Statistics Netherlands

Discussion Paper

Quantifying the effect of classification errors on the accuracy of mixed-source statistics

2015 | 10

Arnout van Delden
Sander Scholtus
Joep Burger

September

Content

Publications in official statistics are more and more 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 non-sampling errors on the accuracy of mixed-source statistics is crucial for decision making, yet not straightforward. Here we simulate the effect of classification errors on the accuracy of turnover 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 due to 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 pre-set 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 to 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.

1. Introduction

Publications in official statistics are more and more often based on a combination of sources, for instance a sample survey combined with an administrative source (Zhang, 2014). This trend is triggered by the need of national statistical institutes (NSIs) to reduce the response burden of households and enterprises and to reduce data collection costs. The combination of data sources sometimes results in the situation that 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). A useful distinction in this context is that between errors in representation (i.e., concerning the definition and demarcation of the population of units) and errors in measurement (i.e., concerning the values of the variables).

We believe that it is important to quantify the implications of those errors on the accuracy of statistical outcomes based on mixed sources. First of all, because the statistical outcomes are used by policy makers, private institutions and the general public for decision-making: it is important to have accurate estimates or at least to have information about the statistical quality. Second, tactical and operational decisions by NSIs are often based on quality information. To give an example of a tactical decision: suppose one would like to decide between two estimation strategies for using administrative data in economic statistics based on tax units. In the first strategy, the values of all administrative units are aggregated directly and the results are then stratified by the economic activity variable as found in the administrative data. In the second strategy, the administrative data are linked first to a central business register (BR) and then stratified using the economic activity classification of the BR. Both strategies come with their own error types; for instance, in the first case the quality of the administrative economic activity may be relatively low and in the second case there may be linkage errors. Hence it is not obvious which approach leads to the most accurate estimates.

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 both micro- and macro-selection approaches for ‘selective editing’ have been developed (de Waal et al., 2011). Macro-selection uses a top-down approach where aggregates are checked and if needed more detailed values are checked. In the micro-selection process the manual editing is limited with the aid of a score function (e.g., Latouche and Berthelot, 1992). This score function combines the risk of an error with its expected influence, 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 non-sampling errors on the accuracy of estimates in practical situations is not yet very straightforward. Depending on the complexity of the combined data sources and the type of non-sampling 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 (2013) proposed to estimate the uncertainty caused by non-sampling 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. In the current paper, we proceed with this work towards a more realistic modelling of the error structure where we use a bootstrap approach, which can also be applied in more complex situations where an analytical solution cannot be found.

To illustrate the method, we look at a case study on the estimation of quarterly turnover in economic statistics based on a combination of a survey and administrative data. The figures are classified by economic activity (according to NACE rev. 2.0) into so-called 'base cells'. A base cell is the smallest building block from which all publication cells can be composed. Determining the correct activity code of economic units is often rather difficult and prone to errors, see for instance Christensen (2008). Reasons for this include: that the units that are surveyed 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 paper aims to quantify the effect of classification errors on the accuracy of the statistical figures under more realistic conditions. In the remainder of the paper the base cells are indicated by the general term 'industry code'.

In any attempt to estimate the accuracy of publication figures as a function of non-sampling errors, not only the technical part of the accuracy estimation *per se* needs attention; finding and estimating an appropriate model for the errors is also crucial. [This may be contrasted with the sampling error (variance) under traditional probability sampling, which is usually estimated directly from the survey data.] It is important to estimate this error model in a cost-effective way, since that is a precondition to achieve practical implementation in statistical production systems. One option is to use an audit sample and try to achieve a gold standard of true values for the set of sampled units, see for instance Zhang (2011). Next, already available data on the same variables in different sources or from different periods could be used in combination with underpinned assumptions. Finally, there are also methods to make use of expert believes and judgements, see for instance Berka et al. (2012) for the use of fuzzy logic and Wiśniowski et al. (2012) on the Delphi framework.

In the present paper we aim to compute the accuracy in a practical case study for the turnover levels in the industry 'car trade' as a function of estimated classification

errors that occur in the BR of Statistics Netherlands. In addition, we want to show how we can use this information to support the editing process at the NSI by studying the extent to which manual selective editing at the micro level can improve the accuracy of the estimates.

The remainder of the paper 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 ($i = 1, \dots, N$) that is divided into industries based on economic activity as derived in a BR. Denote the total set of industries by $\mathcal{H}^*$. Each unit (enterprise) i has an unknown true industry code $s_i = g$ and an observed industry code $\hat{s}_i = h$, where $g, h \in \mathcal{H}^*$. We suppose that random classification errors occur, independently for each unit, 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, e.g., Kuha and Skinner (1997), we consider the true industry code as fixed and the observed industry code as stochastic.

In this paper, 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

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

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, i.e.,

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

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

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

In the situation considered here, likewise to Burger et al. (2015), it is not too difficult to derive analytical expressions for the bias and variance; see Appendix A. 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 non-sampling 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.

Given the transition matrix $\mathbf{P}_i$, the bias and the variance of the turnover level estimates per industry, $\hat{Y}_h$, can be estimated by a bootstrap resampling approach. As described in Burger et al. (2015), in the bootstrap approach, 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: $\hat{Y}_{hr}^* = \sum_{i=1}^{N} \hat{a}_{hir}^* y_i$ $(r = 1, \dots, R)$. The bootstrap bias and variance are then estimated as follows (Efron and Tibshirani, 1993):

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

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

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

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

To explain some of the results found using the bootstrap method, it will be useful to separate the outcomes of the bias and variance per target industry into three 'transition classes'. Let $\hat{Y}_{gh}$ denote the turnover of units with true industry g that were observed in industry h, and let $\hat{Y}_{ghr}^*$ denote its analogue for the r^{th} bootstrap replication. We now define the following transition classes:

- **Inflow from the target set**, with turnover $\hat{Y}_h^{(IT)} = \sum_{\substack{g\in\mathcal{H} \\ g\neq h}} \hat{Y}_{gh}$; its bootstrap analogue is $\hat{Y}_{hr}^{*(IT)} = \sum_{\substack{g\in\mathcal{H} \\ g\neq h}} \hat{Y}_{ghr}^*$, the total turnover of units that enter industry h in bootstrap replication r $(\hat{s}_{ir}^* = h)$ but were originally observed in another industry within the target set.

– **Inflow from other activities**, with turnover $\hat{Y}_h^{(IO)} = \sum_{\substack{g \in \mathcal{H}^* \setminus \mathcal{H} \\ g \neq h}} \hat{Y}_{gh}$ and bootstrap analogue $\hat{Y}_{hr}^{*(IO)} = \sum_{\substack{g \in \mathcal{H}^* \setminus \mathcal{H} \\ g \neq h}} \hat{Y}_{ghr}^*$, the total turnover of units that enter industry h in bootstrap replication r but were originally observed in another industry outside the target set.
– **Outflow**, with turnover $\hat{Y}_h^{(O)} = -\left(Y_h - \hat{Y}_{hh}\right) = \hat{Y}_{hh} - Y_h$ and bootstrap analogue $\hat{Y}_{hr}^{*(O)} = \hat{Y}_{hhr}^* - \hat{Y}_h$, the negative turnover of units that move from industry h where they were observed to another industry in bootstrap replication r.

With these definitions, it clearly holds that

$$\hat{Y}_{hr}^* - \hat{Y}_h = \hat{Y}_{hr}^{*(IT)} + \hat{Y}_{hr}^{*(IO)} + \hat{Y}_{hr}^{*(O)}. \tag{6}$$

Hence, in obvious notation, we can decompose the total estimated bias for industry h as follows:

$$\hat{B}_R^*\left(\hat{Y}_h\right) = m_R\left(\hat{Y}_h^{*(IT)}\right) + m_R\left(\hat{Y}_h^{*(IO)}\right) + m_R\left(\hat{Y}_h^{*(O)}\right). \tag{7}$$

Similarly, noting that $V\left(\hat{Y}_h\right) = V\left(\hat{Y}_h^{(IT)} + \hat{Y}_h^{(IO)} + \hat{Y}_h^{(O)}\right)$, we can analyse the contribution of each transition class to the variance of $\hat{Y}_h$ by estimating the squared correlation between the estimate $\hat{Y}_{hr}^*$ and each component over the bootstrap replications:

$$\begin{aligned} \hat{R}_R^{*2}\left(\hat{Y}_h, \hat{Y}_h^{(j)}\right) &= \frac{\hat{C}_R^{*2}\left(\hat{Y}_h, \hat{Y}_h^{(j)}\right)}{\hat{V}_R^*\left(\hat{Y}_h\right)\hat{V}_R^*\left(\hat{Y}_h^{(j)}\right)} \\ &\equiv \frac{\left\{\frac{1}{R-1}\sum_{r=1}^R \left(\hat{Y}_{hr}^* - m_R\left(\hat{Y}_h^*\right)\right)\left(\hat{Y}_{hr}^{*(j)} - m_R\left(\hat{Y}_h^{*(j)}\right)\right)\right\}^2}{\frac{1}{R-1}\sum_{r=1}^R \left(\hat{Y}_{hr}^* - m_R\left(\hat{Y}_h^*\right)\right)^2 \frac{1}{R-1}\sum_{r=1}^R \left(\hat{Y}_{hr}^{*(j)} - m_R\left(\hat{Y}_h^{*(j)}\right)\right)^2} \end{aligned} \tag{8}$$

where $j \in \{IT, IO, O\}$. The usefulness of this separation into transition classes will be illustrated in Section 4.4.

2.2 Modelling classification errors

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's case study of car trade, they assumed only misclassifications among the nine underlying industries within car trade but no misclassifications between car trade and other types of industry. Secondly, they assumed that the probabilities of misclassification are the same for all units in all industries; i.e., $\mathbf{P}_i = \mathbf{P}$ and all diagonal elements of $\mathbf{P}$ are equal. Thirdly, they assumed that misclassified units are distributed uniformly over the remaining industries; i.e., all off-diagonal elements of $\mathbf{P}$ are also equal. In the current paper we use a more realistic approach. We still suppose 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 Figure 2.2.1, where $g, h \in \{1, \dots, H\}$ stands for the target set of industries $\mathcal{H}$ for which 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, i.e. the union of all industries in $\mathcal{H}^* \backslash \mathcal{H}$. In our case (see Section 3), we are interested to estimate totals of $H = 9$ industries in car trade; the other industries outside car trade, but within the total set of possible NACE codes, are summarised as a 10th 'industry'.

2.2.1 Transition probabilities (subscript *i* omitted)

$$\begin{array}{|cccc|c|} \hline p_{11} & p_{12} & \cdots & p_{1H} & p_{1,H+1} \\ p_{21} & p_{22} & \cdots & p_{2H} & p_{2,H+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{H1} & p_{H2} & \cdots & p_{HH} & p_{H,H+1} \\ \hline p_{H+1,1} & p_{H+1,2} & \cdots & p_{H+1,H} & p_{H+1,H+1} \\ \hline \end{array}$$

To reduce the number of parameters to estimate, we split up the estimation of $\mathbf{P}_i$ into three parts: (a) the diagonal elements $\hat{p}_{ggi}$ with $g \in \{1, \dots, H\}$, (b) the off-diagonal elements $\hat{p}_{ghi}$ ($g \neq h$ and $g, h \in \{1, \dots, H\}$), and (c) 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 $g, h \in \{1, \dots, H\}$. We separate the estimation of the diagonal and non-diagonal 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

$$M_{gh} = \sum_{i=1}^{N} P(\hat{s}_i = h | s_i = g) \cdot I(s_i = g). \tag{9}$$

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:

$$\begin{aligned} P(\hat{s}_i = h | s_i = g) &= P(\hat{s}_i = h, \hat{s}_i \neq g | s_i = g) \\ &= P(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g) \cdot P(\hat{s}_i \neq g | s_i = g) \\ &= P(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g) \cdot (1 - \pi_i) \end{aligned} \tag{10}$$

where $P(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g)$ 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 (9) and (10) it follows that

$$\begin{aligned} M_{gg} &= \sum_{i=1}^{N} \pi_i I(s_i = g), \\ M_{gh} &= \sum_{i=1}^{N} (1 - \pi_i) P(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g) I(s_i = g), \quad (g \neq h). \end{aligned} \tag{11}$$

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

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) according to:

$$E\left(\log\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{l=1}^{L}\beta_l x_{li} \tag{12}$$

where $(x_{1i}, \ldots, x_{Li})'$ is a vector of independent variables that are available for all units in the population. Now suppose we have taken an audit sample of size $n \ll N$ from the population, for which both $\hat{s}_i$ and s_i are observed. We can use this audit sample to obtain estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_L$ of the parameters in model (12). Next, we can estimate π_i for all units in the population by $\hat{\pi}_i = \left(1 + \exp\left\{-(\hat{\beta}_0 + \sum_{l=1}^{L}\hat{\beta}_l x_{li})\right\}\right)^{-1}$.

Modelling the off-diagonal probabilities

To estimate the off-diagonal elements of the $H \times H$ submatrix, we introduce the additional assumption that, given that a unit is misclassified, the off-diagonal probabilities are independent of i:

$$P(\hat{s}_i = h | s_i = g, \hat{s}_i \neq g) = \frac{P(\hat{s}_i = h | s_i = g)}{1-\pi_i} \equiv \psi(g,h), \quad (g \neq h). \tag{13}$$

From (11) it now follows that

$$M_{gh} = \psi(g,h)\sum_{i=1}^{N}(1-\pi_i)I(s_i = g) = \psi(g,h)\left(M_{g+} - M_{gg}\right), \quad (g \neq h), \tag{14}$$

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

$$\psi(g,h) = \frac{M_{gh}}{M_{g+} - M_{gg}}, \quad (g \neq h). \tag{15}$$

Notice 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 (15). However, this would yield unreliable estimates in practice, unless the audit sample was very large or H was very small. Therefore, we propose to reduce 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 completely described by the following saturated log-linear model:

$$\log m_{gh} = u + u_{1(g)} + u_{2(h)} + u_{12(gh)}, \quad (g \neq h), \tag{16}$$

with the identifying restrictions $\sum_{g=1}^{H} u_{1(g)} = \sum_{h=1}^{H} u_{2(h)} = \sum_{g=1}^{H} u_{12(gh)} = \sum_{h=1}^{H} u_{12(gh)} = 0$; see Bishop, Fienberg, and Holland (1975) for more explanation on log-linear models.

From their practical experience, clerical reviewers know 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. Let the indicator $\delta_q(g,h) \in \{0,1\}$ denote whether cell (g,h) is appointed to cluster q, with $\sum_{q=1}^{Q} \delta_q(g,h) = 1$ 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:

$$\log m_{gh} = u + u_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) u_{3(q)}, \quad (g \neq h), \tag{17}$$

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 be different for each industry, 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 formula (19) below.] In addition we account for the effect of the clusters $\delta_q(g,h)$.

Model (17) 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 (17) may then be estimated by maximum likelihood (see Bishop, Fienberg, and Holland, 1975), which gives the estimated values:

$$\hat{m}_{gh} = \exp\left\{\hat{u} + \hat{u}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) \hat{u}_{3(q)}\right\}, \quad (g \neq h). \tag{18}$$

By substituting these values into (15), with $\hat{m}_{gg} = 0$, we obtain estimates of the conditional probabilities, $\hat{\psi}(g,h) = \hat{m}_{gh} / \sum_{h=1}^{H} \hat{m}_{gh}$ $(g \neq h)$.

