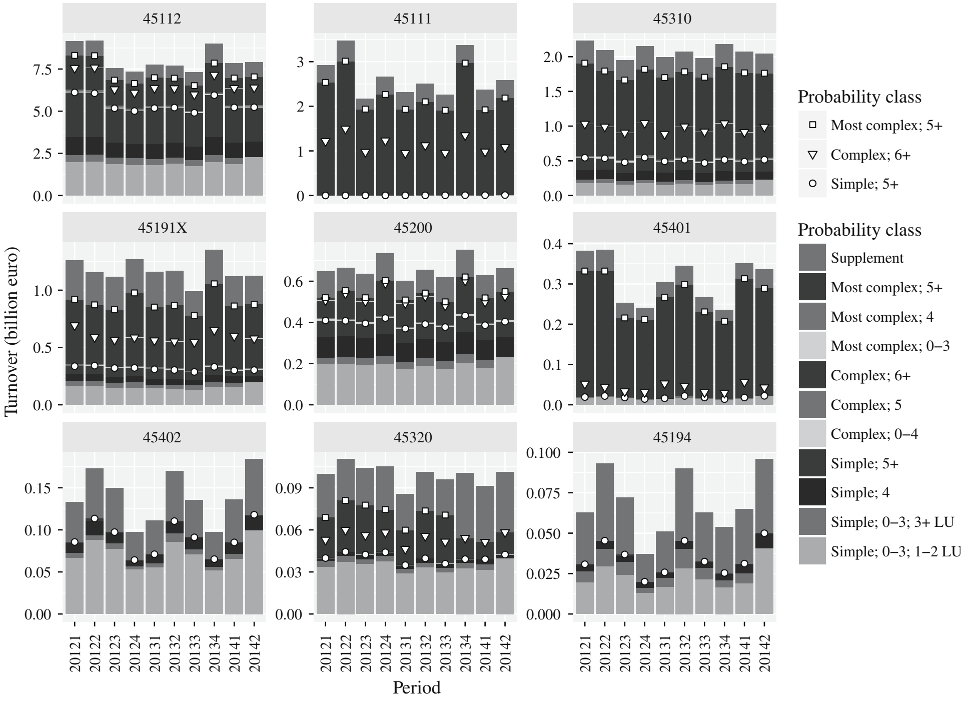


Fig. 1. Distribution of quarterly turnover among the different probability classes (Symbols appear at the upper side of the corresponding bar; Top bar is always the Supplement).

- For the *complex* and *most complex* enterprises we consulted experts at CBS who are responsible for the editing process of the car-trade industry and experts from a special business unit at CBS that deals with the large and complex units. We used expert knowledge for those enterprises, because quality studies reported in 2000 and 2003 that 97% of these enterprises were expected to have a correct three-digit NACE code (Burger et al. 2015). Therefore the transition probabilities for these units are close to 0 and 1, and estimating such small probabilities would have required a very large audit sample and too many resources. The experts were used to estimate the relative levels of classification error and the largest levels were set at five percent, which is in line with a Service Level Agreement that states that the three-digit NACE codes should be correct for 95% of the enterprises (Burger et al. 2015).
- In addition, we used data from our BR on the yearly transitions in NACE code of the enterprises for the years 2009–2014. From these data we computed the relative number of units that are observed in industry g in year t $(\hat{s}_i^t = g)$ given they are observed in h in year $t-1$ $\left(\hat{s}_i^{t-1} = h\right)$ averaged over 2009–2014. The motivation behind this approach was given in Subsubsection 2.2.4. Based on the results of the temporal transitions, we have asked experts to appoint each cell (g, h) to a cluster $q \in \{1, \ldots, Q\}$, where cells within the same cluster have a comparable probability of misclassification.

Details about how these sources were used to estimate the probabilities are given in the next section.

4. Results

4.1. Estimated Probabilities

Diagonal elements. Recall that for the diagonal elements of the $H \times H$ submatrix we try to explain differences in classification-error probabilities between units from their properties. Based on consultations with experts, we identified the following variables that are available for all units in the population and that might affect the level of classification-error probabilities: observed industry, number of legal units, legal form, size class of the enterprise, and being observed in a sample survey (yes/no).

The audit sample contained no classification errors among the simple enterprises with size class 4 or larger (ten working persons or more). We therefore used the audit sample only to estimate the diagonal probabilities for the simple enterprises with size classes 0–3 (0–9 working persons).

We investigated all possible combinations of the background variables using subset selection. To compare the performance of the models, we computed the AIC and deviance values (based on log-likelihood). Table 2 displays the best-fitting models with one, two, and three predictor variables. The fourth column gives the p value of a chi-square test of decrease in deviance (cf. McCullagh and Nelder 1989). Among the three best-fitting models, the model with industry and legal units led to a significant ($p = 0.04$) increase in model fit compared to a model with only industry, whereas adding additional terms did not significantly improve model fit despite a small decrease in AIC. We also verified the model selection results by cross validation (not shown). Taking all results into account, we selected the model with industry and legal units to estimate the diagonal probabilities for the remainder of this study.

The estimated probabilities are given in the bottom two rows of Figure 2. The numbers in the labels "0–3, 4, 5, 5+ , 6+ " stand for the size classes and 1–2 LU and 3+ LU stand for the number of legal units per enterprise. The diagonal probabilities of the upper nine rows of Figure 2 were based on experience of editing experts at CBS. Concerning the background variables affecting those probabilities, we limited ourselves to the complexity and size class of the units and supplementary editing (see below). From now on, the strata defined by these background variables, as shown in Figures 1 and 2, are referred to as probability classes.

The probability class 'supplement' in Figure 2 concerns the enterprises that are edited thoroughly by the statistical division at CBS responsible for the output. Enterprises that belong to the probability class 'supplement' have transition probabilities of 1.0 on the

Table 2. Three best-fitting logistic regression models for the audit sample, size classes 0–3. (Dev = Deviance; df = degrees of freedom).

