Effect of classification errors on the accuracy of business statistics

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

Abstract

National statistical institutes increasingly use administrative data and mixed sources instead of traditional sample surveys. As a result, non-sampling errors become more prominent. In the current paper we estimate the effect of errors in the NACE-code classification for business statistics on the accuracy of register-based level estimates by industry. We address this issue for a case study on quarterly turnover of car trade in the Netherlands, which is used for the short-term statistics, the production statistics and national accounts. We first estimated the current size of classification errors by using an audit sample, data from the business register and expert knowledge. Results were used to estimate the accuracy of quarterly turnover estimates by a parametric bootstrap. Finally, we explored, by simulation, to what extent the accuracy improves with increased editing effort.

Key words: register-based statistics, audit sample, bootstrap, misclassification, NACE

1 Introduction

Good quality of published outcomes is the cornerstone of National Statistical Institutes (NSIs) that are responsible for official statistical figures. The importance of good quality outcomes becomes even more prominent nowadays, because also others – like commercial parties – produce statistics. NSIs increasingly use register-based or mixed-source statistics that encompass nearly all units in the population. Although those estimates do not suffer from sampling errors, that does not imply that they are error-free. The secondary data that are often included in the mixed-source estimates were collected for another intention than producing statistics. Those data may therefore suffer from errors along the measurement side and from errors along the side of the units, e.g. over- and undercoverage. An overview of possible errors is given in Zhang (2012a).

These errors may affect the quality of mixed-source statistics, but quantifying their effect on the published output is not yet straightforward (Agafiţei *et al.*, 2015). There are at least two issues to be solved. The first issue is how to obtain information about the occurrence of non-sampling errors in practical situations. There are many potential non-sampling errors and it may be hard to determine to what extent the observed set of units differs from the target set and the extent to which the data at hand differ from the values according to the target definitions. For quantifying those non-sampling errors, we seek to re-use already available data as much as possible, because collecting additional data is costly and time consuming.

The second issue is how to estimate the accuracy (bias and variance) of the output, given the information we have about the non-sampling errors. Depending on the complexity of the combined data sources and the type of non-sampling errors, sometimes analytical expressions can be found (Burger *et al.*, 2015; Zhang, 2012b). In the current paper, we build on Burger *et al.*'s (2015) approach and use a parametric bootstrap to estimate accuracy.

As a case study, we aim to quantify the accuracy of quarterly turnover (level) estimates of the car trade sector. Those figures are broken down into nine industries, defined by the NACE Rev. 2 classification. The NACE codes are maintained in a central business register (BR) that contains the

[1] Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague and P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. E-mail: adln@cbs.nl, sshs@cbs.nl, jbur@cbs.nl. The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

population of statistical units over time. It is often difficult to appoint the correct NACE code to a unit (Christensen, 2008). At Statistics Netherlands (SN), a service level agreement states that 95% of the large units and 65% of the small units should be correct at three-digit NACE code.

The quarterly turnover is based on two sources. For most units the turnover is derived from value added tax data. In what follows, these units are referred to as the **simple** units. The remaining units are observed through a census survey. Those remaining units are divided into **complex** and **most complex** units. The most complex units refer to very large internationally active companies that are treated by a separate business unit within SN.

2 Basic approach

We are interested in estimating the accuracy of stratum total estimates of a target variable (say turnover) by industry. First consider an enterprise i with an unknown true industry code $s_i = g$ and an observed industry code $\hat{s}_i = h$. There are various reasons why the observed industry code can differ from the true one. For instance an error in the coding practice at the chamber of commerce (COC) could lead to an erroneous NACE code. In addition, the legal unit could deliberately register itself at the COC with an incorrect NACE code. Another option is that the enterprise changed its activity, but it did not report this change to the COC.

We assume that, given the true industry code $s_i = g$, there is a probability distribution that describes the set of (potentially) observed NACE codes in the BR. We suppose that random classification errors occur, independently for each unit. In other words, given the true industry codes, we can have different realisations (observations) of a BR each leading to different estimates of the stratum turnover.