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 $\widehat{N}_{gh,model} = \hat{m}_{gh}(N_{+h}/n_{+h})$. Multiplying the left- and right-hand-sides of (18) by N_{+h}/n_{+h} yields

$$\widehat{N}_{gh,model} = \exp\left\{\hat{v} + \hat{v}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) \hat{v}_{3(q)}\right\}, \quad (g \neq h), \tag{19}$$

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

$$\hat{\psi}_{model}(g,h) = \frac{\widehat{N}_{gh,model}}{\widehat{N}_{g+,model}}, \quad (g \neq h), \tag{20}$$

where $\widehat{N}_{g+,model} = \sum_{h=1}^{H} \widehat{N}_{gh,model}$ and $\widehat{N}_{gg,model} = 0$.

Under the assumption that the transition probabilities are comparable per cluster, this yields an efficient and robust estimation of $\psi(g,h)$. Notice in particular that $\hat{m}_{gh}$ (and thus $\widehat{N}_{gh,model}$) can be positive even when $n_{gh} = 0$. Further notice that two cells, say (g,h) and (j,h) $(g \neq j)$, of the same column h and within the same cluster q will have an identical estimated number $\widehat{N}_{gh,model} = \widehat{N}_{jh,model}$ but their transition

probabilities will differ when their denominators differ ($\widehat{N}_{g+,model} \neq \widehat{N}_{j+,model}$), see equation (20).

Modelling the probabilities in industry $H+1$

Recall that the set of target industries $\{1, \dots, H\}$ 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) base cells 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 correlation 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.

It is not feasible to estimate separate transition probabilities between all pairs of target and non-target industries. Therefore, a simplified model is needed. One option could be to use a two-level model whereby we estimate high granular (say 1-digit) NACE code transitions within the whole BR as the first level and transitions within the underlying (more detailed) industries as the second level. However, data on yearly updates of the BR at Statistics Netherlands showed that this was not a realistic model, because transitions between industries with different 1-digit NACE codes occurred frequently. Instead, we used an alternative two-level model. In the first level we estimated the overall probabilities $p_{g,H+1}$ and $p_{H+1,h}$ (the last column and row of Figure 2.2.1) and in the second level we modelled the transitions to specific industries within industry $H+1$. In fact, the observed distribution of those temporal transitions within the BR varied considerably among the $h \in \{1, \dots, H\}$ industries.

For the first level, consider the last row in Figure 2.2.1 with the transition probabilities of units that are observed in industry $h \neq H+1$ but with true industry $H+1$ (outside 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 $q \in \{1, \dots, Q\}$ just like the other off-diagonal cells.)

Next, we consider units that are observed in industry $H+1$ but with true industry $h \neq H+1$. 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 $\{1, \dots, H\}$ is equal to the number of "wrong units" in the observed industries $\{1, \dots, H\}$, i.e., 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 $\{1, \dots, H\}$ would be structurally too high or too low.

Under the above assumption it should hold that

$$\widehat{N}_{+,H+1,model} \equiv \sum_{g=1}^{H} \widehat{N}_{g,H+1,model} = \sum_{h=1}^{H} \widehat{N}_{H+1,h,model} \equiv \widehat{N}_{H+1,+,model}. \qquad (21)$$

If we extend the log-linear model of the previous subsection also to the last column, the resulting estimates $\widehat{N}_{g,H+1,model}$ are given by

$$\widehat{N}_{g,H+1,model} = \exp\left\{\hat{v} + \hat{v}_{2(H+1)} + \sum_{q=1}^{Q} \delta_q(g, H+1)\hat{v}_{3(q)}\right\}, \quad (g \neq H+1). \qquad (22)$$

Note that the cluster parameters $\hat{v}_{3(q)}$ are estimated on the cells (g, h) where $h \in \{1, \dots, H\}$ and their values are extrapolated to $h = H + 1$.

As it stands, we cannot use expression (22) because we cannot estimate the effect $\hat{v}_{2(H+1)}$ directly from the audit sample. However, taking the sum of (22) over all cells in this column we obtain:

$$\sum_{g=1}^{H} \widehat{N}_{g,H+1,model} = \exp\{\hat{v}_{2(H+1)}\} \sum_{g=1}^{H} \exp\left\{\hat{v} + \sum_{q=1}^{Q} \delta_q(g, H+1)\hat{v}_{3(q)}\right\}. \qquad (23)$$

According to (21), the left-hand sum should be equal to $\widehat{N}_{H+1,+,model}$ which is known after 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 (23). Next, the underlying estimates $\widehat{N}_{g,H+1,model}$ can be obtained from (22). Finally, we can use all estimated counts $\widehat{N}_{gh,model}$ to obtain estimates of $\hat{\psi}_{model}(g, h)$ as in (20). This completes the first level of the model for industry $H + 1$.

Subdividing units in $H + 1$ into underlying industries

The model from the previous subsection allows us to estimate $P(\hat{s}_i \in \mathcal{H}^* \backslash \mathcal{H} | s_i = h)$ and $P(\hat{s}_i = h | s_i \in \mathcal{H}^* \backslash \mathcal{H})$, 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", to take into account the fact that

- the transition probabilities $P(\hat{s}_i = h | s_i = g)$ with $h \in \mathcal{H}$ are not identical for all $g \in \mathcal{H}^* \backslash \mathcal{H}$;
- the distribution of turnover is different for each industry $g \in \mathcal{H}^* \backslash \mathcal{H}$.

In other words: the accuracy of turnover estimates for a target industry will depend in general on the specific set of non-target industries that contribute most to the "inflow of turnover".

We therefore introduce a second level of the model for transitions between the sets of target and non-target industries. As there is little information available about classification errors at this detailed level, we propose a simplified model that consists of the following steps:

1. Given that $N_{H+1,+}$ units that actually belong to one of the non-target industries are misclassified in one of the target-industries, we draw $(N_{H+1,1}, \dots, N_{H+1,H})$

from the multinomial distribution with parameters $N_{H+1,+}$ (total inflow) and $\psi(H+1,1), \ldots, \psi(H+1,H)$. By construction, $\sum_{h \in \mathcal{H}} N_{H+1,h} = N_{H+1,+}$.

2. Given that $N_{H+1,h}$ units that actually belong to one of the non-target industries are misclassified in target industry h, we draw a specification of corresponding numbers $\left(N_{gh}\right)$ for all $g \in \mathcal{H}^* \backslash \mathcal{H}$ from the multinomial distribution with parameters $N_{H+1,h}$ and f_{gh}. Here, f_{gh} denotes the relative contribution of each non-target industry to the total number of misclassifications in target industry h. By construction, $\sum_{g \in \mathcal{H}^* \backslash \mathcal{H}} N_{gh} = N_{H+1,h}$.
3. Given that N_{gh} units that actually belong to non-target industry $g \in \mathcal{H}^* \backslash \mathcal{H}$ are misclassified in target industry h, we draw their turnover values from a log-normal distribution $LN\left(\mu_g, \sigma_g^2\right)$.

This approach also has an important computational advantage, as data on the individual units outside the target set of industries are not required during bootstrap simulation. We already discussed how to estimate the parameters for the first step in the previous subsection. In the remainder of this subsection, we briefly summarise how the remaining parameters were estimated in the case study to be discussed below.

For the second step, we used the data of the yearly transitions in the BR to estimate f_{gh}. As mentioned above, we conjecture that the distribution of these transitions is similar to the distribution of misclassifications in the BR. In particular, we assume that there is an overlap between units that change their observed economic activity between times $t-1$ and t, and units that were misclassified at time $t-1$. With this in mind, we estimated f_{gh} by the number of units that were observed in industry h within the target set in year $t-1$ ($\hat{s}_i^{t-1} = h$) and in industry g outside the target set in year t ($\hat{s}_i^t = g$), taken as a fraction of the total number of units with $\hat{s}_i^{t-1} = h$ and $\hat{s}_i^t \in \mathcal{H}^* \backslash \mathcal{H}$. To obtain more data, we averaged these numbers over five years (2009–2014). Moreover, we ordered the industries outside the target set by the total number of units that entered these industries from within target set. For the second and third step, we restricted attention to the subset of industries in $\mathcal{H}^* \backslash \mathcal{H}$ with the largest contributions, in such a way that for each of the industries $h \in \mathcal{H}$ at least 70 per cent of the outflowing units were covered. Thus, as an approximation we assumed that all misclassifications with respect to the target set of industries were confined to this subset of the total NACE domain.

For the third step, we fitted a log-normal distribution to the observed values of turnover within each non-target industry (restricted to the above subset), separately for each time period. We also made a distinction between units with size class 0–3 and other units, obtaining separate estimates of μ_g and σ_g^2 for both groups. To obtain robust estimates, outliers were identified and removed prior to estimation.[1]

Note that, when during the bootstrap simulations turnover values are drawn from a log-normal distribution to simulate units that enter target industry h by mistake, it

[1] We used the following criterion: within each group/industry-combination observations were considered as outliers if they deviated more than 7 from the 0.05 two-sided trimmed mean (on a log scale).

could happen that some of these values are relatively large compared to the values of turnover that are usually observed in this industry. As a result, the bootstrap replications of total turnover for that industry would be highly volatile and we might conclude that misclassifications have a large impact on the estimated turnover. However, this scenario is not realistic: in practice, if a unit entered an industry with an unusually large turnover compared to the other units, this would trigger follow-up actions by subject-matter experts and, most likely, the classification error would be identified. To take this into account in our bootstrap simulations, we added the following restriction: when drawing from the log-normal distribution to simulate a unit that enters target industry h, if we obtained a value that was larger than the largest observed value within industry h, then we rejected the value and drew a new one, repeating this procedure if necessary until an appropriate value was obtained. Effectively, this means that turnover values were drawn from a truncated log-normal distribution.

2.3 Bias correction

A technical assumption used in the derivation of (4) is that $p_{hhi} > \max_{g \neq h} p_{ghi}$ for all $h \in \mathcal{H}^*$ and $i = 1, \ldots, N$; see Burger et al. (2015). Even when this assumption is satisfied, $\hat{B}_R^*(\hat{Y}_h)$ in (4) is generally 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 easily be corrected. In our case it is also possible to compute an unbiased bootstrap estimator of $B(\hat{Y}_h)$; see Appendix A. In terms of the notation in Appendix A, we denote the original bias 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 $\hat{B}_{\omega R}^*(\hat{Y}_h)$:

$$\hat{B}_{\omega R}^*(\hat{Y}_h) = \omega \hat{B}_{1R}^*(\hat{Y}_h) + (1 - \omega)\hat{B}_{0R}^*(\hat{Y}_h), \tag{24}$$

where the relative weight ω is determined by minimising the mean squared error of $\hat{B}_{\omega R}^*(\hat{Y}_h)$. The exact procedure – which actually involves optimal weights at a more detailed level than indicated in (24) – is given in Appendix A. 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 $\hat{B}_R^*(\hat{Y}_h)$; see Appendix A 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.

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 car trade are subdivided into nine industries, the base cells (see the Introduction). The quarterly turnover is estimated from a mixed-source production system, see, 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**. The remaining units are observed in a census survey. On the 1st of January 2013 these remaining units corresponded to 8403 enterprises within the whole domain of economic activities and 239 within car trade. For a subset of this group, there is a special business unit at Statistics Netherlands with centralised data collection and data editing. This concerned 2305 enterprises within the whole domain of economic activities and 49 within car trade. This latter subset is referred to as the complexity class **most complex units**. The other units observed in a survey but not treated by this special business unit are referred to as the **complex units**.

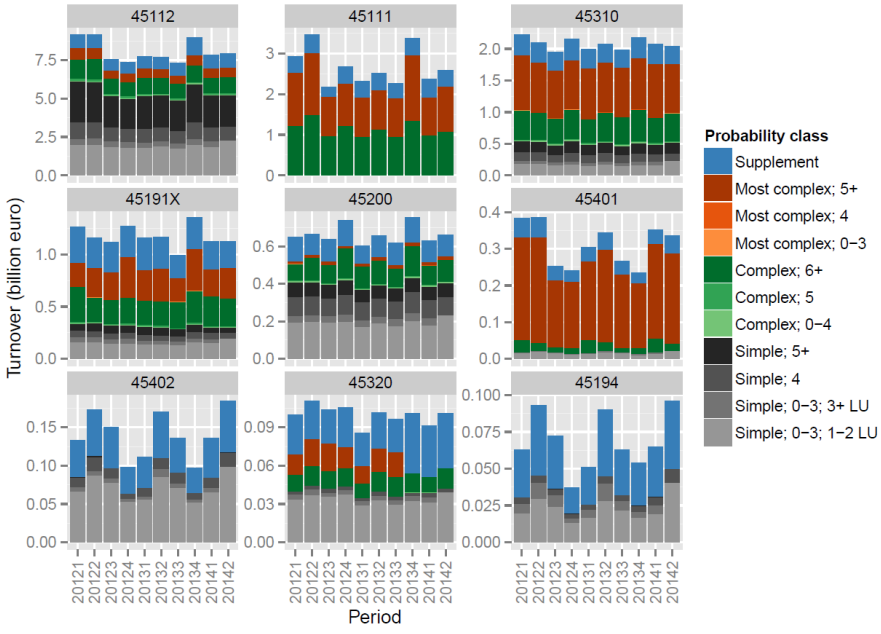
The quarterly outcomes are published in different releases: 30 days (flash), 60 days (early), 90 days (late) and one year (final) after the end of the reference period. The computations in the current paper concern the most recent releases that were available. For 2012 and 2013 this concerns the final release and for Q1 and Q2 of 2014 this concerns the late release. The available microdata covered nearly the complete target population. At a late release, quarterly non-respondents are missing and units that report their VAT on a yearly basis. The latter group corresponds with 2–3% of the total turnover. Missing values are imputed. At 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 (we do not compute the effect of the imputation process on the accuracy).

The nine industries within car trade vary considerably in number of enterprises, total turnover and turnover per enterprise (see Table 3.1.1). In the first quarter of 2013, total turnover varies from 7748.7 million euro (industry code 45112) to 51.04 million euro (industry code 45194). The division of total turnover in the different complexity classes also varies considerably over the nine industry codes (see Figure 3.1.2; some terms used in this figure will be explained in Section 4.1). Note that throughout this paper the industry classes are ordered from the largest to the smallest total turnover per industry.