	Model terms	Dev (df)	ΔDev (Δdf)	p value	AIC
0	NULL	257.27 (210)			259.27
1	Industry	170.94 (202)	86.34 (8)	<0.0001	188.94
2	Industry + Legal units	166.51 (201)	4.43 (1)	0.04	186.51
3	Industry + Legal units + Observed (Y/N)	164.36 (200)	2.15 (1)	0.14	186.36

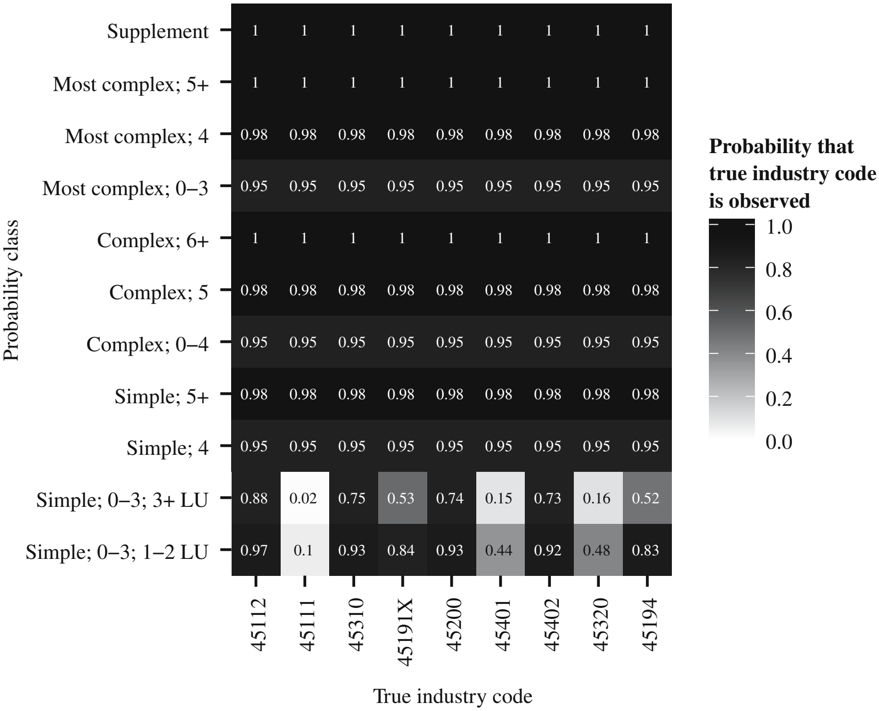
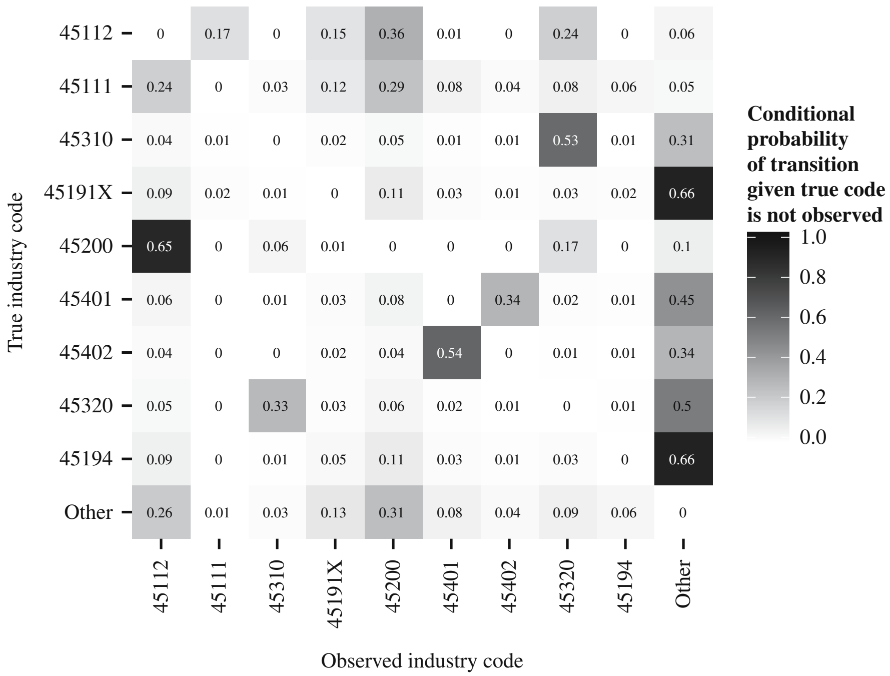


Fig. 2. Estimated transition probabilities for the diagonal elements.

main diagonal (first row of Figure 2), regardless of the further characteristics of the unit. For each target industry the size of this supplement was set to the 25 enterprises with the largest turnover, which approximately resembles the actual situation at the statistical division.

Off-diagonal elements. Using the average of the yearly transitions of the NACE codes over 2009–2014, experts appointed four clusters. Based on these $Q = 4$ clusters, we fitted a log-linear model to the off-diagonal numbers found in the audit sample, according to Equation (13). The model fitted well with a likelihood ratio of 85.92 with $p = 0.082$ at 69 degrees of freedom (df). The likelihood-ratio statistic compares the fit of the posited model to that of a saturated log-linear model, which reproduces the original table exactly (Bishop et al. 1975, 125); nonsignificant values indicate that all relevant factors are included in the model. There was one outlier that dominated the values for cluster 4. We therefore placed that outlying value in a separate fifth cluster. The model adjusted for this outlier had a likelihood ratio of 43.44 ($p = 0.991$ at 68 df). The adjusted model had expected numbers that fit the observed numbers in the audit sample better. Using those expected numbers and the sampling fractions n_{+h}/N_{+h}, the off-diagonal probabilities were estimated according to Equations (15) and (16) (Figure 3). Recall that the probabilities for the $(H+1)^{\text{th}}$ industry (column) were derived from Equation (17)–(18).