First suppose that the random classification errors occur according to a known (or previously estimated) unit-specific 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 regard 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

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

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

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 using bootstrap resampling (Efron and Tibshirani, 1993). We use bootstrap resampling, because we can extend this method in the future to handle more complicated situations such as interactions between different non-sampling errors. As described in Burger *et al.* (2015), we apply the transition matrix $\mathbf{P}_i$ to the observed $\hat{s}_i$, which results in a new industry assignment variable for all units in the population, denoted by $\hat{s}_{ir}^*$, for each bootstrap replicate ($r = 1, \ldots, R$). The bootstrap bias and variance are then estimated as follows: $\hat{B}_R^*(\hat{Y}_h) = m_R(\hat{Y}_h^*) - \hat{Y}_h$ and $\hat{V}_R^*(\hat{Y}_h) = \frac{1}{R-1} \sum_{r=1}^{R} \{\hat{Y}_{hr}^* - m_R(\hat{Y}_h^*)\}^2$, with $m_R(\hat{Y}_h^*) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_{hr}^*$. These bootstrap estimates are biased because they are based on the observed $\hat{s}_i$ rather than the true s_i. This bias is largest for $\hat{B}_R^*(\hat{Y}_h)$; we corrected that bias, as described in Van Delden *et al.* (2015).

3 Estimating the transition matrix

We estimated the transition matrices $\mathbf{P}_i$ using three sources. As a first source, an audit sample was drawn of 25 enterprises from each of the nine industries of car trade, from the simple units. For each of the 25 units, an expert determined the correct NACE code by consulting information from the

internet and by contacting the enterprise, if needed. A second source of information was the yearly changes of NACE codes in the BR. We assumed that the pairs of NACE codes for which there are frequent yearly changes are also the (pairs of) codes where misclassifications occur often. A third source of information we used was expert knowledge. Experts divided the combinations of NACE-codes into four groups. NACE codes belonging to the same group are supposed to have a comparable probability of misclassification (given that the observed NACE code is incorrect). Experts were also consulted to estimate the probability for misclassifications for the complex and most complex units.

We first estimated the diagonal elements of the transition matrix $\mathbf{P}_i$ – *i.e.*, the probabilities $P(\hat{s}_i = g | s_i = g)$ – for the simple units, using the results of the audit sample. We modelled these probabilities by logistic regression, as a function of background variables of the units. We used subset selection to find a good fitting model; see van Delden *et al.* (2015) for detailed results. The estimated diagonal probabilities (see Figure 1) are smaller for small (label 0–3) than larger size classes (4, 5+), and smaller for units that consist of 1-2 legal units (LU) than for units that consist of 3 LU or more. Each row in Figure 1 represents a "probability class", i.e., a set of units with the same set of diagonal elements. Thus, different probability classes can have different transition matrices. In fact there are only five matrices with distinct values.

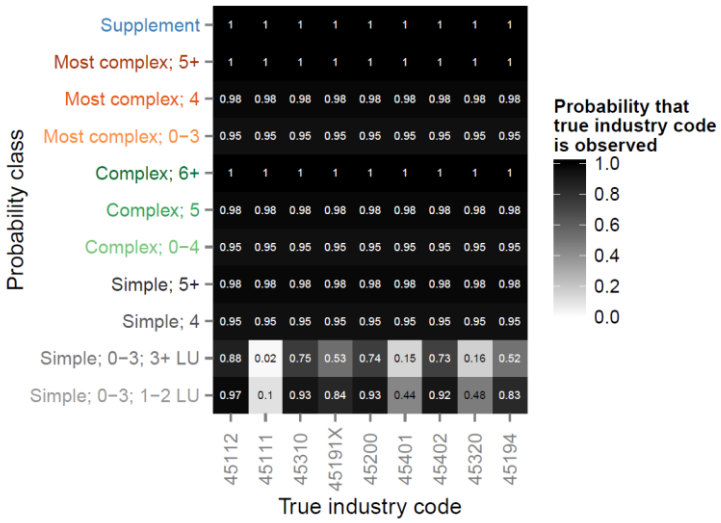

Figure 1. Estimated transition probabilities for the diagonal elements

Next, we fitted a log-linear model to the results of the audit sample, in order to estimate the off-diagonal conditional probabilities, i.e. the transition probabilities given that a unit is misclassified. The observed NACE code and the groups appointed by the experts were used as explanatory factors in the model. This model fitted well with p=0.08 for the likelihood ratio (compared to the saturated model), meaning that all relevant factors were in the model. To account for one outlier, we added a fifth group, which further improved the model fit. Note that, to reduce the number of parameters, we assumed that these conditional transition probabilities do not depend on other, unit-specific background variables.