3.1.1 Main characteristics of the nine industries in car trade (estimates from Q1 2013).

Industry code	Economic activity (description)	$\widehat{Y}_h$ ($\times 10^6$€)	$\widehat{N}_h$	$\widehat{\bar{Y}}_h=\widehat{Y}_h/\widehat{N}_h$ ($\times 10^3$€)
45112	Sale and repair of passenger cars and light motor vehicles (no import of new cars)	7 748.7	17 458	443.8
45111	Import of new passenger cars and light motor vehicles	2 313.0	132	17 523.0
45310	Wholesale and commission trade of motor vehicle parts and accessories	1 993.1	1 922	1 037.0
45191X	Sale and repair of trucks and trailers	1 159.0	1 309	885.4
45200	Specialised repair of motor vehicles	602.1	5 512	109.2
45401	Wholesale and commission trade of motorcycles and related parts	303.4	420	722.5
45402	Retail trade and repair of motorcycles and related parts	111.0	1 063	104.4
45320	Retail trade of motor vehicle parts and accessories	85.6	692	123.6
45194	Sale and repair of caravans	51.0	336	151.9

3.1.2 Distribution of quarterly turnover.



The probabilities that were used to model the classification errors were estimated using three sources:

– We took an audit sample from the population of the simple enterprises within car trade that existed on the 1st of July 2014 according to our BR. From each of the nine industries we randomly sampled 25 enterprises. Next the true NACE codes

were determined by two experts, examining the Chamber of commerce information, internet data and by contacting the enterprise in case of doubts.

- For the complex and most complex enterprises we consulted experts at SN that are responsible for the editing process of the car trade industry and experts from a special business unit at SN that treats the large and complex units. We used expert knowledge for those enterprises, because, based on quality studies reported in 2000 and 2003, 97% of these enterprises were expected to have a correct three-digit NACE code (see 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 much capacity. The experts were used to estimate the relative levels of classification error and the largest levels were set at 5 per cent 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 (see Burger et al., 2015).
- In addition, we used data of our BR for the years 2009–2014 on the yearly transitions in NACE code of the enterprises. 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 $t-1$ ($\hat{s}_i^{t-1} = h$) averaged over 2009–2014. The motivation behind this approach was given in Section 2.2. Based on the results of the temporal transitions, we have asked experts to appoint each cell (g, h) to a cluster $q \in \{1, \dots, Q\}$, 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 for the 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 code;
- number of legal units;
- legal form;
- size class of the enterprise;
- being observed in a sample survey (yes/no).

The industry may play a role for a number of reasons: some economic activities are easily confused with each other; in some industries units can change their economic activities relatively easily and these changes may enter the BR with a substantial

delay, while other industries are more stable (e.g., because changing activities requires a large investment); there may be incentives for units to try to obtain a particular industry code (e.g., for fiscal reasons). The number of legal units of an enterprise may play a role because each legal unit may have its own economic activity, and errors can be made while deriving the main economic activity. The legal form is another potential indicator for the complexity of a unit. The size class may have an additional effect (apart from other indicators of the complexity of an enterprise), because larger units are more influential on the outcomes and therefore more editing effort is put into them. Finally, units that are observed in a sample survey are more prone to editing checks in the production system than units that are not observed, so we expect less classification errors for those units.

According to the audit sample, the simple units of the industries 45111, 45401 and 45320 were most prone to classification errors (Table 4.1.1). Note that the net sample size per (observed) stratum was sometimes below 25 because a few units had ceased to exist between the moment the sample was drawn (1 July 2014) and the moment the review was done (August 2014).

4.1.1 Results of the audit sample (prefix '45' suppressed in column header)

true\obs	'112	'111	'310	'191X	'200	'401	'402	'320	'194
45112	23	19	1	2	1	1		4	2
45111		4							
45310			23					2	
45191X		1		21					
45200			1		22			6	
45401						11			
45402						11	23		
45320								10	
45194					1				21
other	1	1		2		2	2	1	2
total	24	25	25	25	24	25	25	23	25

We started by analysing the results of the audit sample (Table 4.1.1), taken from the simple enterprises. We aimed to find a parsimonious logistic regression model of the form (12) to explain classification errors. In fact, the audit sample contained no classification errors among the simple enterprises with size class 4 or larger (10 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). The probabilities for the larger simple units were modelled together with those of the complex and most complex enterprises by using expert knowledge; see below.

We investigated all possible combinations of the background variables (subset selection). Table 4.1.2 displays the best fitting models with one, two, and three predictor variables. To compare the models, we looked at the deviance values (based on the log-likelihood). For the nested models 1, 2, and 3, the third column in Table

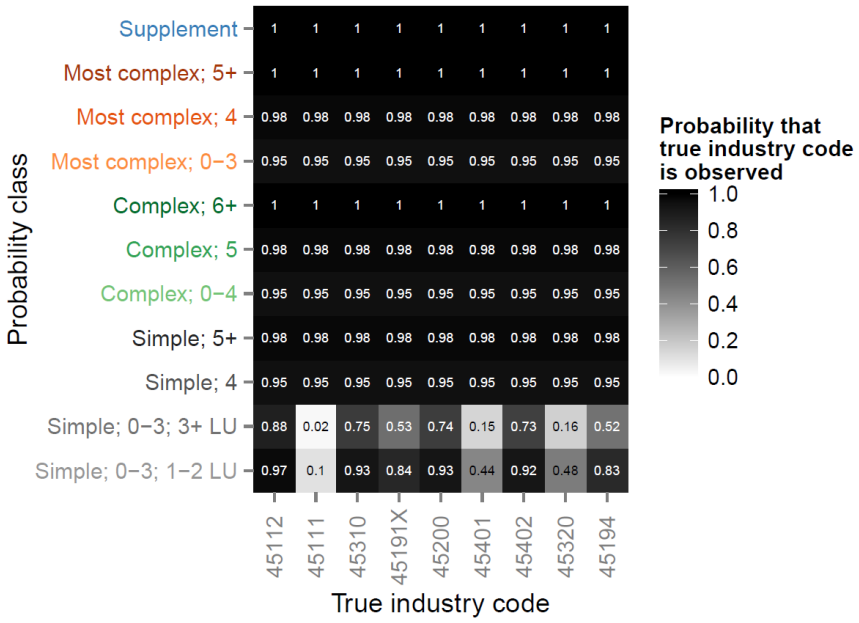
4.1.2 shows the decrease in deviance compared to the model directly above it. The fourth column shows the p value of a chi-square test of this difference (cf. McCullagh and Nelder, 1989). It is seen that the model fit was improved significantly by adding the observed industry and the number of legal units (in two classes: 1–2 units and ≥ 3 units). The next model, which also included a 0-1 indicator for units that were recently observed in a survey, did not yield a significantly better fit according to the chi-square test ($p = 0.14$), although it did have a marginally better AIC value. We also verified the model selection results by cross-validation: by fitting the model on 90% of the data and predicting the other 10% and repeating this procedure ten times such that in the end all units are predicted. Taking the results of this cross-validation procedure (not shown) into consideration as well, we decided to use model 2 from Table 4.1.2 to estimate the diagonal probabilities for the remainder of this case study.

4.1.2 Selected results of 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

The estimated probabilities based on model 2 are given in the bottom two rows of Figure 4.1.3. The numbers in the labels “0–3”, “4”, “5”, “5+”, “6+” stand for the size classes and “1–2 LU” and “3+ LU” stands for the number of legal units per enterprise.

4.1.3 Estimated transition probabilities for the diagonal elements



The diagonal probabilities of the upper nine rows of Figure 4.1.3 were based on experience from experts at SN in editing. We limited ourselves to three background variables of the enterprises: the size class, the complexity class of the units and supplementary editing (yes/no). From now on, the strata defined by these background variables, corresponding to the rows of Figure 4.1.3, will be referred to as probability classes (PCs).

The variable supplement (yes/no) concerns the enterprises that are edited thoroughly by the statistical division at SN that is responsible for the output. Enterprises that belong to the PC 'supplement=yes' have transition probabilities of 1.0 on the main diagonal (first row of Figure 4.1.3), regardless of the further characteristics of the unit. In practice, for each base cell the size of this supplement approximately equals the 25 enterprises with the largest turnover. A more precise definition will be given in Section 5.1.

4.2 Estimated probabilities for the off-diagonal elements

The average over the yearly transitions of the NACE codes over 2009–2014 are given in Table 4.2.1. The NACE codes in the old year are given in the top row, the NACE codes of the new year are in the left column, and for each year the transitions in each column were normalised to 1 and then the results were averaged over the years. The grey colouring indicates the four clusters that were appointed by the experts. The largest group is those for which the cells have the smallest transition probability.

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 (17). 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, p. 125); non-significant values indicate that all relevant factors have been included in the model.

There was one cell (observed NACE of 45111 and true NACE 45112) that had a large number of cases (19 of the sample of 25 in observed NACE 45111; see Table 4.1.1) that dominated the estimates 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 better fit the observed numbers in the audit sample.

Table 4.2.2 shows the final model-based expected off-diagonal elements after correcting for the sampling fractions n_{+h}/N_{+h} using equation (19). The resulting estimated off-diagonal probabilities according to equation (20) are shown in Figure 4.2.3. Recall that the probabilities for the $(H+1)^{\text{th}}$ industry (column) were derived using equations (21)–(23).

4.2.1 Relative transition of enterprises for the off-diagonal elements according to the BR 2009–2014 and four clusters (dark background: high probability, light background: low probability; "-" means exactly zero). Each column adds up to 1.

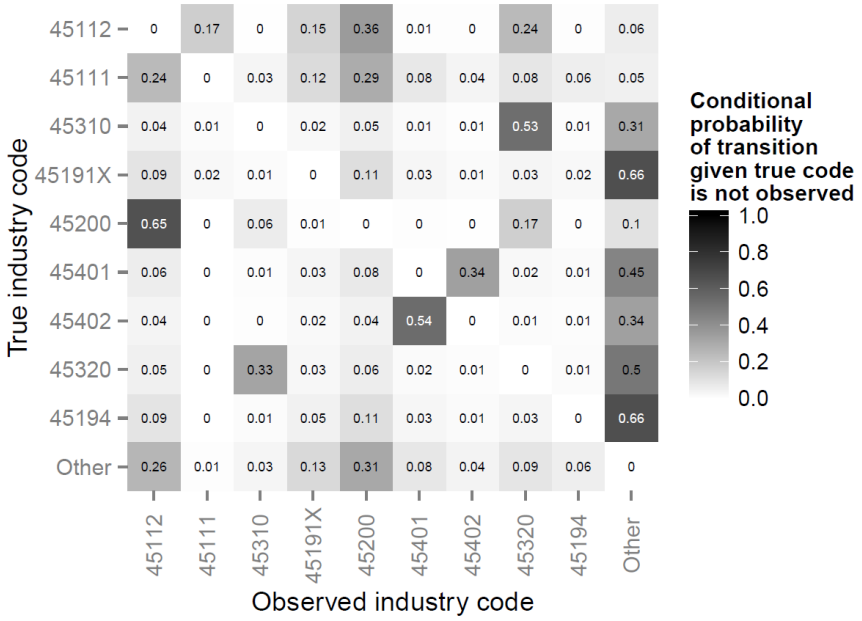
$t \setminus t-1$	45112	45111	45310	45191X	45200	45401	45402	45320	45194	other
45112		0.684	0.339	0.355	0.365	0.011	0.047	0.086	0.066	0.392
45111	0.010		0.003	0.005	-	-	-	0.002	0.016	0.005
45310	0.035	0.053		0.009	0.015	-	-	0.045	-	0.068
45191X	0.020	0.053	0.003		0.006	-	-	0.002	-	0.036
45200	0.112	0.018	0.054	0.028		-	0.004	0.041	-	0.106
45401	0.002	-	0.003	-	-		0.043	-	-	0.024
45402	0.008	-	0.003	-	0.005	0.157		0.002	-	0.116
45320	0.017	0.018	0.058	0.005	0.055	-	0.009		-	0.221
45194	0.004	-	-	-	0.002	-	-	0.002		0.031
other	0.791	0.175	0.537	0.598	0.553	0.831	0.897	0.819	0.918	

4.2.2 Estimated off-diagonal elements $\hat{N}_{gh,model}$, for the audit sample, size classes 0–3.

true\obs	'112	'111	'310	'191X	'200	'401	'402	'320	'194	other
45112	-	91.2	1.4	81.0	192.3	3.4	1.6	131.7	2.5	32.3
45111	10.7	-	1.4	5.4	12.9	3.4	1.6	3.7	2.5	2.2
45310	10.7	2.7	-	5.4	12.9	3.4	1.6	131.7	2.5	76.6
45191X	10.7	2.7	1.4	-	12.9	3.4	1.6	3.7	2.5	76.6
45200	514.9	0.1	50.2	5.4	-	3.4	1.6	131.7	2.5	76.6
45401	10.7	0.1	1.4	5.4	12.9	-	58.2	3.7	2.5	76.6
45402	10.7	0.1	1.4	5.4	12.9	165.4	-	3.7	2.5	104.1
45320	10.7	0.1	68.2	5.4	12.9	3.4	1.6	-	2.5	104.1
45194	10.7	0.1	1.4	5.4	12.9	3.4	1.6	3.7	-	76.6
other	159.7	3.7	21.1	81.0	192.3	51.3	24.5	55.5	36.5	-
total	749.7	100.8	148.0	200.0	475.0	240.8	94.2	469.1	56.2	625.7

These 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 misclassified – to be observed as 45320 (retail trade of motor vehicle parts and accessories). Likewise, misclassified units from industry 45320 have a probability of 0.33 to be 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 misclassified units from the car trade industries 45194, 45320, 45402, 45401, 45191X, and 45310 all have a probability > 0.30 to be observed outside car trade.

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



4.3 Estimated probabilities for the industries outside car trade

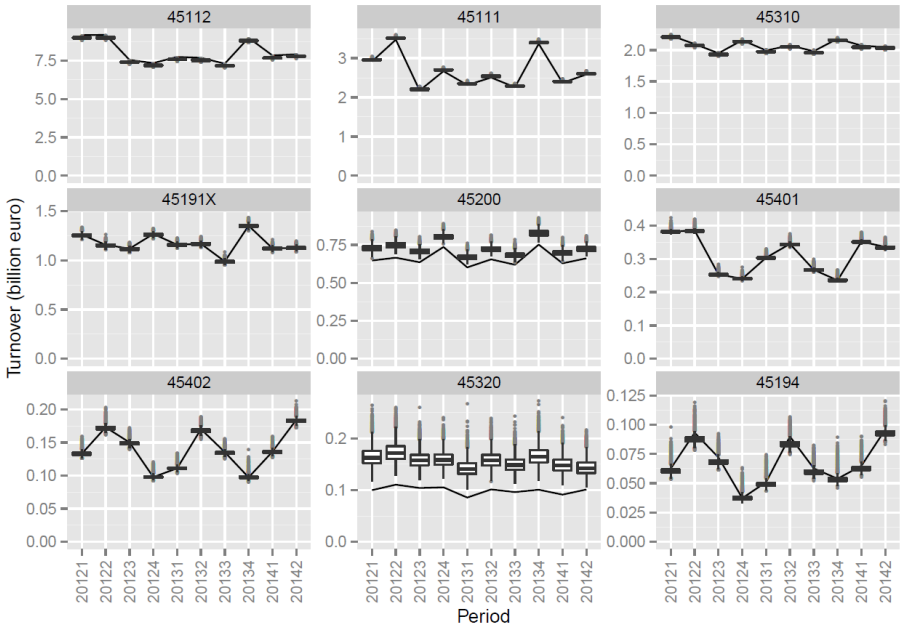
Based on the data of the yearly transitions, we found a set of 54 base cells that covered, for each of the car trade industries, at least 70 per cent of the total number of yearly outflowing units to industries outside car trade (see also Section 2.2). The relative contributions (f_{gh}) among the 54 base cells to each car trade industry are found in Appendix B. The estimated parameters of the log-normal distributions, $LN(\mu_g, \sigma_g^2)$, are not shown here.