Results show that there are pairs of industries with relatively high conditional classification-error probabilities. For instance, a unit from industry 45310 (wholesale trade of motor vehicle parts and accessories) has a probability of 0.53 – given that it is



True industry code \ Observed industry code	45112	45111	45310	45191X	45200	45401	45402	45320	45194	Other
45112	0	0.17	0	0.15	0.36	0.01	0	0.24	0	0.06
45111	0.24	0	0.03	0.12	0.29	0.08	0.04	0.08	0.06	0.05
45310	0.04	0.01	0	0.02	0.05	0.01	0.01	0.53	0.01	0.31
45191X	0.09	0.02	0.01	0	0.11	0.03	0.01	0.03	0.02	0.66
45200	0.65	0	0.06	0.01	0	0	0	0.17	0	0.1
45401	0.06	0	0.01	0.03	0.08	0	0.34	0.02	0.01	0.45
45402	0.04	0	0	0.02	0.04	0.54	0	0.01	0.01	0.34
45320	0.05	0	0.33	0.03	0.06	0.02	0.01	0	0.01	0.5
45194	0.09	0	0.01	0.05	0.11	0.03	0.01	0.03	0	0.66
Other	0.26	0.01	0.03	0.13	0.31	0.08	0.04	0.09	0.06	0

True industry code
Observed industry code
Conditional probability of transition given true code is not observed
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 3. Estimated transition probabilities for the off-diagonal elements. Each row adds up to 1.

misclassified – of being observed as 45320 (retail trade of motor vehicle parts and accessories). Likewise, misclassified units from industry 45320 have a probability of 0.33 of being observed as 45310. Similar high conditional probabilities of misclassification exist between the industries 45401 (wholesale trade in maintenance and repair of motor cycles) and 45402 (retail trade in maintenance and repair of motor cycles). Finally, note that in six of the nine car trade industries, misclassified units have a probability over 0.30 of being observed outside the car-trade.

Probabilities for the industries outside the car trade. We applied the approach of Subsubsection 2.2.5 to estimate the parameters of the second-level model. Details can be found in van Delden et al. (2015a).

4.2. Simulation of Accuracy

Having modelled the probabilities of classification errors for the data in our case study, we applied the bootstrap method from Subsection 2.1. We applied 10,000 bootstrap replicates. We implemented this method within the R environment for statistical computing. The code used for these simulations is available from the authors upon request.

We summarised the results in terms of the following accuracy measures, derived from (4) and (5):

- the relative bias (RB) $\hat{B}_R^*(\hat{Y}_h)/\hat{Y}_h$,
- the coefficient of variation (CV) $\sqrt{\hat{V}_R^*(\hat{Y}_h)}/\hat{Y}_h$,
- the relative root mean squared error (RRMSE) $\sqrt{\left\{\left[\hat{B}_R^*(\hat{Y}_h)\right]^2+\hat{V}_R^*(\hat{Y}_h)\right\}}/\hat{Y}_h$.

These results are shown in Figure 4 (expressed as percentages). The RRMSE varies from about 1.0% for the industries 45401 and 45310 to about 60% for industry 45320. The variance (CV) dominates in the industries 45191X, 45401, 45402, and 45194, in the other industries the bias dominates. The industries 45112 and 45310 both have a negative bias. A negative bias means that the values of bootstrap simulations $\left(\hat{Y}^*_{hr}\right)$ are smaller on average than the estimated value ($\hat{Y}_h$), which in turn implies that $\hat{Y}_h$ underestimates the (unknown) true target value Y_h.

We found that industry 45320 has a very large RRMSE: on average 62% (Figure 4). This industry has a relatively large probability of classification error on the diagonal elements (Figure 2) of the complexity class "simple", and this class constitutes about one third of the total turnover in this industry (Figure 1). Industry 45111 has an even larger probability on classification errors in the complexity class "simple" (Figure 2) but does not have a large RRMSE. The latter is because the turnover of the simple enterprises in industry 45111 is very small compared to the other complexity classes (Figure 1). The RRMSE for the car trade as a whole is about 0.33% and was relatively stable over the ten periods (Figure 5). The CV was also relatively stable (about 0.29%). The RB varied most and ranged between −0.2% and −0.1%.

The RRMSE for the car trade as a whole is judged as acceptable by the owner of the production process, whereas that of industry 45320 is far too large. Fortunately, turnover levels for industry 45320 are not published separately but combined with industry 45310. The combined quarterly turnover-level estimates have an average RRMSE of 1.9% (see industry 45300 in Figure 2 by van Delden et al. 2015b). The least accurate industry that is

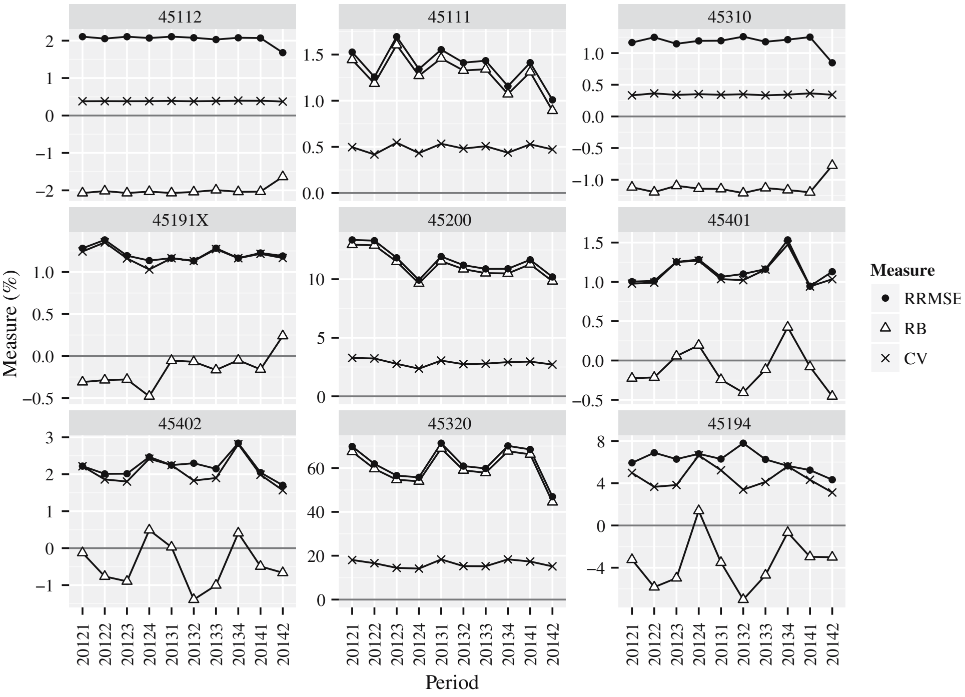


Fig. 4. RRMSE, RB and CV for ten periods per industry.