We found pairs of industries with relatively high conditional classification error probabilities (van Delden *et al.*, 2015). For instance, a misclassified unit from industry 45310 (wholesale trade of motor vehicle parts and accessories) has a conditional probability of 0.53 to be observed as 45320 (retail trade of motor vehicle parts and accessories). Likewise, misclassified units from industry 45320 have a conditional probability of 0.33 to be observed as 45310.

Finally we estimated the transitions between car trade industries and other activities. To that end, we used the same groupings and effects as found in the log-linear model, and the enterprises in the audit sample that were observed in car trade but in fact belonged to other activities. We assumed that the number of units that belong to other activities and that are wrongly classified in car trade is equal

to the number of units that do belong to car trade but are misclassified as other activities. The exact computation can be found in van Delden *et al.* (2015).

4 Estimated accuracy

The transition matrices $\mathbf{P}_i$ were used to estimate the accuracy for car trade for ten consecutive quarters: 2012 Q1 – 2014 Q2. The relative root mean squared error (RRMSE) for the car trade sector was about 0.3% (not shown). The RRMSE of the nine underlying industries varied between 1.0% for industries 45401 and 45301 to about 60% for industry 45320. The outcomes of industry 45320 are not published separately for the short-term statistics (STS), but are combined with 45310 into STS-publication cell 45300. The quarterly turnover for publication cell 45300 has an RRMSE of about 2% (Figure 2). The least accurate publication cell is 45200 with an RRMSE of 10%. This publication cell has a large RRMSE because it has a large probability of misclassification for the small units, and the small units encompass about one-third of the total turnover of this publication cell. In some industries, the RRMSE is dominated by the variance, in others by the bias (see van Delden *et al.* 2015 for details).