4.4 Simulation of accuracy

Having modelled the probabilities of classification errors for the data in our case study, we applied the bootstrap method from Section 2.1. We chose $R = 10{,}000$ as the number of replications. The observed versus simulated values (box plots) per base cell are given in Figure 4.4.1. We summarised the results also as 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) $\sqrt{\left\{\left[\hat{B}_R^*(\hat{Y}_h)\right]^2 + \hat{V}_R^*(\hat{Y}_h)\right\}}/\hat{Y}_h$.

4.4.1 Observed versus simulated values per base cell

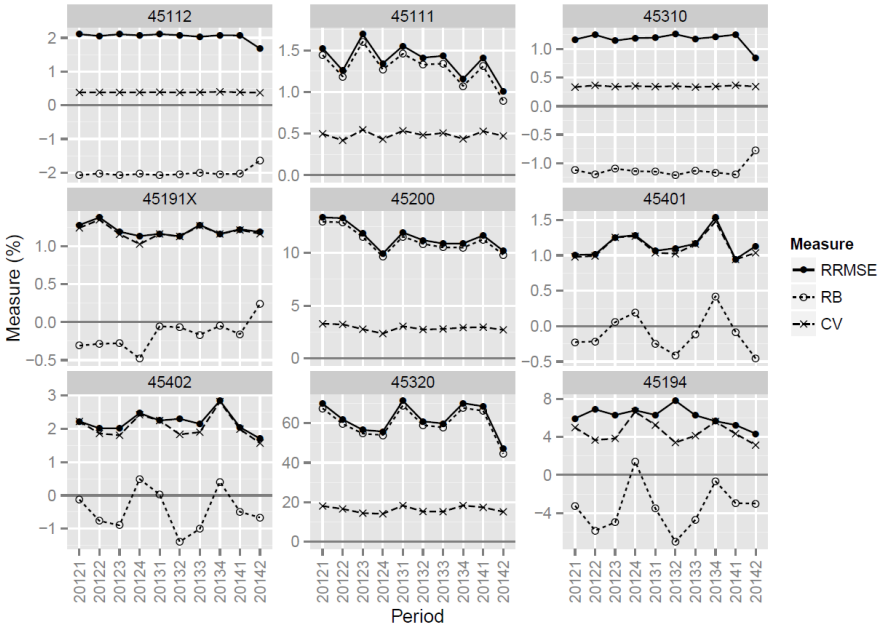


These results are shown in Figure 4.4.2 (expressed as percentages). The RRMSE varies from about 1.0% for the industries 45401 and 45310 to about 60% for industry 45320. The variance (CV) dominates in the industries 45191X, 45401, 45402 and 45194, in the other industries the bias dominates. The industries 45112 and 45310 both have a negative bias. A negative bias means that the values of bootstrap simulations ($\hat{Y}^*_{hr}$) 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).

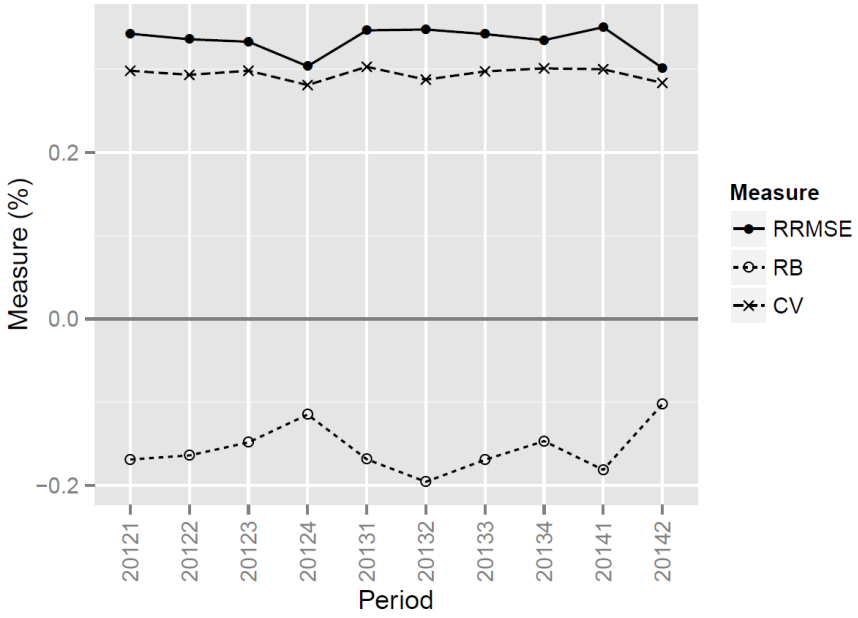
We found that industry 45320 has a very large RRMSE: on average 62% (Figure 4.4.2). This industry has a relatively large probability of classification error on the diagonal elements (Figure 4.1.3) of the complexity class “simple”, and this class constitutes about one third of the total turnover in this industry (see Figure 3.1.2). Industry 45111 has an even larger probability on classification errors in the complexity class “simple” (Figure 4.1.3) but it does not have a large RRMSE. The latter is true because the turnover of the simple enterprises in industry 45111 is very small compared to the other complexity classes (see Figure 3.1.2).

The RRMSE for car trade as a whole is about 0.33% and was relatively stable over the ten periods (Figure 4.4.3). The CV was also relatively stable (about 0.29%). The RB varied most and ranged between –0.2% and –0.1%.

4.4.2 RRMSE, RB and CV for ten periods per base cell

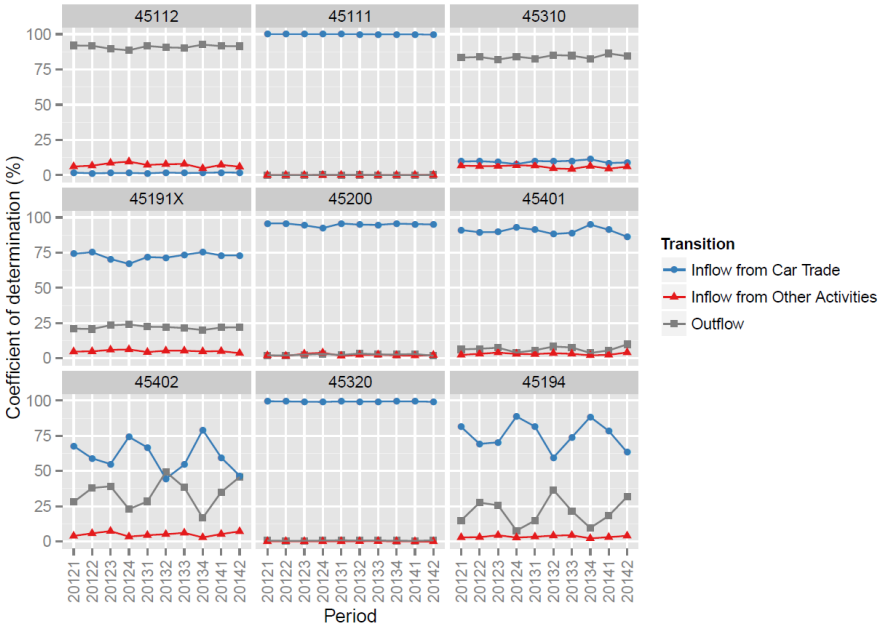


4.4.3 RRMSE, RB and CV for ten periods for car trade as a whole

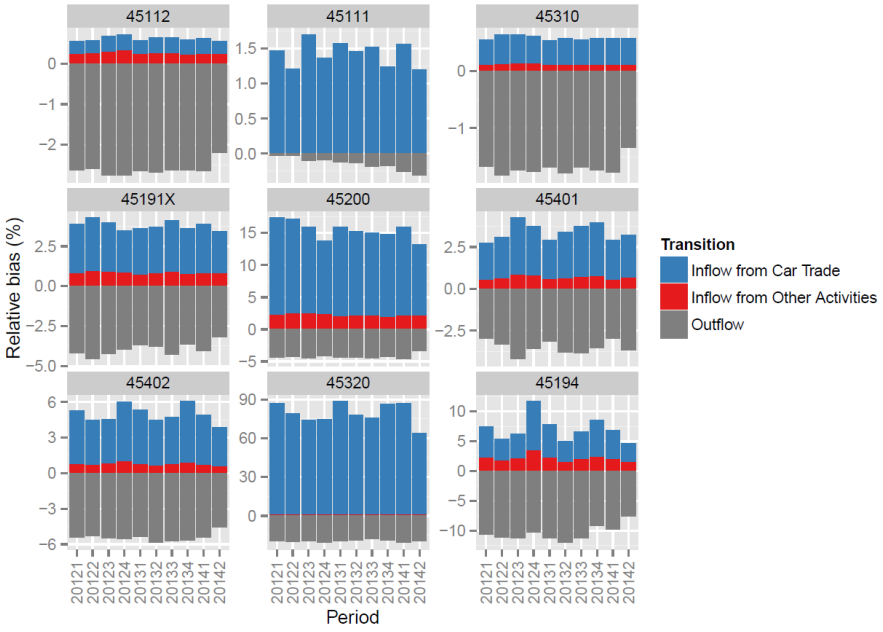


To better understand the differences in accuracy among the industries, we separated the outcomes of the variance and the bias per industry into three transition classes, as explained in Section 2.1. In the context of this case study, “Inflow from the target set” may be read as “Inflow from car trade”. The results are shown in Figure 4.4.4 and Figure 4.4.5. Recall that the contribution of the class “Outflow” to the overall bias in (7) is negative and therefore partly balances the contributions of the two “Inflow” classes.

4.4.4 Coefficient of determination of three transition classes



4.4.5 Relative bias of three transition classes



Using the three transition classes, we can better understand the 'causes' of the variance and bias for each industry. First consider the results for industry 45112. The variance and the bias are dominated by the transition class "outflow", see Figure 4.4.4 and Figure 4.4.5. How can we explain this? Inflow within car trade to industry 45112 consists mostly of units from 45200 (see Table 4.2.2). Industry 45200 has a smaller turnover per enterprise than 45112 (Table 3.1.1). The effect of outflow from 45112 therefore dominates the (main components) of inflow. For industry 45111 we found that the variance and the bias are dominated by the transition "inflow from car

trade". This base cell has by far the largest inflow from 45112 (see Table 4.2.2). The units that enter industry 45111 from 45112 (mainly simple units) will have a larger turnover than the units that stay in 45111 (not shown). That explains that "inflow from car trade" dominates in industry 45111.

Industry 45112 and 45200 are two examples of industries with a bias that is unequal to zero, whereas the bias of 45401 and 45402 varies around 0 (Figure 4.4.2). What is the reason for this difference? Figure 4.4.5 reveals that in 45401 and 45402 the negative bias due to outflow is balanced by positive bias due to inflow from car trade and from other activities, yielding a net bias of around zero. In industry 45112 however, the negative bias due to outflow is much larger than the positive bias from inflow probably because the turnover per enterprise is larger in 45112 (as observed code) than in other industries, at least for the most influential PCs (not shown). In industry 45200 the bias is dominated by positive bias of the inflow from car trade. In this industry the inflow is mostly from 45112 (Table 4.2.2). The average turnover per enterprise in 45112 is larger than in 45200, so the inflow from car trade will dominate the transition by outflow.

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** (see Figure 3.1.2). 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. Below we will explain the difference between our approximation and the (true) effect of editing.

We will compare four levels of supplementary editing, namely 0, 225, 450, and 675 edited enterprises in car trade. Since our results on accuracy were reasonably consistent over the 10 quarters, we only computed the results for one quarter: the first quarter of 2013. The second level (225 units) corresponds reasonably well with the actual situation at Statistics Netherlands. We distinguish 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 correspond to 0, 25, 50, and 75 enterprises per industry.
2. Pro rata: the number of enterprises to be edited per base cell (n_h^E) is in proportion to the product of $RMSE(\hat{Y}_h) = \sqrt{\left\{\left[\hat{B}_R^*(\hat{Y}_h)\right]^2 + \hat{V}_R^*(\hat{Y}_h)\right\}}$ and the population size per industry (N_h):

$$n_h^E = \frac{RMSE(\hat{Y}_h)N_h}{\sum_{h=1}^{H} RMSE(\hat{Y}_h)N_h} n^E, \tag{25}$$

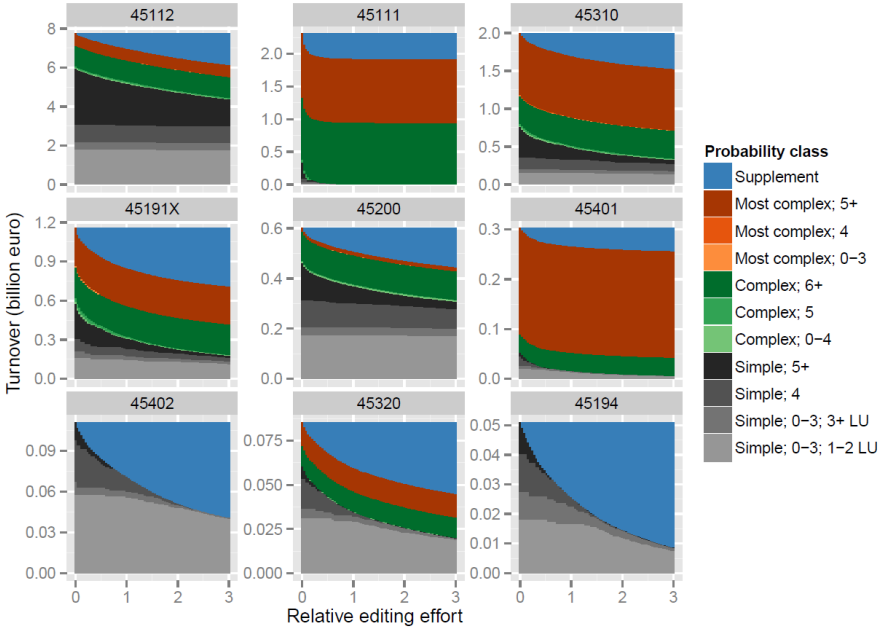
where n^E denotes the total number of units selected for supplementary editing. Notice that equation (25) resembles the so-called Neyman allocation of a survey sample over its underlying industries (see, e.g., Cochran, 1977, pp. 98–99). Because of this analogy, one might expect the accuracy of the estimated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

In the above strategy we ignored possible differences in editing time between different units. In fact the editing time per enterprise depends on the complexity class of the unit. Experts at SN estimated that a clerical reviewer can edit eight simple units per hour and six complex or most complex units per hour. A slightly more realistic editing scenario would therefore fix the total available editing time, say t^E. Under the pro rata strategy, industry h would then be allocated an editing time t_h^E according to equation (25) with n^E replaced by t^E. Given this allocated editing time, an optimal selection strategy for units would work as follows. Let t_i denote the editing time required for unit i. Within each industry the units are ordered according to the ratio y_i/t_i from large to small. Next, select those units that have the largest values for the ratio y_i/t_i and that did not already have a diagonal probability of 1, until the available editing time is reached. This strategy is equivalent to maximising the total amount of turnover that can be edited in a given amount of editing time. We did not implement this refinement in our case study, because the differences in editing time per unit were small.