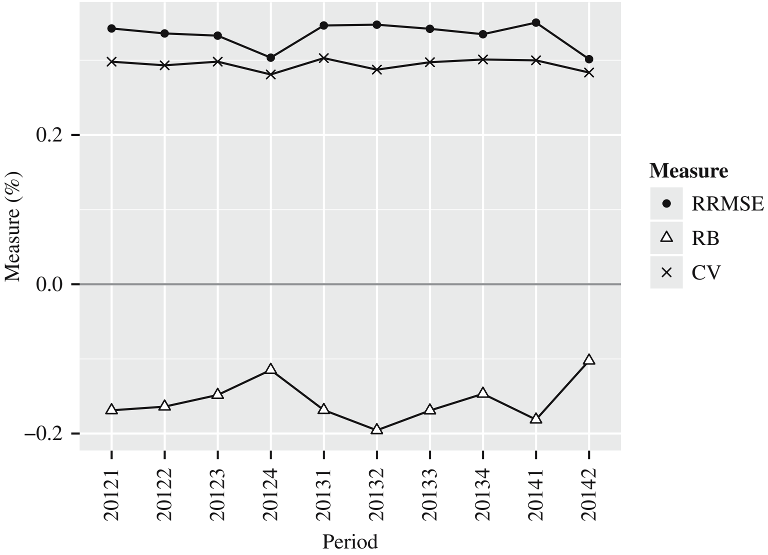


Fig. 5. RRMSE, RB and CV for ten periods for car trade as a whole.

actually published is 45200 with an RRMSE for the quarterly turnover slightly larger than ten percent. CBS aims to have a maximum uncertainty margin of three percent points on turnover levels. This means that in the car trade, an additional editing effort is needed to improve industry 45200. Below we investigate different selective editing strategies.

5. Editing Scenarios

5.1. Scenarios of Editing

We would also like to study to what extent the accuracy is improved when the editing effort is increased. An exact computation of those results is in fact only possible when we actually have a set of data that are free of classification errors. That information is needed because we need to know the true NACE code for each of the individual units. Since we do not have a data set for the whole population that is error free, we used an approximation. We assumed that with additional editing effort, those units that are checked and edited (on top of the starting situation) have a diagonal transition probability of 1, in other words a classification-error probability of zero. The edited units are called the “supplement” (Figure 1). They are called supplement because they are edited by the clerical reviewers of the production unit supplementary to the editing that is done by our central business unit on large and complex units. The exact difference between our approximation and the (true) effect of editing is explained in van Delden et al. (2015a). Nonetheless, we are convinced that our approximation is good enough to compare different editing scenarios in a qualitative way.

We compared four levels of supplementary editing, namely 0, 225, 450, and 675 edited enterprises in the car trade (relative editing effort 0, 1, 2, and 3). Since our results on accuracy were reasonably consistent over the ten quarters, we only computed the results for one quarter: the first quarter of 2013. The second level, 225 units, corresponds reasonably well to the actual situation at CBS. We distinguished between two editing scenarios that differ in how those enterprises are allocated over the nine industries:

1. Fixed: each industry is allocated an equal number of enterprises for supplementary editing. So the four levels are equal to 0, 25, 50, and 75 enterprises per industry.
2. Pro rata: the number of enterprises to be edited per industry $\left(n_h^E\right)$ is in proportion to the product of $RMSE\left(\hat{Y}_h\right) = \sqrt{\left\{\left[\hat{B}_R^*\left(\hat{Y}_h\right)\right]^2 + \hat{V}_R^*\left(\hat{Y}_h\right)\right\}}$ and the population size per industry (N_h):

$$n_h^E = \frac{RMSE\left(\hat{Y}_h\right)N_h}{\sum_{h=1}^{H} RMSE\left(\hat{Y}_h\right)N_h} n^E, \tag{20}$$

where n^E denotes the total number of units to be selected for supplementary editing. Note that Equation (20) resembles the so-called Neyman allocation of a survey sample over its underlying industries (e.g., Cochran 1977, 98–99). Because of this analogy, one might expect the accuracy of the estimated turnover for the car trade as a whole to improve more under the pro-rata scenario than under the fixed scenario. For the $RMSE\left(\hat{Y}_h\right)$ values in Equation (20) we used the bootstrap estimates when 25 enterprises per industry were edited. Within each industry h, we selected the n_h^E units with the largest quarterly turnover for editing.

5.2. *Simulation of Editing*

The change in the accuracy measures with increased relative editing effort and the two editing scenarios showed several interesting results (Figure 6). First of all, as expected, the CV decreased with increased relative editing effort. Moreover, the absolute value of the RB) often decreased with increased editing effort. However, there were also many examples of situations where this relative bias increased. A prominent example is industry 45401, where the absolute RB clearly increased between the relative editing effort 1 and 2 for the fixed scenario, and between the relative editing effort 2 and 3 for the pro-rata scenario. The overall effect of the change of CV and RB with increased editing effort is that the RRMSE does not always decrease with increased editing effort.