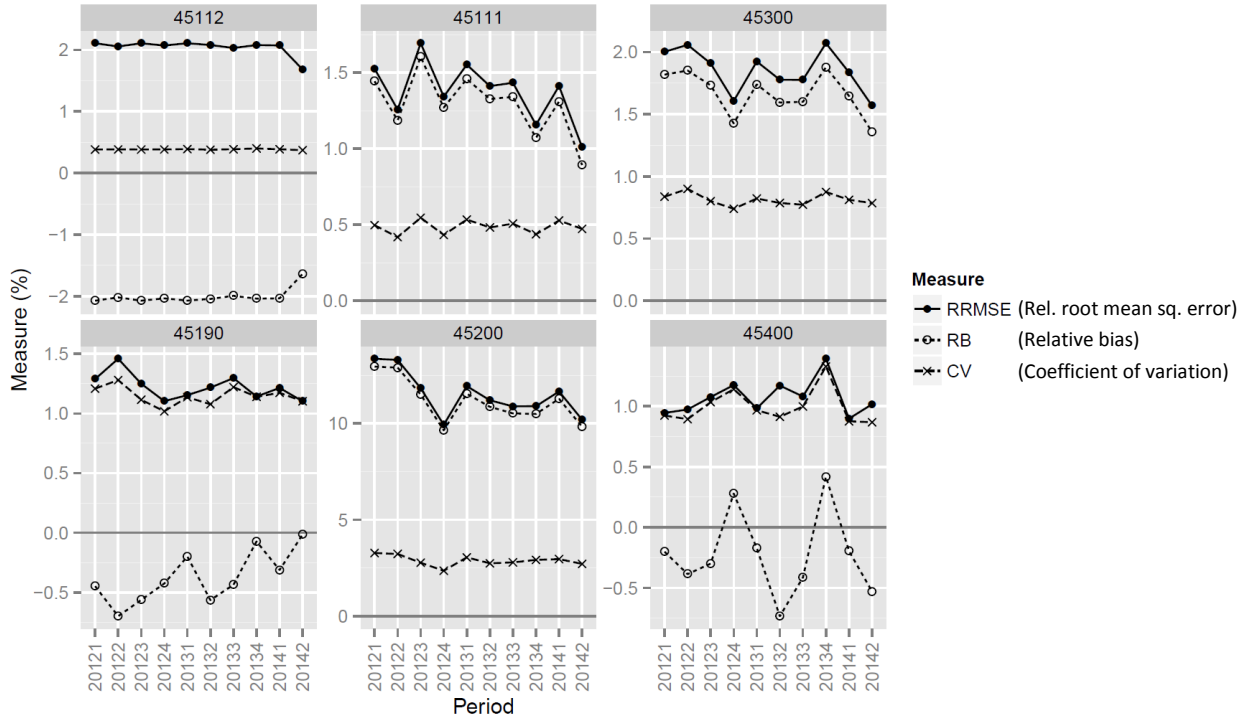

Figure 2. Accuracy measures for the six STS-publication cells of car trade

5 Improving accuracy

We explored the extent to which the accuracy of the outcomes could be improved with increased editing effort. This effort is expressed in terms of the size of the probability class "supplement" in Figure 1. The probability class "supplement" stands for the (set of) enterprises with the largest turnover, enterprises that are known and checked very well by the persons that are responsible for the statistical production. Those enterprises are therefore assumed to have no NACE code error. In the current situation this concerns the 25 largest enterprises per industry. We differentiated between four levels of editing effort, with $9\times25\times k$ enterprises included in the probability class "supplement" for car trade as a whole at level k (k = 0,1,2,3). We explored two editing scenarios: **fixed** meaning that the size of the supplement is equal for each of the nine industies, and **pro rata** meaning that the size of the supplement per industry is obtained by a Neyman-like allocation. With the pro rata scenario, we aimed

to improve the accuracy at car trade sector level in a more efficient way than with the fixed scenario. We applied the scenarios to one quarter: 2013 Q1.

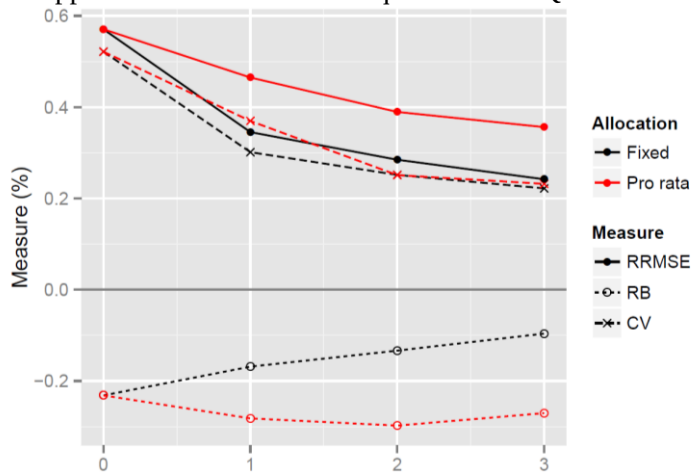

Figure 3. Quality for car trade as a whole in relation to editing effort, for Q1 2013

For the fixed allocation, the RRMSE for car trade as a whole decreased slowly with increased editing effort, as expected. Surprisingly, the pro rata allocation was less effective in reducing the RRMSE of car trade as a whole than the fixed allocation. In particular, the RB of the pro rata allocation does not improve very well with increased editing effort. This effect occurred because the accuracy of an industry is also affected by misclassification errors in other industries (see van Delden *et al.*, 2015).

6 Conclusions

In conclusion, our method can reveal which industries do not meet pre-set quality criteria. It is an open question how we can effectively improve the accuracy of estimates of specific industries. Future work will focus on quantifying the effect of classification errors on growth rate estimates, and on extending the method to other types of non-sampling error.

References

Agafiţei, M., Gras, F., Kloek, W., Reis, F. and Vaju, S. (2015). Measuring output quality for multisource statistics in official statistics: Some directions. Statistical Journal of the IAOS 31 (2015) 203–211.

Burger, J., van Delden, A. and Scholtus, S. (2015, accepted). Sensitivity of mixed-source statistics to classification errors. Accepted for publication in a special issue of Journal of Official Statistics.

Christensen, J.L. 2008. Questioning the precision of statistical classification of industries. Conference on entrepreneurship and innovation – organizations, institutions, systems and regions, Copenhagen, Denmark, June 17–20, 2008.

Delden, A. van, Scholtus, S. and Burger, J. (2015). Quantifying the effect of classification errors on the accuracy of mixed-source statistics. Discussion paper, Statistics Netherlands.

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

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.

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. European Conference on Quality in Official Statistics (Q2012), May 30–June 1, Athens.