The total turnover as a function of the four levels of supplementary editing ('relative editing effort') is given in Figure 5.1.1.

The difference between our approximation and the (potential) true effect of editing is presented in Figure 5.1.2. Suppose we are interested in estimating the accuracy of industry 45320. First consider the "starting situation" (25 units per base cell in the supplement). The transition probabilities of Figure 4.2.3 are given. In the starting point of the bootstrap simulations the units that are observed in 45320 and the relative distribution of those units over the industries after a bootstrap stimulation are given by the values in the row 45320 in Figure 5.1.2. Now, when we are interested in splitting up the accuracy of $\hat{Y}_h$ in the three transition classes, the outflow is given by the grey row, and the inflow from car trade by the blue column and the inflow from other activities by the red cell in Figure 5.1.2.

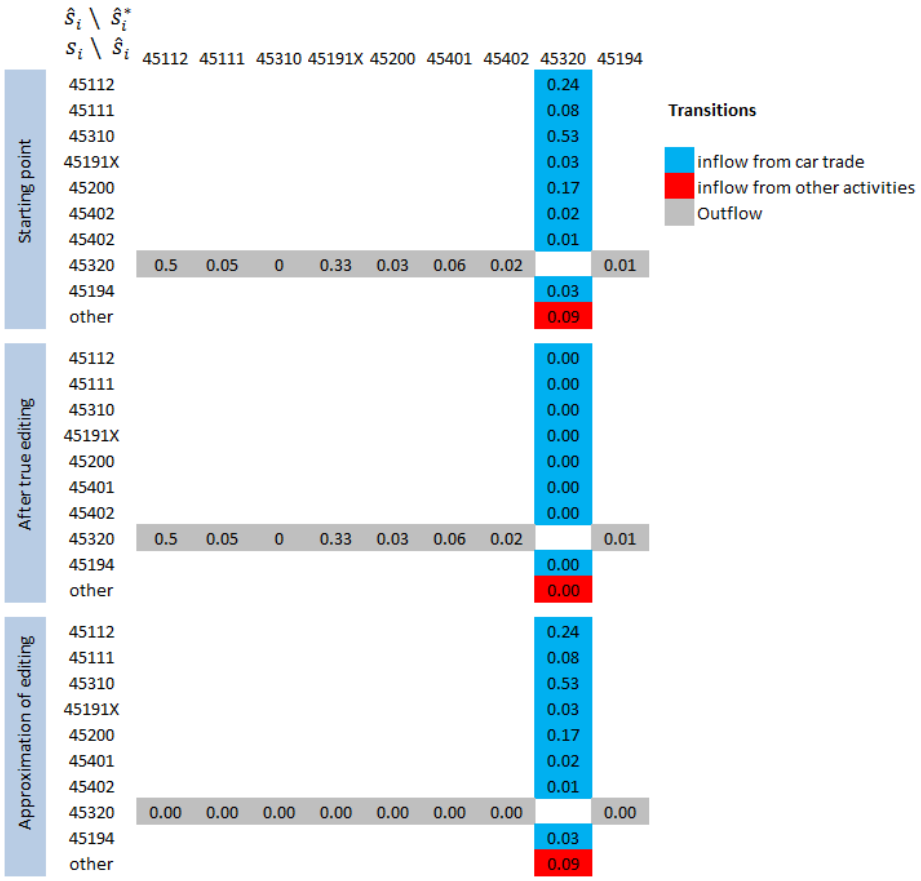
5.1.1 The turnover distribution per unit type for the four relative editing efforts (scenario fixed)



To illustrate the problem clearly, we consider the extreme scenario that *all* units in base cell 45320 are submitted to supplementary editing, and no supplementary editing is done in the other base cells. The middle part of Figure 5.1.2 gives the (hypothetical) situation after truly editing all observed units in 45320. Column 45320 stands for the units that are observed in 45320; the rows are the true values. Since there are no classification errors any more, the off-diagonal elements in the column of 45320 are turned to zero. The row-values are not turned into zero, because the units observed in other industries were not edited.

Finally, the bottom part of Figure 5.1.2 represents our approximation of supplementary editing under this extreme scenario. We now assume that our observations (as found in the BR) are error-free as the starting point of our simulations. This is equivalent to using values of zero in the off-diagonal elements of the row of 45320 in Figure 5.1.2. Now consider the effect of the transitions for industry 45320 in Figure 4.4.5. In case of truly editing this industry, we expect the bias to *decrease*, because the inflow from car trade (blue part) reduces to zero. In our approximation however, the bias *increases* because in our simulations the outflow (the grey part) reduces to zero but the blue part remains.

5.1.2 True and simulated effects of supplementary editing on classification error probabilities.

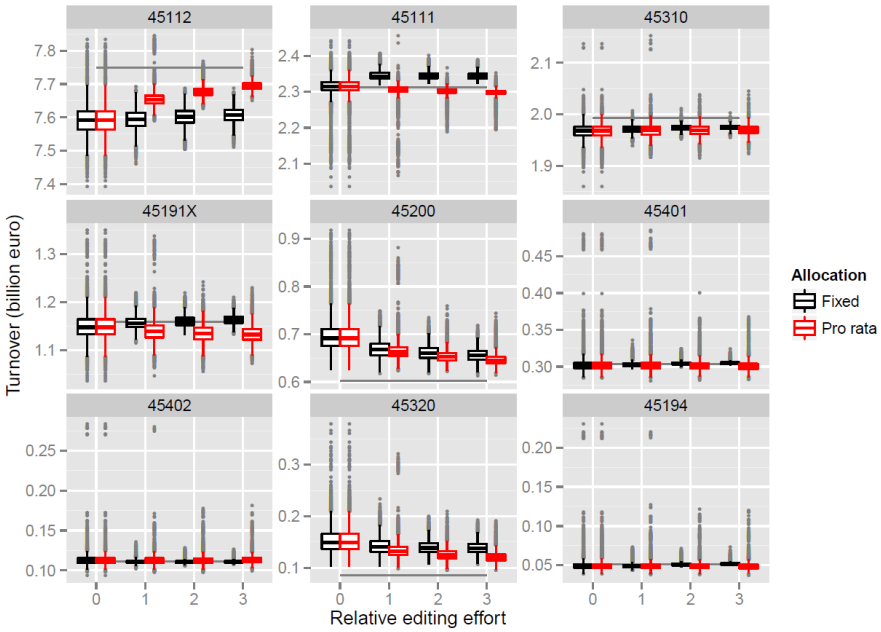


This shows that our approximation cannot be used directly to estimate the accuracy effects of a supplementary editing of units. Still, the next section presents surprising results on the difference between the two scenarios that we believe are useful to know when considering true editing. We also simulated a *reduction* in the editing of units, by reducing the size of the supplementary class from 225 to 0 units. For the latter case our results are directly valid, because we know the true NACE code values of the edited units.

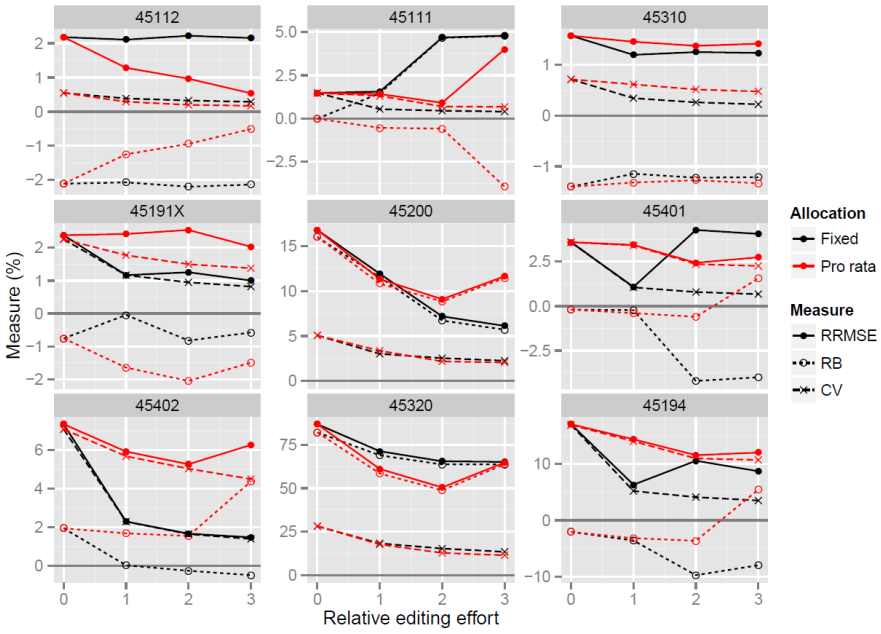
5.2 Simulation of editing

The progress of the accuracy measures with increased relative editing effort for the nine industries (Figure 5.2.1 and Figure 5.2.2) and the two editing scenarios shows different interesting results. First of all, as expected, the coefficient of variation (CV) decreased with increased relative editing effort. Also the absolute value of the relative bias (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 45111 where the absolute RB clearly increased between the relative editing effort 1–3 for the fixed scenario, and between the relative editing effort 2 and 3 for the pro rata scenario.

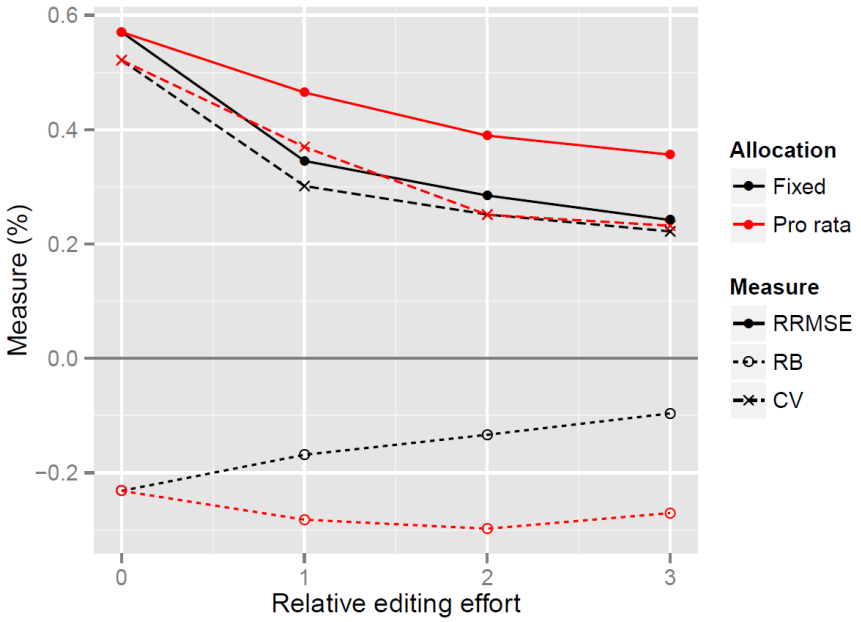
5.2.1 Simulating the effect of editing on accuracy: simulated (box plot) vs. observed (line) for Q1 2013



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



5.2.3 Simulating the effect of editing on accuracy: overall effect on car trade



We can understand this surprising phenomenon looking at the size of the bias for the three transitions, see Figure 4.4.5. For each of the editing scenarios the outflow component per industry will be reduced (remember Figure 5.1.2), while the inflow from car trade will stay at a relatively high level. Figure 4.4.5 indicates that a reduction in outflow for industry 45111 will lead to an increase of the bias. Of course also the absolute level of inflow from car trade will decrease because the outflow from all other car trade industries will be reduced, but still the balance between out- and inflow on the bias becomes less favourable. The overall effect of the change of CV and RB with increased editing effort is that the RRMSE is not always decreasing with increased editing effort.

Figure 5.2.2 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 car trade as a whole (sum of nine base cells) is larger for the fixed scenario than for the scenario pro rata; see Figure 5.2.3. This can be understood as follows. The pro rata scenario is inspired upon the Neyman allocation that assumes that the errors $\hat{Y}_h - Y_h$ are independent of each other. This assumption however does not hold in the case of classification errors since many off-diagonal transition probabilities are larger than zero. By changing the number of edited units in an industry, we can control the expected size of the "outflow" component in that industry, but not the "inflow" components. As can be seen in Figure 4.4.4 and Figure 4.4.5, in some industries the total error is mainly determined by "inflow". 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

In the current paper we computed estimates of accuracy measures of register-based outcomes due to classification errors. The results showed that the quarterly turnover level estimates for car trade as a whole have an RRMSE of 0.3%, which is judged as an acceptable accuracy by the owner of the production process. One of the underlying base cells (industry 45320) was very inaccurate, with an average RRMSE of 62%. 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% (not shown). The least accurate industry that is actually published is 45200 with an RRMSE for the quarterly turnover slightly larger than 10%. Statistics Netherlands aims to have a maximum uncertainty margin of 3 per cent points on turnover levels. This means that in car trade an additional editing effort is needed to improve industry 45200 (see more below) or, alternatively, to combine results of industry 45200 with another industry.

We summarised our results in terms of RB, CV, and RRMSE. It would be slightly more accurate to use a bootstrap confidence interval based on the simulated data compared to the observed data, because results of the simulation show that this interval was not symmetrical in our case (see Figure 4.4.1). Nonetheless, using RB and CV proved to be a convenient approach to explain differences between industries and of changes in accuracy with editing effort.

We 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), Chipperfield and Chambers (2015), or are planning to use a similar approach (Bryant and Graham, 2013). One of the most challenging parts of this approach is to achieve good estimates of the parameters values of the error model. In our case study, we could rely on an audit sample for the simple unit types, where we expected large error rates, supplemented with data from the BR and with expert knowledge for the off-diagonal elements. At least we were able to obtain a model in which the effects of the independent variables that were found to explain the diagonal probabilities were in line with those expected by experts. Also certain large transition probabilities among industries (off-diagonal probabilities) could be explained by experts, such as the ‘confusion’ between wholesale and retail trade.

The weakest part of our parameterisation was our reliance on expert knowledge for the diagonal probabilities for the complex and most complex units. One way to improve on this might be to use a more systematic approach in combining information from different experts, for instance by the Delphi method (e.g., Wiśniowski et al., 2012). Another way would be to use a data base on editing information of the complex and most complex units in our BR. This editing information concerns approved requests to change the NACE code of enterprises, including the corresponding registration and application dates. Analysing the relative

frequencies of those classification changes as a function of the background variables of the enterprises will probably give an improved estimation of the parameters for the complex and most complex units. On the other hand, the problem remains that we are dealing with 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 practical reasons, the parameterisation was based on a limited number of sampling units. That means that the estimated parameters themselves are prone to uncertainty. This uncertainty in the (estimated) transition probabilities does not affect the true but unknown accuracy of the quarterly turnover estimates themselves, it only implies that we are uncertain about our accuracy estimates. We could estimate this uncertainty by repeating our (10,000) bootstraps for different values of our transition probabilities, and those different values are then obtained by sampling from the distribution of the estimated parameters. A disadvantage of this approach is that it would be computationally (very) intensive. Another idea is that within each of our 10,000 bootstraps we apply a slightly different version of our transition matrices. We could then compare the resulting accuracy estimates to the accuracy estimate as given in the present paper, and the difference between the two would give an impression of the sensitivity of the accuracy estimates for parameter uncertainty.