We can understand this surprising phenomenon by analysing the transition of units among the industries. To this end we distinguish between inflow and outflow of turnover in industry h. Inflow of turnover occurs when units that were originally observed in another industry enter industry h in bootstrap replication r. Outflow of turnover occurs when units move to another industry from industry h where they were observed. The bias of a turnover estimate for an industry is the net result of the effects of the turnover inflow and outflow. Accordingly, when there are no classification errors (as a result of perfect editing of the units in all existing industries), the inflow and outflow components are zero and there is no bias. Likewise, when the transition probabilities happen to be such that turnover inflow and outflow to industry h are perfectly balanced, there is also no bias. In van Delden et al. (2015a) we describe and quantify the observed bias (and variance) patterns in each of the industries as the net result of inflow and outflow. In some industries we found a reasonable

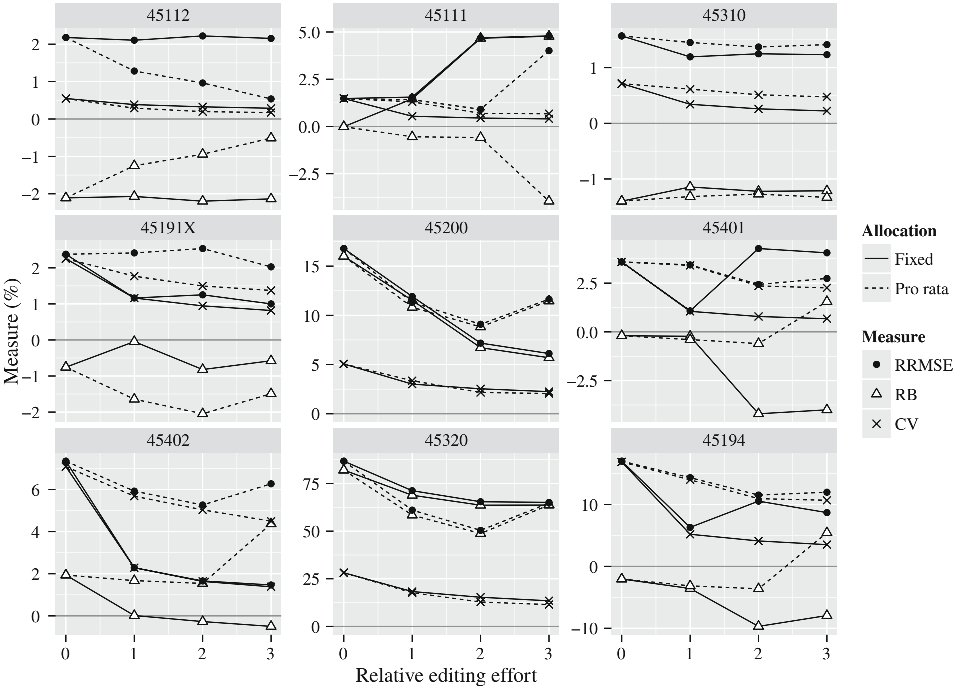


Fig. 6. Simulating the effect of editing on accuracy: the three measures (RRMSE, RB, and CV).

balance between inflow and outflow, while for others the total error is mainly determined by inflow (see Figure 4.4.5 in van Delden et al. 2015a). By changing the number of edited units in an industry, we can control the expected size of the outflow from that industry – that is, how many errors remain in industry h – but not the inflow. Due to this effect, the balance between outflow and inflow can become less favourable, leading to an increased bias.

In the above example, as the total amount of editing is increased, the absolute level of inflow in industry 45401 will decrease because the outflow from all the other car-trade industries will be reduced. Nonetheless, the balance in industry 45401 between out- and inflow on the bias becomes less favourable (van Delden et al. 2015a). In fact, with increased overall editing effort the turnover inflow in 45401 decreased at a smaller rate than the outflow, resulting in an increased bias. This effect is enhanced under the pro-rata scenario, because industry 45401 has the largest turnover inflow from industry 45402, which is more accurate than industry 45401 to begin with.

Figure 6 shows that in some industries the pro-rata scenario reduces the RRMSE further than the fixed scenario, while in other industries the opposite is the case. This is of course due to differences in editing effort per industry in the pro-rata scenario. Surprisingly, the decrease of the RRMSE for the car trade as a whole (sum of nine industries) is *larger* for the fixed scenario than for the scenario pro rata (Figure 7). This can be understood as follows. The pro-rata scenario, inspired by the Neyman allocation, assumes that the errors $\hat{Y}_h - Y_h$ are *independent* of each other. This assumption, however, does not hold in the

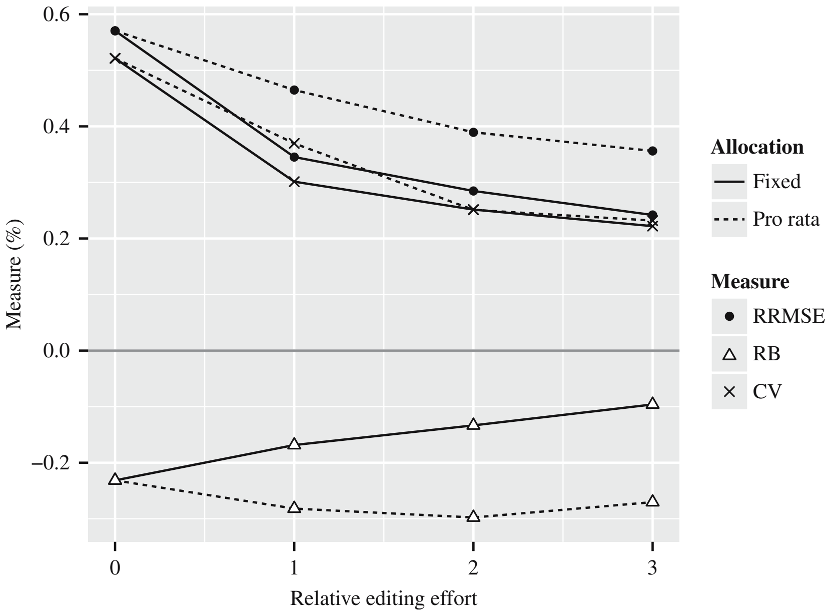


Fig. 7. Simulating the effect of editing on accuracy: overall effect on car trade.

case of classification errors, since many off-diagonal transition probabilities are larger than zero. We conclude that a simple ‘fixed’ scenario is more effective in reducing the overall RRMSE than the pro-rata approach. It remains to be seen whether a more efficient scenario than ‘fixed’ can be found, without introducing complexities that render the approach impractical.

6. Discussion

The long-term aim of our research is to develop a practical method for assessing as well as improving the accuracy of register-based estimates affected by nonsampling errors. In this article, we have estimated the accuracy of register-based outcomes for classification errors using a bootstrap method. Others have also used resampling to estimate the accuracy of statistical outcomes for certain error types, such as Zhang (2011), Lumme et al. (2015), and Chipperfield and Chambers (2015). A key challenge is to obtain good estimates of the parameters of the postulated error model.

How to handle the complex and most complex units in this respect is a difficult question. We have relied on expert knowledge when setting the diagonal probabilities of these units in our study. This is a relatively small group of units for which classification errors are rare. Furthermore, these units are not ‘mutually interchangeable’, given their large individual shares in the total turnover. Fundamentally, it may be asked whether a random classification-error model is appropriate for the group of complex and most complex units.

For the simple units where the model parameters can be estimated empirically, audit data can only be obtained at some additional cost. The question is then how to combine editing and estimation efficiently in practice. An option could be to use information obtained during regular production instead of audit data to estimate the model parameters, similarly to the use of paradata in social surveys (Kreuter 2013). Maybe we can combine selective editing for the most influential units with a probability sample of less influential units. The result of this two-phase design can be used to estimate the bias and the variance of the resulting estimator as a result of the editing process (e.g., Ilves and Laitila 2009). Such an approach might also offer the possibility of extending the procedure to other industries. The development of a robust and efficient selective editing strategy for classification errors, which accounts for the in- and outflow components of the target variable due to misclassified units, is a point for future research.

Two key extensions are still needed to achieve our long-term aim. These are the extension to other types of estimators and the extension to other sources of nonsampling errors. The use of an overarching modelling framework in which the observations reflect measurements of unobserved (true) values, like in latent class models and like the Bayesian approach by Bryant and Graham (2015), might be helpful in this respect.

Appendix

Bias Correction

Correction for the Bias in $\hat{B}^*_R\left(\hat{Y}_h\right)$

We use the notation that was introduced in Section 2. In addition, let $\boldsymbol{a}_i$ denote the vector $\left(a_{1i}, \ldots, a_{H+1,i}\right)^T$ that contains the values of the indicator variable $a_{hi} = I(s_i = h)$ of which one element per unit i is equal to 1. Similarly, we define $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{a}}_i^*$ on the basis of $\hat{s}_i$ and $\hat{s}_i^*$. Recall that $\mathbf{P}_i$ stands for the $(H+1) \times (H+1)$ matrix with the transition probabilities for unit i. Under the classification-error model, the expectation of $\hat{\boldsymbol{a}}_i$ for enterprise i equals $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}_i^T \boldsymbol{a}_i$. Similarly it holds that $E\left(\hat{\boldsymbol{a}}_i^* | \hat{\boldsymbol{a}}_i\right) = \mathbf{P}_i^T \hat{\boldsymbol{a}}_i$. Denote the vectors with the true, observed and bootstrap values for the total turnover per industry as $\boldsymbol{y} = \sum_{i=1}^N \boldsymbol{a}_i y_i$, $\hat{\boldsymbol{y}} = \sum_{i=1}^N \hat{\boldsymbol{a}}_i y_i$ and $\hat{\boldsymbol{y}}* = \sum_{i=1}^N \hat{\boldsymbol{a}}_i^* y_i$. Using an argument similar to that in Burger et al. (2015), the following expressions may be derived for the true bias and variance-covariance matrix of $\hat{\boldsymbol{y}}$ as an estimator for $\boldsymbol{y}$:

$$B(\hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}) - \boldsymbol{y} = \sum_{i=1}^N \left(\mathbf{P}_i^T - \mathbf{I}\right) \boldsymbol{a}_i y_i, \tag{21}$$

$$V(\hat{\boldsymbol{y}}) = \sum_{i=1}^N \left\{ \operatorname{diag}\left(\mathbf{P}_i^T \boldsymbol{a}_i y_i^2\right) - \mathbf{P}_i^T \operatorname{diag}\left(\boldsymbol{a}_i y_i^2\right) \mathbf{P}_i \right\}, \tag{22}$$

where $\mathbf{I}$ stands for the $(H+1) \times (H+1)$-identity matrix. Here, we use the assumption that only the observed classifications $\hat{\boldsymbol{a}}_i$ may be erroneous, while the other components of $\hat{\boldsymbol{y}}$ are fixed.

In the bootstrap approach, the above bias and variance are estimated by the conditional bias and variance of $\hat{\boldsymbol{y}}*$ as an estimator for $\hat{\boldsymbol{y}}$. Letting $R \rightarrow \infty$ in Expressions (4) and (5), we would obtain:

$$\hat{B}^*_\infty(\hat{\boldsymbol{y}}) = B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}} = \sum_{i=1}^N \left(\mathbf{P}_i^T - \mathbf{I}\right) \hat{\boldsymbol{a}}_i y_i, \tag{23}$$

$$\hat{V}^*_\infty(\hat{\boldsymbol{y}}) = V(\hat{\boldsymbol{y}}^*|\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \left\{\mathrm{diag}\left(\mathbf{P}_i^T\hat{\boldsymbol{a}}_i y_i^2\right) - \mathbf{P}_i^T\mathrm{diag}\left(\hat{\boldsymbol{a}}_i y_i^2\right)\mathbf{P}_i\right\}; \tag{24}$$

cf. Burger et al. (2015). In our case study, we did not use these analytical formulas directly. We preferred to use Monte Carlo simulation to have more flexibility in the modelling of classification errors, in particular for industry $H+1$. (Note that the sum in Expressions (23) and (24) is over all units in the BR, including all industries outside the target set.)

Focussing on the bias, we see that $E\left\{\hat{B}^*_\infty(\hat{\boldsymbol{y}})\right\} = \sum_{i=1}^N \mathbf{P}_i^T\left(\mathbf{P}_i^T - \mathbf{I}\right)\boldsymbol{a}_i y_i$. This implies that $\hat{B}^*_\infty(\hat{\boldsymbol{y}})$ is biased as an estimator for $B(\hat{\boldsymbol{y}})$; the same follows for $\hat{B}^*_R(\hat{\boldsymbol{y}})$ based on a finite number of replications.