A further drawback of the parameterisation of the classification errors is that in our approach we used net errors that are implicitly corrected for the effects of editing. It may be more natural to distinguish explicitly between the probability of occurrence of an error and the probability to correct an error, given that it occurred. Then we would probably find that the probability of classification errors before editing increases over the whole domain with size and complexity of the unit with its effect depending on the NACE code, and the probability that a unit is checked and corrected for an error then depends solely on the editing strategy. In practice however at SN it is hard to obtain a raw data set without any cleaning for classification errors because we work with a central unit that maintains the BR, so it is difficult to estimate the 'raw' probability of occurrence of classification errors. For measurement errors, it is in practice much easier to obtain raw data, so there this strategy may be more feasible.

In modelling the classification errors, we introduced a number of assumptions. Unfortunately, many of these assumptions could not be tested with the data we had. The fact that we estimated different parts of the model separately also makes it difficult to assess the uncertainty in the estimated error probabilities. An interesting alternative approach might be to use latent class analysis to model classification errors. This would at least provide a single modelling framework, with the possibility of computing an overall measure of model fit and standard errors for the estimated parameters.

We have at least two challenges concerning the 'theoretical part' of our study. The first issue concerns the potential bias in our bootstrap estimates. We did find an alternative estimator that corrects for this bias, but this estimator has a large variance; we therefore used a combined estimator. Results without this bias correction were very similar to those with the bias correction, because the combined estimator only selected the bias-corrected estimator when the estimated bias was small. Studies with artificial data (not shown) however demonstrated that our combined estimator provides more robust results when computed over different sets of realisations of observed values given a set of true values. The bias-corrected and combined bootstrap estimators were derived here using an approximate analytical solution for the bias and variance due to classification errors, so they cannot be generalised to more difficult problems (considering, e.g., 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).

The second theoretical issue concerns our simulation of the editing scenarios. Unfortunately we cannot correctly simulate the exact accuracy outcomes in case of increased editing, because we do not know the true NACE code of the units. Turnover of business statistics can be very dependent on those true values, because the distribution of turnover is very skewed. In fact, for our study to be "fully exact" we would need to have both the raw data and the true data of NACE codes for the population of enterprises. Still, we are convinced that our approximation is good enough for the main patterns of changes in accuracy over the industries. We believe that our main finding that the pro rata scenario is no improvement compared to the fixed scenario holds, because the changes in accuracies with editing effort in the different industries are not independent of each other.

Our research has focused on developing a method for quantifying the accuracy of estimates due to classification errors. Given that some estimates are not accurate enough, one needs an effective and efficient strategy to improve those estimates. Unfortunately, efficient editing of classification variables is still an underdeveloped research area. In the application of Section 5, it was seen that increased selective editing actually led to less accurate estimates of turnover in some cases. Of the two editing strategies considered here, the fixed scenario gave the best overall results, but this is an inefficient strategy. An open question is therefore: which editing strategy would be more effective than the fixed strategy in improving the overall accuracy of car trade as a whole? Probably we need a strategy that selects additional units such that the bias for each industry is equal to or smaller than the starting point. This bias depends on the transition probabilities and on the corresponding turnover values per enterprise of the complexity classes. In addition, we would like to have an effective editing strategy for improving the RRMSE of specific industries that are currently not accurate enough.

A second, related, question is: is there an alternative to manual editing as a means to improve the accuracy of the estimates? This is for instance relevant for industry 45200. Manual clerical review is very labour intensive: going from relative editing

time 1 to 3, which would involve about 60 hours of additional editing, the overall RRMSE over the nine industries of car trade decreased from 0.35% to 0.24%, and from 12% to 6% for industry 45200 (under the 'fixed' scenario). Maybe a more effective approach would be to use internet robots (web scraping) to try to automatically classify the industry of all enterprises (ten Bosch and Windmeijer, 2014). A third question is: can we use the results to numerically correct the estimates for the bias of the current estimates per industry, $\hat{Y}_h$ compared to Y_h? This would be an interesting 'macro-level' alternative to editing the individual NACE codes. The answer depends, among others, on the variance of our bias estimates and the stability of the estimates over time. It is a point for future research.

The long-term aim of our research is to obtain a practical method that can be implemented in production to estimate the accuracy of register-based estimates. There are a number of issues still to be solved before this can be achieved. The first issue is that we are interested in estimating the accuracy of the growth rates, in addition to those of the level estimates. To that end we need to understand more about the time-relations between classification errors. In fact, in the case study discussed here, we applied the bootstrap method separately to each quarter, so the results do not provide any information on the accuracy of estimated growth rates. A second important issue is: how can we obtain measurements during the production process to compute the accuracy without having to rely on audit samples? We already tried to make use of the yearly updates in the NACE codes within the BR; probably we can make use of more information that is available in our BR about causes of changes in NACE code. This resembles the use of paradata in social surveys (Kreuter, 2013). Another option may be to collect information during the regular editing process. This resembles ideas to 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).

A third issue is: how can we extend the procedure to other industries? It is practically impossible to estimate the transition probabilities for the complete set of industries. At SN about 325 base cells are distinguished, so we would obtain a 325 × 325 matrix $\mathbf{P}_i$. Our ideas now are that we could estimate a separate matrix for each economic sector (retail trade, car trade, wholesale trade, manufacturing industry, etc.). For each sector-matrix we could include all industries within that sector supplemented by those NACE codes among which misclassifications occur most often and one remaining "other activities" category. For the other activities category we could use a hierarchical approach: we could estimate transitions among the 1-digit level NACE codes (0–9). That would yield a 10 × 10 1-digit level-matrix. The relative frequencies within that matrix could be estimated from the relative frequencies of yearly transitions of NACE codes in the BR.

A final issue is: how can we extend the procedure to other error sources and their interactions? We could potentially use a similar approach of modelling the errors, but collecting information to parameterise those models will be the challenging part.

References

Berka, C., Humer, S., Moser, M., Lenk, M., Rechta, H., and Schwerer, E. (2012), Combination of Evidence from Multiple Administrative Data Sources: Quality Assessment of the Austrian Register-Based Census 2011. *Statistica Neerlandica* **66**, 18–33.

Biemer, P. and Lyberg, L. (2003), *Introduction to Survey Quality*. John Wiley & Sons.

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

Bryant, J. and Graham, P. (2013), A Bayesian Method for Deriving Population Statistics from Multiple Imperfect Data Sources. Paper presented at the World Statistics Congress, August 25–30, Hong Kong. Available at http://www.statistics.gov.hk/wsc/IPS027-P4-S.pdf (accessed December 2013).

Burger, J., van Delden, A., and Scholtus, S. (2015), Sensitivity of Mixed-Source Statistics to Classification Errors. *Journal of Official Statistics* (forthcoming).

Chipperfield, J. and Chambers, R. (2015), Using the Bootstrap to Analyse Binary Data Obtained Via Probabilistic Linkage. *Journal of Official Statistics* (forthcoming).

Christensen, J.L. (2008), Questioning the Precision of Statistical Classification of Industries. Paper presented at the Conference on entrepreneurship and innovation – organizations, institutions, systems and regions, Copenhagen, CBS, Denmark, June 17–20, 2008.

Cochran, W.G. (1977), *Sampling Techniques* (3rd Edition). John Wiley & Sons, New York.

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

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

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2009), *Survey Methodology* (2nd Edition). John Wiley & Sons, Inc.

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

Ilves, M. and Laitila, T. (2009), Probability-Sampling Approach to Editing. *Austrian Journal of Statistics* **38**, 171–182.

Kreuter, F. (2013), *Improving Surveys with Paradata. Analytic Use of Process Information*. John Wiley & Sons.

Kuha, J. and Skinner, C. (1997), Categorical Data Analysis and Misclassification. In: Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwarz, and Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, pp. 633–670.

Latouche, M. and Berthelot, J.-M. (1992), Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics* **8**, 389–400.

Lumme, S., Sund, R., Leyland, A.H., and Keskmäki, I. (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. Published online 18 February 2015.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models* (2nd Edition). Chapman & Hall, London.

Stoer, J. and Bulirsch, R. (2002), *Introduction to Numerical Analysis* (3rd Edition). Springer, New York.

ten Bosch, O. and Windmeijer, D. (2014), On the Use of Internet Robots for Official Statistics. MSIS, 14–16 April 2014, Dublin. Available at http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf.

van Delden, A. and de Wolf, P.P. (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.

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

Wiśniowski, A., Keilman, N., Bijak, J., Christiansen, S., Forster, J.J., Smith, P.W.F., and Raymer, J. (2012), Augmenting Migration Statistics with Expert Knowledge. Norface migration Discussion Paper 2012–05. Accessed at www.norface-migration.org/publ_uploads/NDP_05_12.pdf.

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.

Zhang, L.-C. (2012b), On the Accuracy of Register-Based Census Employment Statistics. Paper presented at the European Conference on Quality in Official Statistics (Q2012), May 30–June 1, Athens.

Zhang, L.-C. (2014), Data integration. *The Survey Statistician* **70**, 15–24.

Appendix A

A.1 Bias-corrected bootstrap estimates of bias

We use the notation that was introduced in Section 2. In addition, let $\boldsymbol{a}_i$ denote the vector $(a_{1i}, \ldots, a_{H+1,i})^T$ that contains the values of the indicator $a_{hi} = I(s_i = h)$ 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 $E(\hat{\boldsymbol{a}}_i^* | \hat{\boldsymbol{a}}_i) = \mathbf{P}_i^T \hat{\boldsymbol{a}}_i$. Denote the vectors with the true, observed and bootstrap values for the total turnover per industry as, respectively, $\boldsymbol{y} = \sum_{i=1}^N \boldsymbol{a}_i y_i$, $\hat{\boldsymbol{y}} = \sum_{i=1}^N \hat{\boldsymbol{a}}_i y_i$, and $\hat{\boldsymbol{y}}^* = \sum_{i=1}^N \hat{\boldsymbol{a}}_i^* y_i$. Using a similar argument as 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 $\boldsymbol{y}$:

$$B(\hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}) - \boldsymbol{y} = \sum_{i=1}^{N} (\mathbf{P}_i^T - \mathbf{I}) \boldsymbol{a}_i y_i, \tag{26}$$

$$V(\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \{\mathrm{diag}(\mathbf{P}_i^T \boldsymbol{a}_i y_i^2) - \mathbf{P}_i^T \, \mathrm{diag}(\boldsymbol{a}_i y_i^2) \, \mathbf{P}_i\}, \tag{27}$$

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 $R \to \infty$ in expressions (4) and (5), we would obtain:

$$B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = E(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}} = \sum_{i=1}^{N} (\mathbf{P}_i^T - \mathbf{I}) \hat{\boldsymbol{a}}_i y_i, \tag{28}$$

$$V(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \{\mathrm{diag}(\mathbf{P}_i^T \hat{\boldsymbol{a}}_i y_i^2) - \mathbf{P}_i^T \, \mathrm{diag}(\hat{\boldsymbol{a}}_i y_i^2) \, \mathbf{P}_i\}; \tag{29}$$

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 (28) and (29) is over all units in the BR, including all industries outside car trade.]

Turning first to the bias (we will return to the variance in Appendix A.4), we see that $E\{B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}})\} = \sum_{i=1}^N \mathbf{P}_i^T (\mathbf{P}_i^T - \mathbf{I}) \boldsymbol{a}_i y_i$. This implies that $B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}})$ is biased as an estimator for $B(\hat{\boldsymbol{y}})$; the same follows for $\hat{B}_R^*(\hat{Y}_h)$ based on a finite number of replications.

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

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

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

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

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

A.2 Practical issues in the computation of the bias correction

In our case study the population consisted of a limited number of probability classes (PCs) with the same transition matrix. We can use this to compute $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{z}}^*$ 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 $\mathbf{Q}_k = (\mathbf{P}_k^T)^{-1}$. Now, $\hat{\boldsymbol{z}}$ can be computed according to:

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

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, $\hat{\boldsymbol{z}}^*$ can be computed as $\hat{\boldsymbol{z}}^* = \sum_{k=1}^K \hat{\boldsymbol{z}}_k^* \equiv \sum_{k=1}^K \mathbf{Q}_k \hat{\boldsymbol{y}}_k^*$, with $\hat{\boldsymbol{y}}_k^* = \sum_{i\in U_k} \hat{\boldsymbol{a}}_i^* y_i$. This means that in each bootstrap replication we can compute $\hat{\boldsymbol{y}}_{kr}^*$ and $\hat{\boldsymbol{z}}_r^* = \sum_{k=1}^K \mathbf{Q}_k \hat{\boldsymbol{y}}_{kr}^*$. Using R bootstrap replications, we can estimate the bias of the vector $\hat{\boldsymbol{y}}$ by $\hat{B}_R(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}}) = R^{-1} \sum_{r=1}^R \hat{\boldsymbol{z}}_r^* - \hat{\boldsymbol{z}}$.

Note that the $(H+1)^{\text{th}}$ elements of the vectors $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, etc. contain the totals for all industries outside the subset of target industries. It is in fact possible to compute $\hat{B}_R(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}})$ without knowing the total turnover of other activities, which is very useful in practice. To see this, we rewrite $\hat{B}_R(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}})$ as follows:

$$\begin{aligned}
\hat{B}_R(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}}) &= \frac{1}{R}\sum_{r=1}^R \sum_{k=1}^K \boldsymbol{Q}_k \hat{\boldsymbol{y}}_{kr}^* - \sum_{k=1}^K \boldsymbol{Q}_k \hat{\boldsymbol{y}}_k \\
&= \sum_{k=1}^K \boldsymbol{Q}_k \left(\frac{1}{R}\sum_{r=1}^R \hat{\boldsymbol{y}}_{kr}^* - \hat{\boldsymbol{y}}_k \right) \\
&= \sum_{k=1}^K \boldsymbol{Q}_k \left(\frac{1}{R}\sum_{r=1}^R \hat{\boldsymbol{e}}_{kr}^* \right),
\end{aligned}$$

where $\hat{\boldsymbol{e}}_{kr}^* = \hat{\boldsymbol{y}}_{kr}^* - \hat{\boldsymbol{y}}_k$.

We can define transition components “Inflow from the target set” (IT), “Inflow from other activities” (IO), and “Outflow” (O) for each element of $\hat{\boldsymbol{y}}_{kr}^*$, similarly to Section 2.1 but now separately for each PC. We collect these transition components into vectors $\hat{\boldsymbol{y}}_{kr}^{*(j)}$, with $j \in \{IT, IO, O\}$. Now, analogously to expression (6) it holds that:

$$\hat{\boldsymbol{e}}^{*}_{kr} = \hat{\boldsymbol{y}}^{*(IT)}_{kr} + \hat{\boldsymbol{y}}^{*(IO)}_{kr} + \hat{\boldsymbol{y}}^{*(O)}_{kr}. \qquad (30)$$

Hence, to obtain $\hat{B}_R(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}})$ we only need to know the amounts of turnover that 'move' between industries in each replication, not the total values.