Now assume that the matrix $\mathbf{P}_i^T$ can be inverted and denote its inverse as $\mathbf{Q}_i = \left(\mathbf{P}_i^T\right)^{-1}$. It follows directly that $\hat{\boldsymbol{b}}_i = \mathbf{Q}_i\hat{\boldsymbol{a}}_i$ is an unbiased estimator for $\boldsymbol{a}_i$:

$$E(\hat{\boldsymbol{b}}_i) = E(\mathbf{Q}_i\hat{\boldsymbol{a}}_i) = \mathbf{Q}_i E(\hat{\boldsymbol{a}}_i) = \mathbf{Q}_i\mathbf{P}_i^T\boldsymbol{a}_i = \boldsymbol{a}_i.$$

Similarly for $\hat{\boldsymbol{b}}_i^* = \mathbf{Q}_i\hat{\boldsymbol{a}}_i^*$ it holds that $E\left(\hat{\boldsymbol{b}}_i^*|\hat{\boldsymbol{b}}_i\right) = E\left(\hat{\boldsymbol{b}}_i^*|\hat{\boldsymbol{a}}_i\right) = \mathbf{Q}_i\mathbf{P}_i^T\hat{\boldsymbol{a}}_i = \hat{\boldsymbol{a}}_i$. Analogously to $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}*$, we can define the turnover-related vectors $\hat{z} = \sum_{i=1}^N \hat{\boldsymbol{b}}_i y_i$ and $\hat{z}* = \sum_{i=1}^N \hat{\boldsymbol{b}}_i^* y_i$. Now, consider the conditional bias of $\hat{z}*$ as an estimator for $\hat{z}$:

$$B(\hat{z}*|\hat{z}) = E(\hat{z}*|\hat{z}) - \hat{z} = \sum_{i=1}^{N}\left\{E\left(\hat{\boldsymbol{b}}_i^*|\hat{\boldsymbol{b}}_i\right) - \hat{\boldsymbol{b}}_i\right\}y_i = \sum_{i=1}^{N}(\hat{\boldsymbol{a}}_i - \hat{\boldsymbol{b}}_i)y_i.$$

It follows that $E\{B(\hat{z}*|\hat{z})\} = \sum_{i=1}^N \{E(\hat{\boldsymbol{a}}_i) - E(\hat{\boldsymbol{b}}_i)\}y_i = \sum_{i=1}^N \left(\mathbf{P}_i^T - \mathbf{I}\right)\boldsymbol{a}_i y_i = B(\hat{\boldsymbol{y}})$. Hence $B(\hat{z}*|\hat{z})$ is an unbiased estimator for the bias of $\hat{\boldsymbol{y}}$.

In our case study the population is divided into a limited number of probability classes (PCs) with the same transition matrix. We can exploit this to compute $\hat{z}$ and $\hat{z}*$ in an efficient manner. Divide the population into the PCs of units $U_1, \ldots, U_K$, where the transition matrix for the k^{th} PC is denoted by $\mathbf{P}_k$, with the corresponding inverse being $\mathbf{Q}_k = \left(\mathbf{P}_k^T\right)^{-1}$. Now $\hat{z}$ can be computed according to:

$$\hat{z} = \sum_{i=1}^{N}\hat{\boldsymbol{b}}_i y_i = \sum_{k=1}^{K}\sum_{i\in U_k}\hat{\boldsymbol{b}}_i y_i = \sum_{k=1}^{K}\mathbf{Q}_k\sum_{i\in U_k}\hat{\boldsymbol{a}}_i y_i = \sum_{k=1}^{K}\mathbf{Q}_k\hat{\boldsymbol{y}}_k \equiv \sum_{k=1}^{K}\hat{z}_k,$$

with $\hat{\boldsymbol{y}}_k = \sum_{i\in U_k}\hat{\boldsymbol{a}}_i y_i$ the vector of industry-turnover totals for the k^{th} PC. Analogously, $\hat{z}*$ can be computed as $\hat{z}* = \sum_{k=1}^K \hat{z}_k^* \equiv \sum_{k=1}^K \mathbf{Q}_k\hat{\boldsymbol{y}}_k^*$, with $\hat{\boldsymbol{y}}_k^* = \sum_{i\in U_k}\hat{\boldsymbol{a}}_i^* y_i$. Some other practical issues related to the computation of the bootstrap estimator and its bias correction are discussed in Appendix A2 of van Delden et al. (2015a).

Similarly to the bias, the bootstrap estimator of the variance is also biased. In section A4 van Delden et al. (2015a) derive a formula for this bias, explain how it can be corrected and argue that this bias is likely to be small. We therefore did not apply the bias correction for the variance.

Adjusted Bias Correction for "increased Accuracy"

The corrected bootstrap estimator for the bias $B(\hat{z}*|\hat{z})$ is unbiased, but may yield inaccurate estimates of $B(\hat{\boldsymbol{y}})$ in practice. Unbiased bootstrap estimation of $B(\hat{\boldsymbol{y}})$ may come at the cost of an increased variance, to such a degree that the bias correction is not an improvement in all cases. Results on simulated data (not shown here) suggest that the bias-corrected bootstrap estimator tends to be unstable when some of the probabilities of classification errors are large.

In fact, it turns out that when some of the diagonal probabilities in $\mathbf{P}_k$ are much smaller than 1, the so-called condition number $\operatorname{cond}\left(\mathbf{P}_k^T\right) = ||\mathbf{P}_k^T||\ ||\mathbf{Q}_k||$ can become much larger than 1. Here, the symbol $||.||$ denotes a matrix norm. Since $\hat{\boldsymbol{y}}_k^* = \mathbf{P}_k^T \hat{\boldsymbol{z}}_k^*$, it follows from a standard result in numerical analysis that

$$\left|\text{rel. change}\left(\hat{\boldsymbol{z}}_k^*\right)\right| \leq \operatorname{cond}\left(\mathbf{P}_k^T\right) \times \left|\text{rel. change}\left(\hat{\boldsymbol{y}}_k^*\right)\right|,$$

where rel. change(.) denotes a relative change in the value of its argument (e.g., Stoer and Bulirsch 2002, 211). Hence, when $\operatorname{cond}\left(\mathbf{P}_k^T\right)$ is large, a small uncertainty in the simulated values of $\hat{\boldsymbol{y}}_k^*$ can be propagated as a large uncertainty in the derived values of $\hat{\boldsymbol{z}}_k^*$. This provides a heuristic explanation for why the bias-corrected bootstrap estimator (based on $\hat{\boldsymbol{z}}_k^*$) can be less accurate than the original bootstrap estimator (based on $\hat{\boldsymbol{y}}_k^*$) in situations where some units have a relatively large probability of being misclassified.