A particular practical problem arose in our case study in evaluating expression (30) for the last element of the vector, which corresponds to 'industry' $H+1$. Here, the first term in (30) represents the total amount of turnover in PC k which moves from the target set to other industries, which is known exactly. Also, the second term is zero by definition, because we do not distinguish between activities within the set of other industries. The third term ($\hat{Y}^{*(O)}_{H+1,kr}$, say) represents the negative value of the total amount of turnover in PC k which moves from other industries to the target set: $\hat{Y}^{*(O)}_{H+1,kr} = -\left(\hat{Y}_{H+1,kr} - \hat{Y}^{*}_{H+1,H+1,kr}\right)$. In our case study, $\hat{Y}^{*(O)}_{H+1,kr}$ was unknown because the turnover observed in other activities was drawn from a log-normal distribution at a less granular level than that of the PCs (see Section 2.2). Hence, we had $\hat{Y}^{*(O)}_{H+1,r} = \hat{Y}^{*(O)}_{H+1,A,r} + \hat{Y}^{*(O)}_{H+1,B,r}$, where the first term refers to the PCs in the bottom two rows of Figure 4.1.3 and the second term to the remaining nine PCs. Since these terms were not differentiated further during the simulations, we approximated the last element of $\hat{\boldsymbol{y}}^{*(O)}_{kr}$ by

$$\hat{Y}^{*(O)}_{H+1,kr} \approx \begin{cases} \hat{Y}^{*(O)}_{H+1,A,r}/K_A & \text{if } k \in A \\ \hat{Y}^{*(O)}_{H+1,B,r}/K_B & \text{if } k \in B \end{cases}$$

with $K_A = 2$ and $K_B = 9$ the number of PCs within the groups A and B.

A final point is that that in order to compute $\mathbf{Q}_k = (\mathbf{P}_k^T)^{-1}$ we need to know the $(H+1)^{\text{th}}$ diagonal element of $\mathbf{P}_k$: the probability that a unit that is classified outside the target set actually belongs to that group. An estimate of this probability can be obtained indirectly from the audit sample, as we will illustrate for our case study. From the estimated log-linear model for the off-diagonal elements $\hat{N}_{gh,model}$ we obtained an estimated number of 625.7 units that are wrongly classified in car trade (see Table 4.2.2). The total number of units with other activities in the BR was 1,000,484 around the time that the audit sample was taken, yielding a proportion of 0.000625 of the units with other activities that is wrongly classified in car trade. That means that a fraction of 1 – 0.000625 = 0.999375 is correctly classified in other activities. Strictly speaking we should have computed this value for each PC separately, but the outcome of $\mathbf{Q}_k$ is not very sensitive to small deviations from 1, so we ignored this.

A.3 Adjusted bias correction for increased accuracy

The corrected bootstrap estimator for the bias $B(\hat{\boldsymbol{z}}^*|\hat{\boldsymbol{z}})$ in Appendix A.1 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 $\mathrm{cond}(\mathbf{P}_k^T) = \|\mathbf{P}_k^T\|\|\mathbf{Q}_k\|$ can become much larger than 1. Here, the symbol $\|.\|$ denotes a matrix norm. Since $\widehat{\boldsymbol{y}}_k^* = \mathbf{P}_k^T \widehat{\boldsymbol{z}}_k^*$, it follows from a standard result in numerical analysis that

$$|rel.change(\widehat{\boldsymbol{z}}_k^*)| \leq cond(\boldsymbol{P}_k^T) \times |rel.change(\widehat{\boldsymbol{y}}_k^*)|,$$

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

In this section, we propose an alternative correction method by minimising the mean square error of the estimated bias. We restrict attention to the situation of our case study, where the units are grouped into PCs with equal transition matrices.

Denote the original and the corrected bootstrap estimators for $B(\widehat{\boldsymbol{y}})$ as

$$\widehat{\boldsymbol{B}}_0 \equiv B(\widehat{\boldsymbol{y}}^*|\widehat{\boldsymbol{y}}) \equiv \sum_{k=1}^{K} \widehat{\boldsymbol{B}}_{0k} \left[= \sum_{k=1}^{K} (\boldsymbol{P}_k^T - \boldsymbol{I})\widehat{\boldsymbol{y}}_k \right],$$
$$\widehat{\boldsymbol{B}}_1 \equiv B(\widehat{\boldsymbol{z}}^*|\widehat{\boldsymbol{z}}) \equiv \sum_{k=1}^{K} \widehat{\boldsymbol{B}}_{1k} \left[= \sum_{k=1}^{K} \boldsymbol{Q}_k \widehat{\boldsymbol{B}}_{0k} \right].$$

We now consider a combined estimator for the bias of $B(\widehat{\boldsymbol{y}})$, by taking a convex combination of $\widehat{\mathbf{B}}_{0k}$ and $\widehat{\mathbf{B}}_{1k}$ for each PC:

$$\widehat{\boldsymbol{B}}_\lambda \equiv \sum_{k=1}^{K} \widehat{\boldsymbol{B}}_{\lambda_k k} \equiv \sum_{k=1}^{K} \{\lambda_k \widehat{\boldsymbol{B}}_{1k} + (1-\lambda_k)\widehat{\boldsymbol{B}}_{0k}\},$$

where each $\lambda_k \in [0,1]$. Notice that $\lambda_k = 0$ yields $\widehat{\mathbf{B}}_{0k}$ and $\lambda_k = 1$ yields $\widehat{\mathbf{B}}_{1k}$. Within this family of combined estimators we seek to find the most accurate estimator of the bias of $B(\widehat{\boldsymbol{y}})$, as a compromise between the unbiasedness of $\widehat{\mathbf{B}}_1$ and the higher precision of $\widehat{\mathbf{B}}_0$. We will first derive the mean squared error of $\widehat{\mathbf{B}}_\lambda$ and then find its minimum.

Using $\widehat{\mathbf{B}}_{1k} = \mathbf{Q}_k \widehat{\mathbf{B}}_{0k}$ in the above equation gives an alternative expression for $\widehat{\mathbf{B}}_{\lambda_k k}$:

$$\widehat{\boldsymbol{B}}_{\lambda_k k} = \lambda_k \boldsymbol{Q}_k \widehat{\boldsymbol{B}}_{0k} + (1-\lambda_k)\widehat{\boldsymbol{B}}_{0k} = \{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}\widehat{\boldsymbol{B}}_{0k}.$$

This then gives

$$\begin{aligned} E(\widehat{\boldsymbol{B}}_{\lambda_k k}) &= \{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}E(\widehat{\boldsymbol{B}}_{0k}) \\ &= \{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}\boldsymbol{P}_k^T\boldsymbol{B}_k \\ &= \{\boldsymbol{P}_k^T + \lambda_k(\boldsymbol{I} - \boldsymbol{P}_k^T)\}\boldsymbol{B}_k, \\ V(\widehat{\boldsymbol{B}}_{\lambda_k k}) &= \{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}V(\widehat{\boldsymbol{B}}_{0k})\{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}^T \\ &= \{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T\{\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})\}^T, \end{aligned}$$

where $\mathbf{B}_k = B(\widehat{\boldsymbol{y}}_k)$ stands for the true bias and $\boldsymbol{\Omega}_k = V(\widehat{\boldsymbol{y}}_k)$ for the true variance-covariance matrix of $\widehat{\boldsymbol{y}}_k$. From (26) and (27) we have: $\mathbf{B}_k = (\mathbf{P}_k^T - \mathbf{I})\boldsymbol{y}_k$ and $\boldsymbol{\Omega}_k = \mathrm{diag}(\mathbf{P}_k^T \boldsymbol{c}_k) - \mathbf{P}_k^T \,\mathrm{diag}(\boldsymbol{c}_k)\,\mathbf{P}_k$, with $\boldsymbol{c}_k = \sum_{i \in U_k} \boldsymbol{a}_i y_i^2$ the vector of sums of squared turnover values for units in PC k. To obtain the expression for $\boldsymbol{\Omega}_k$, we used

that $V(\hat{\boldsymbol{y}}) = \sum_{k=1}^{K} V(\hat{\boldsymbol{y}}_k)$ since classification errors are mutually independent across units.

The mean squared error of the estimated bias of the h^{th} element of $\hat{\boldsymbol{y}}_k$ equals the sum of the squared value of the h^{th} element of $B(\widehat{\mathbf{B}}_{\lambda_k k}) = E(\widehat{\mathbf{B}}_{\lambda_k k}) - \mathbf{B}_k = (1-\lambda_k)(\mathbf{P}_k^T - \mathbf{I})\mathbf{B}_k$ and the h^{th} diagonal element of $V(\widehat{\mathbf{B}}_{\lambda_k k})$:

$$mse\left\{\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)_h\right\} = \left\{B\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\}_h^2 + \left\{V\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\}_{hh}.$$

The total mean squared error within PC k across all target industries is then given by:

$$\begin{aligned} mse\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right) &\equiv \sum_{h=1}^{H} mse\left\{\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)_h\right\} \\ &= \left\{B\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\}_{1:H}^{T}\left\{B\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\}_{1:H} + tr_{1:H}\left\{V\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\} \\ &= (1-\lambda_k)^2\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{B}_k\}_{1:H}^{T}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{B}_k\}_{1:H} \\ &\quad + tr_{1:H}\{[\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})](\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T[\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})]^T\}. \end{aligned}$$

where the subscript $1{:}H$ means that only the first H elements should be taken, thus $\text{tr}_{1:H}$ is the trace over the first H elements.

We now aim to find the value λ_k that minimises $\text{mse}\left(\widehat{\mathbf{B}}_{\lambda_k k}\right)$. We consider that estimator as the optimal estimator for the bias of $\hat{\boldsymbol{y}}_k$ and the sum of these optimal estimators over the PCs as the optimal estimator for the bias of $\hat{\boldsymbol{y}}$.

Using the fact that $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA})$, we can rewrite $\text{tr}_{1:H}\left\{V\left(\widehat{\mathbf{B}}_{\lambda_k k}\right)\right\}$ as

$$\begin{aligned} tr_{1:H}\left\{V\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right)\right\} &= tr_{1:H}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T[\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})]^T[\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k - \boldsymbol{I})]\} \\ &= tr_{1:H}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T[\boldsymbol{I} + \lambda_k(\boldsymbol{Q}_k + \boldsymbol{Q}_k^T - 2\boldsymbol{I}) \\ &\quad + \lambda_k^2(\boldsymbol{Q}_k - \boldsymbol{I})^T(\boldsymbol{Q}_k - \boldsymbol{I})]\}. \end{aligned}$$

That means that $\text{mse}\left(\widehat{\mathbf{B}}_{\lambda_k k}\right)$ can be expressed as follows:

$$\begin{aligned} mse\left(\widehat{\boldsymbol{B}}_{\lambda_k k}\right) &= m_1(1-\lambda_k)^2 + (m_2 - 2m_3 + m_4)\lambda_k^2 + 2(m_3 - m_4)\lambda_k + m_4, \\ m_1 &= \{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{B}_k\}_{1:H}^{T}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{B}_k\}_{1:H}, \\ m_2 &= tr_{1:H}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T\boldsymbol{Q}_k^T\boldsymbol{Q}_k\}, \\ m_3 &= tr_{1:H}\left\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T\frac{1}{2}(\boldsymbol{Q}_k + \boldsymbol{Q}_k^T)\right\}, \\ m_4 &= tr_{1:H}\{(\boldsymbol{P}_k^T - \boldsymbol{I})\boldsymbol{\Omega}_k(\boldsymbol{P}_k^T - \boldsymbol{I})^T\}. \end{aligned}$$

Notice that m_1, m_2, m_3 and m_4 are scalars. Setting the derivative of $\text{mse}\left(\widehat{\mathbf{B}}_{\lambda_k k}\right)$ with respect to λ_k equal to 0 yields

$$\tilde{\lambda}_k = \frac{m_1 - m_3 + m_4}{m_1 + m_2 - 2m_3 + m_4}.$$

Now, taking into account that $\lambda_k \in [0,1]$, we find for the optimal weight that minimises $\text{mse}\left(\widehat{\mathbf{B}}_{\lambda_k k}\right)$:

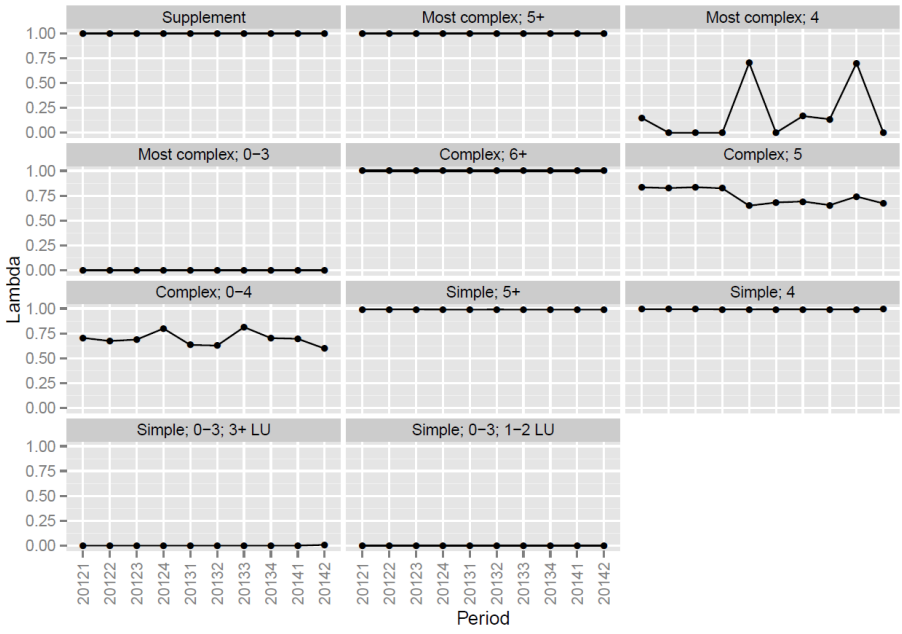
$$\lambda_k = max\{0, min[1, \tilde{\lambda}_k]\}.$$

A practical problem in calculating $\tilde{\lambda}_k$ is that it depends on the unknown true values for the bias $\mathbf{B}_k$ and the variance $\boldsymbol{\Omega}_k$ of $\hat{\boldsymbol{y}}_k$. In order to estimate $\tilde{\lambda}_k$ we can replace the true value $\boldsymbol{\Omega}_k$ by its bootstrap estimator $V(\hat{\boldsymbol{y}}_k^*|\hat{\boldsymbol{y}}_k)$. Finding a good value for the bias $\mathbf{B}_k$, which only occurs in m_1, is more difficult since $\widehat{\mathbf{B}}_{0k} = B(\hat{\boldsymbol{y}}_k^*|\hat{\boldsymbol{y}}_k)$ in itself is also biased and reducing that bias was the objective of using the combined estimator. We propose the following iterative approach:

1. Start with $\widehat{\mathbf{B}}_k = \widehat{\mathbf{B}}_{0k} = B(\hat{\mathbf{y}}_k^*|\hat{\mathbf{y}}_k)$, $\widehat{\mathbf{\Omega}}_k = V(\hat{\mathbf{y}}_k^*|\hat{\mathbf{y}}_k)$ and $\lambda_k = 0$.
2. Determine the optimal λ_k given $\widehat{\mathbf{B}}_k$ and $\widehat{\mathbf{\Omega}}_k$.
3. When the value of λ_k is stable, e.g., when the difference with the previous value of λ_k in step 2 is smaller than 10^{-6}, then stop. Otherwise, set $\widehat{\mathbf{B}}_k = \widehat{\mathbf{B}}_{\lambda_k k}$ and return to step 2.