In Appendix A3 of van Delden et al. (2015a) an alternative correction method is proposed that uses a combined estimator

$$\hat{\mathbf{B}}_{\lambda}^* = \sum_{k=1}^{K} \left\{\lambda_k \hat{\mathbf{B}}_{1k}^* + (1 - \lambda_k)\hat{\mathbf{B}}_{0k}^*\right\}, \tag{25}$$

where $\hat{\mathbf{B}}_{0k}^* = B\left(\hat{\boldsymbol{y}}_k^* | \hat{\boldsymbol{y}}_k\right)$ and $\hat{\mathbf{B}}_{1k}^* = B\left(\hat{\boldsymbol{z}}_k^* | \hat{\boldsymbol{z}}_k\right)$ denote the original and bias-corrected bootstrap estimators of the bias of $\hat{\boldsymbol{y}}_k$. It is shown there how the weights $\lambda_k \in [0, 1]$ can be obtained by minimising the mean square error of the estimated bias.

7. References

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/0471458740.

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bryant, J.R. and P. Graham. 2015. “A Bayesian Approach to Population Estimation with Administrative Data.” *Journal of Official Statistics* 31: 475–487. Doi: http://dx.doi.org/10.1515/JOS-2015-0028.

Burger, J., A. van Delden, and S. Scholtus. 2015. “Sensitivity of Mixed-Source Statistics to Classification Errors.” *Journal of Official Statistics* 31: 489–506. Doi: http://dx.doi.org/10.1515/jos-2015-0029.

Chipperfield, J. and R. Chambers. 2015. “Using the Bootstrap to Analyse Binary Data Obtained via Probabilistic Linkage.” *Journal of Official Statistics* 31: 397–414. Doi: http://dx.doi.org/10.1515/JOS-2015-0024.

Christensen, J.L. 2008. “Questioning the Precision of Statistical Classification of Industries.” Paper presented at the DRUID Conference on Entrepreneurship and Innovation, 17–20 June 2008, Copenhagen. Available at: http://www2.druid.dk/conferences/viewpaper.php?id=3419&cf=29 (accessed April 2016).

Cochran, W.G. 1977. *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons.

Delden, A. van, S. Scholtus, and J. Burger. 2015a. “Quantifying the Effect of Classification Errors on the Accuracy of Mixed-Source Statistics.” Discussion Paper 2015-10. Available at: https://www.researchgate.net/publication/281450992_Quantifying_the_effect_of_classification_errors_on_the_accuracy_of_mixed-source_statistics (accessed April 2016).

Delden, A. van, S. Scholtus, and J. Burger. 2015b. “Effect of Classification Errors on the Accuracy of Business Statistics.” Paper presented at the European Establishment Statistics Workshop, 7–9 September 2015, Poznan. Available at: https://enbes.wikispaces.com/file/view/Effect+of+classification+errors+on+the+accuracy+of+business+statistics+Arnout+van+Delden,+Sander+Scholtus+and+Joep+Burger.pdf (accessed April 2016).

Delden, A. van and P.P. de Wolf. 2013. “A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data.” Paper presented at the New Techniques and Technologies for Statistics (NTTS) conference, 5–7 March 2013, Brussels. Available at http://ec.europa.eu/eurostat/cros/sites/crosportal/files/NTTS2013%20Proceedings_0.pdf (accessed April 2016).

Delden, A. van, S. Scholtus, P.P. de Wolf, and J. Pannekoek. 2014. “Methods to Assess the Quality of Mixed-Source Estimates.” Internal report PPM-2014-09-26-ADLN-SSHS-PWOF-JPNK, Statistics Netherlands, The Hague.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC. Doi: http://dx.doi.org/10.1007/978-1-4899-4541-9.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, 2nd ed. New York: John Wiley and Sons.

Hall, P. and T. Maiti. 2006. “On Parametric Bootstrap Methods for Small Area Prediction.” *Journal of the Royal Statistical Society B* 68: 221–238.

Ilves, M. and T. Laitila. 2009. “Probability-Sampling Approach to Editing.” *Austrian Journal of Statistics* 38: 171–182. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00541.x.

Kreuter, F., ed. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/9781118596869.

Kuha, J. and C. Skinner. 1997. “Categorical Data Analysis and Misclassification.” In *Survey Measurement and Process Quality*, edited by L.E. Lyberg, P.P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 633–670. New York: John Wiley and Sons. Doi: http://dx.doi.org/10.1002/9781118490013.

Lumme, S., R. Sund, A.H. Leyland, and I. Keskmäki. 2015. “A Monte Carlo Method to Estimate the Confidence Intervals for the Concentration Index Using Aggregated Population Register Data.” *Health Services and Outcomes Research Methodology* 15: 82–98. Doi: http://dx.doi.org/10.1007/s10742-015-0137-1.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall. Doi: http://dx.doi.org/10.1007/978-1-4899-3242-6.

Stoer, J. and R. Bulirsch. 2002. *Introduction to Numerical Analysis*, 3rd ed. New York: Springer. Doi: http://dx.doi.org/10.1007/978-0-387-21738-3.

Waal, T. de, J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley and Sons.

Zhang, L.-C. 2005. “On the Bias in Gross Labour Flow Estimates Due to Nonresponse and Misclassification.” *Journal of Official Statistics* 21: 591–604.

Zhang, L.-C. 2011. “A Unit-Error Theory for Register-Based Household Statistics.” *Journal of Official Statistics* 27: 415–432.

Zhang, L.-C. 2012a. “Topics of Statistical Theory for Register-Based Statistics and Data Integration.” *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. 2012b. “On the Accuracy of Register-Based Census Employment Statistics.” Paper presented at the European Conference on Quality in Official Statistics, May 30–June 1 2012, Athens. Available at: http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (accessed April 2016).

Received October 2015
Revised April 2016
Accepted May 2016