We found that this algorithm converges already after a few cycles. Tests on simulated data suggest that the resulting value for λ_k is close to the optimal value that would have been found when $\mathbf{B}_k$ and $\mathbf{\Omega}_k$ were known. Notice that we start with $\lambda_k = 0$, because we know that $\widehat{\mathbf{B}}_{0k}$ yields more stable estimates than $\widehat{\mathbf{B}}_{1k}$.

A.3.1 Values for λ_k per PC in the case study for car trade



There are two main findings concerning the estimated values for λ_k in the case study for car trade (Figure A.3.1):

– A value $\lambda_k \approx 1$ (yielding the bias-corrected estimator $\widehat{\mathbf{B}}_{1k}$) was selected in those PCs with a small probability of classification errors, thus for the classes with a small bias, since for those units $\hat{\boldsymbol{a}}_i = \boldsymbol{a}_i$ with high probability. On the other hand, $\lambda_k \approx 0$ was selected for those PCs with a large probability of classification errors.

– PCs with equal probability of classification errors, e.g., "most complex SC 4", "complex SC 5" and "simple SC 5+" all with diagonal transition probabilities of 0.98 (see Figure 4.1.3), had increasing values for λ_k with increasing size class (in this example: SC 4, SC 5, SC 5+). Mathematically, this was explained by the fact that, with increasing size class, the variance $\mathbf{\Omega}_k$ increased but the bias $\mathbf{B}_k$ increased even more. It can be shown that the latter result stems from two effects. Firstly, $\hat{\boldsymbol{e}}_{kr}^*$ contained zero values for these PCs in a substantial subset of the replications, due to the high diagonal probabilities and the small number of units in these classes. This fraction of zero values decreased with size class. Secondly, the turnover per enterprise increased with size class. Analyses showed that the net result of these two effects was that the value of $\mathbf{B}_k$ increased faster

than the variance $\boldsymbol{\Omega}_k$ (for these three PCs). An implication of this finding is that it would have been better to combine the PCs with an equal probability on classification errors, because we would then avoid the situation with zero turnover values and thus λ_k would have been larger, allowing to correct for the bias.

A.4 Bias-corrected bootstrap estimates of variance

Likewise to the bias, the bootstrap estimator of the variance is also biased. In the current section we derive a formula for this bias, explain how it can be corrected and show that this bias is likely to be small in practice. Again, we restrict attention to the situation of our case study.

As mentioned in the previous subsection, the true variance-covariance matrix of $\hat{\boldsymbol{y}}$ equals

$$V(\hat{\boldsymbol{y}}) = \sum_{k=1}^{K} V(\hat{\boldsymbol{y}}_k) = \sum_{k=1}^{K} \{diag(\boldsymbol{P}_k^T \boldsymbol{c}_k) - \boldsymbol{P}_k^T \, diag(\boldsymbol{c}_k) \, \boldsymbol{P}_k\},$$

with $\boldsymbol{c}_k = \sum_{i \in U_k} \boldsymbol{a}_i y_i^2$. In particular, the variance of $\hat{Y}_h = (\hat{\boldsymbol{y}})_h$ can be written as:

$$V(\hat{Y}_h) = \sum_{k=1}^{K} \sum_{g=1}^{H+1} p_{ghk}(1 - p_{ghk}) C_{gk},$$

where p_{ghk} is the $(g,h)^{\text{th}}$ element of $\mathbf{P}_k$ and C_{gk} the g^{th} element of $\boldsymbol{c}_k$.

The bootstrap estimator of $V(\hat{\boldsymbol{y}})$ for an infinite number of replications equals $V(\hat{\boldsymbol{y}}^*|\hat{\boldsymbol{y}}) = \sum_{k=1}^{K} V(\hat{\boldsymbol{y}}_k^*|\hat{\boldsymbol{y}}_k)$. For the estimated variance of $\hat{Y}_h$ this gives:

$$V(\hat{Y}_h^*|\hat{\boldsymbol{y}}) = \sum_{k=1}^{K} \sum_{g=1}^{H+1} p_{ghk}(1 - p_{ghk}) \hat{C}_{gk},$$

where $\hat{C}_{gk}$ is the g^{th} element of $\hat{\boldsymbol{c}}_k = \sum_{i \in U_k} \hat{\boldsymbol{a}}_i y_i^2$ [cf. expression (29)]. It follows directly that the bias of the bootstrap estimator can be expressed in terms of the estimated sum of squared turnover values:

$$\begin{aligned} B\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\} &= E\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\} - V(\hat{Y}_h) \\ &= \sum_{k=1}^{K} \sum_{g=1}^{H+1} p_{ghk}(1 - p_{ghk}) E(\hat{C}_{gk}) - \sum_{k=1}^{K} \sum_{g=1}^{H+1} p_{ghk}(1 - p_{ghk}) C_{gk} \\ &= \sum_{k=1}^{K} \sum_{g=1}^{H+1} p_{ghk}(1 - p_{ghk}) B(\hat{C}_{gk}). \end{aligned}$$

So, in order to correct the bias in $V(\hat{Y}_h^*|\hat{\boldsymbol{y}})$, it is sufficient to have an unbiased estimator of $B(\hat{\boldsymbol{c}}_k)$. Since $B(\hat{\boldsymbol{c}}_k)$ has the same form as $B(\hat{\boldsymbol{y}}_k)$ – both estimators are linear functions of $\hat{\boldsymbol{a}}_i$ and we assume that $\hat{\boldsymbol{a}}_i$ is the only source of errors – such an unbiased estimator could be obtained analogously to the bias-corrected estimator of $B(\hat{\boldsymbol{y}}_k)$ from Appendix A.2. Or, to obtain a more stable estimate, we could construct a combined bootstrap estimator of $B(\hat{\boldsymbol{c}}_k)$ by minimizing its mean squared error, as we

did for $B(\hat{\boldsymbol{y}}_k)$ in Appendix A.3. However, if the size of the bias in $V(\hat{Y}_h^*|\hat{\boldsymbol{y}})$ is expected to be small, it is probably better leave out this bias correction.

We will now look into the conditions under which the bias of $B\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\}$ is small. The relative bias equals:

$$\begin{aligned} RB\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\} &= \frac{B\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\}}{V(\hat{Y}_h)} \\ &= \frac{\sum_{k=1}^{K}\sum_{g=1}^{H+1} p_{ghk}(1-p_{ghk})C_{gk}\cdot RB(\hat{C}_{gk})}{\sum_{k=1}^{K}\sum_{g=1}^{H+1} p_{ghk}(1-p_{ghk})C_{gk}} \\ &= \sum_{k=1}^{K}\sum_{g=1}^{H+1} w_{ghk} RB(\hat{C}_{gk}), \end{aligned}$$

where $RB(\hat{C}_{gk}) = B(\hat{C}_{gk})/C_{gk}$ is the relative bias in $\hat{C}_{gk}$ and the weights w_{ghk} have the property that $\sum_{k=1}^{K}\sum_{g=1}^{H+1} w_{ghk} = 1$. This means that the absolute value of the relative bias $|RB\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\}|$ is at most equal to the maximum of $|RB(\hat{C}_{1k})|, \ldots, |RB(\hat{C}_{H+1,k})|$. In practice $RB\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\}$ does not attain this maximum, as it is impossible for all terms in the above sum to have the same sign (see below).

Similar to $B(\hat{\boldsymbol{y}}_k) = (\mathbf{P}_k^T - \mathbf{I})\boldsymbol{y}_k$ it holds that $B(\hat{\boldsymbol{c}}_k) = (\mathbf{P}_k^T - \mathbf{I})\boldsymbol{c}_k$. In particular:

$$B(\hat{C}_{gk}) = \sum_{l=1}^{H+1} p_{lgk} C_{lk} - C_{gk}.$$

The sum of $B(\hat{C}_{gk})$ over all industries equals zero:

$$\sum_{g=1}^{H+1} B(\hat{C}_{gk}) = \sum_{g=1}^{H+1}\sum_{l=1}^{H+1} p_{lgk} C_{lk} - \sum_{g=1}^{H+1} C_{gk} = \sum_{l=1}^{H+1} C_{lk} \sum_{g=1}^{H+1} p_{lgk} - \sum_{g=1}^{H+1} C_{gk} = 0,$$

since $\sum_{g=1}^{H+1} p_{lgk} = 1$ for each row in $\mathbf{P}_k$. Note that this only holds because we assume that classification errors are the only errors that occur. Hence, both positive and negative values of $B(\hat{C}_{gk})$ occur in the above expression for $B\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\}$.

Now we introduce the following notation:

$$\begin{aligned} \pi_{ghk} &= p_{ghk}(1-p_{ghk}), \\ \bar{\pi}_{hk} &= \frac{1}{H+1}\sum_{g=1}^{H+1} \pi_{ghk}, \\ \sigma_{\pi,hk}^2 &= \frac{1}{H+1}\sum_{g=1}^{H+1} \{\pi_{ghk} - \bar{\pi}_{hk}\}^2, \\ \sigma_{BC,k}^2 &= \frac{1}{H+1}\sum_{g=1}^{H+1} \{B(\hat{C}_{gk})\}^2. \end{aligned}$$

Using $\sum_{g=1}^{H+1} B(\hat{C}_{gk}) = 0$ we find:

$$\begin{aligned} B\{V(\hat{Y}_h^*|\hat{\boldsymbol{y}})\} &= \sum_{k=1}^{K}\sum_{g=1}^{H+1} \pi_{ghk} B(\hat{C}_{gk}) \\ &= \sum_{k=1}^{K}\sum_{g=1}^{H+1} (\pi_{ghk} - \bar{\pi}_{hk}) B(\hat{C}_{gk}) \end{aligned}$$

$$= (H+1)\sum_{k=1}^{K} \rho_{hk}\sigma_{\pi,hk}\sigma_{BC,k}$$

where ρ_{hk} denotes the correlation between π_{ghk} and $B(\hat{C}_{gk})$ with h and k held fixed. This result shows that the bias in $V(\hat{Y}_h^*|\hat{\boldsymbol{y}})$ as an estimator for $V(\hat{Y}_h)$ is small when for each PC k at least one of the following conditions holds:

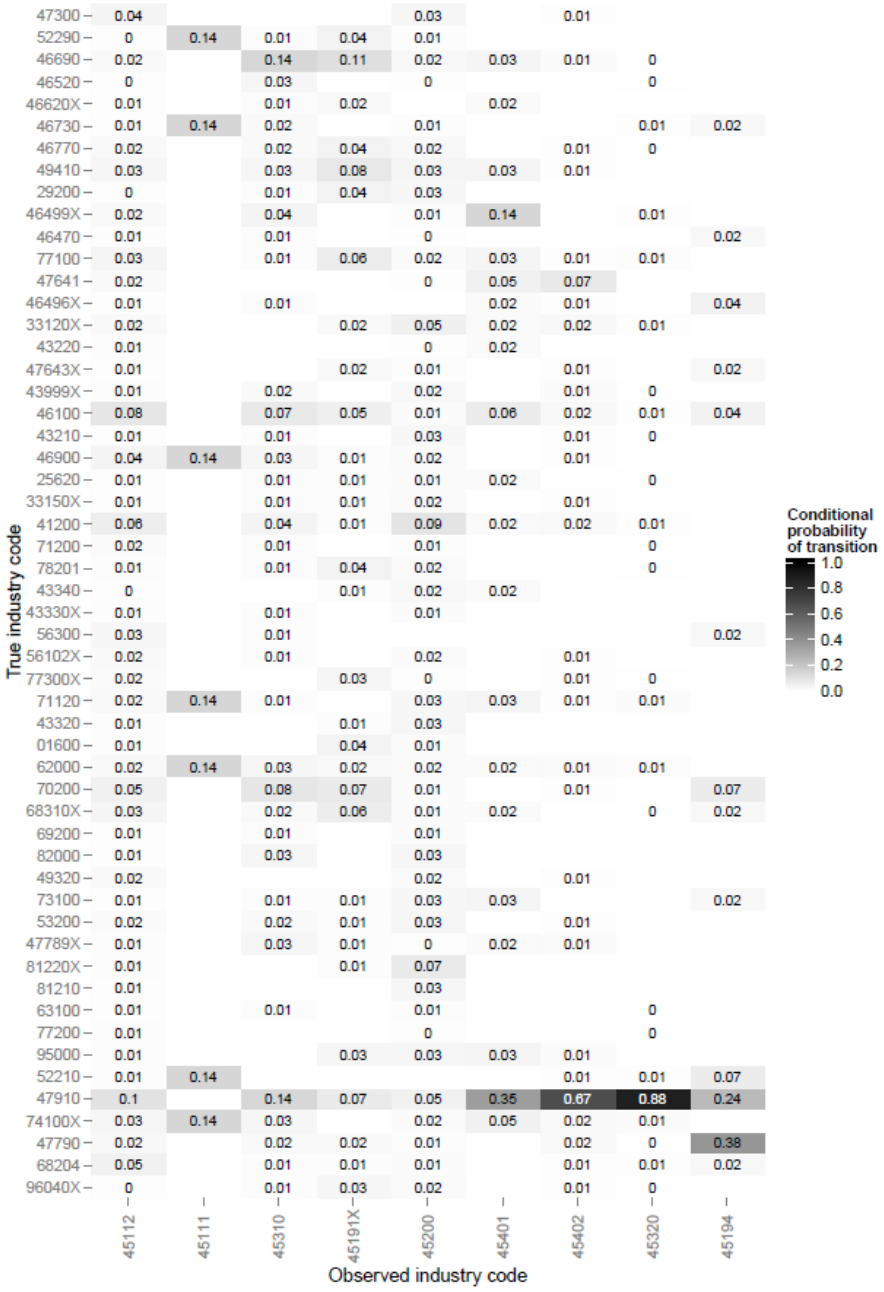
- there is little correlation between π_{ghk} and $B(\hat{C}_{gk})$, i.e., $|\rho_{hk}| \ll 1$;
- there is little variation in the values $\pi_{1hk}, \dots, \pi_{H+1,hk}$, i.e., $\sigma_{\pi,hk}$ is small;
- there is little variation in the values $B(\hat{C}_{1k}), \dots, B(\hat{C}_{H+1,k})$, i.e., $\sigma_{BC,k}$ is small.

The first condition amounts to having little correlation between the ordered set of transition probabilities $\{p_{1hk}, \dots, p_{H+1,hk}\}$ and the values in the ordered set $\{B(\hat{C}_{1k}), \dots, B(\hat{C}_{H+1,k})\}$, which is likely to hold in practice because the first set depends on the industry h whereas the second does not.

In fact, for the bias in $V(\hat{Y}_h^*|\hat{\boldsymbol{y}})$ to be small, it is sufficient in practice that the above conditions hold for those PCs that have the largest share of turnover. Moreover, positive and negative values for the bias contributions per PC can also compensate each other in the above expression.

Appendix B

B.1.1 Conditional probabilities of units observed in car trade but with true industry outside car trade. Each column adds up to 1. Industries outside car trade are in descending order of average turnover.



Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
Empty cel	Not applicable
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information