Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 489–506, http://dx.doi.org/10.1515/JOS-2015-0029

Sensitivity of Mixed-Source Statistics to Classification Errors

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

For policymakers and other users of official statistics, it is crucial to distinguish real differences underlying statistical outcomes from noise caused by various error sources in the statistical process. This has become more difficult as official statistics are increasingly based upon a mix of sources that typically do not involve probability sampling. In this article, we apply a resampling method to assess the sensitivity of mixed-source statistics to source-specific classification errors. Classification errors can be seen as coverage errors within a stratum. The method can be used to compare relative accuracies between strata and releases, it can assist in deciding how to optimally allocate resources in the statistical process, and it can be applied in evaluating potential estimators. A case study on short-term business statistics shows that bias occurs especially for those strata that deviate strongly from the mean value in other strata. It also suggests that shifting classification resources from small and medium-sized enterprises to large enterprises has virtually no net effect on accuracy, because the gain in precision is offset by the creation of bias. The resampling method can be extended to include other types of nonsampling error.

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

1. Introduction

Official statistics provide information to policymakers, researchers and the general public on a country's social and economic development. Traditionally, the information is collected through sample surveys. Nowadays, National Statistical Institutes (NSIs) increasingly use administrative data. Administrative sources provide a population frame from which samples can be drawn, and auxiliary information that can be used to correct for selective nonresponse in sample surveys (Bethlehem 2009). Moreover, statistics can be based entirely on administrative data (UNECE 2007). The main advantages of administrative data are a reduced response burden and lower costs for the NSI. The costs per inhabitant of censuses based on administrative data or virtual censuses are one or two orders of magnitude smaller than those of traditional censuses (Chamberlain and Schulte Nordholt 2004), without any additional burden on respondents. On the other hand, administrative data are not designed for

[1] Statistics Netherlands, Department of Process Development and Methodology, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: j.burger@cbs.nl
[2] Statistics Netherlands, Department of Process Development and Methodology, Henri Faasdreef 312, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Emails: a.vandelden@cbs.nl and s.scholtus@cbs.nl
Acknowledgments: We thank Arjen de Boer for providing the raw data, Daniel Lewis and Piet Daas for their discussion and Bart Bakker, Peter van der Heijden, the associate editor and two anonymous referees for their useful comments on earlier drafts. This work was supported by the ESSnet on the Use of Administrative and Accounts Data for Business Statistics.

statistical purposes. They may suffer from selective undercoverage, and administrative units and variables may not match statistical definitions (Bakker and Daas 2012). In other words, they are prone to nonsampling errors along the lines of the representation side and the measurement side (Zhang 2012a). The representation side of nonsampling error addresses units, which can be redundant (out-of-scope), missing, misidentified, misclassified, and so on. The measurement side of nonsampling error addresses variables, which can be proxy, unstable, mismeasured, wrongly processed, and so forth.

To benefit from the best of both worlds, survey and administrative data can be combined at unit level through data integration techniques, such as record linkage, statistical matching, and microintegration processing. Using the strength of both sources, with the administrative data covering a large part of the population and the survey data matching statistical definitions, NSIs tend to publish statistical information at a more detailed level than with survey data alone.

It is unclear how accurate the estimates based on administrative data or mixed sources are. Knowledge of the accuracy of those estimates is crucial, both for users of the statistical output and for NSIs. For users of statistical output, statistical estimates need to be precise and approximately unbiased to achieve sound decision making. For NSIs, quantification of the accuracy can be used in the design phase of a new statistical production process to compare possible estimators and select the 'best' one. After the implementation of the statistical process, knowledge about the effects of various nonsampling errors on accuracy can be used to improve the production process.

The present article provides an example of the use of mixed-source estimates in business statistics. In business statistics, an important source of nonsampling errors is the classification of statistical units into economic activity or industry code. The correct industry code of a unit is hard to determine because units often perform a mixture of economic activities and their activities may change over time. For statistical purposes, the correct code can be determined using operational derivation rules and different sources, such as internet and chamber of commerce data, but finding the correct code often requires expert knowledge. NSIs often focus their editing effort on the largest and most complex units and have neither the time nor the resources to verify the industry codes for the numerous small units. Consequently, it is to be expected that some units – small units in particular – are assigned to the wrong economic activity stratum. Such classification errors can be seen as coverage errors within a stratum; a coverage error occurs when a unit is unjustly included (overcoverage) or excluded (undercoverage) from the target population. In Zhang's (2012a) classification, these errors fall along the line of representation.

A well-developed theory for estimating the accuracy of estimates as a function of probability sampling exists that has been applied in many practical situations (e.g., Särndal et al. 1992). Far less advanced is the current theory on how to estimate the accuracy of outcomes as a function of nonsampling errors, in particular for the case of mixed sources. This theory needs to be elaborated further before it can be applied easily in practical situations. Several authors have posited ideas about this topic. Bryant and Graham (2013), for instance, proposed to estimate the uncertainty caused by nonsampling errors using a Bayesian approach. Zhang (2012b) used analytical formulas to compare the accuracy of two estimators, whereas Zhang (2011) used formulas combined with bootstrap resampling to assess uncertainty due to errors in the grouping of persons into households.

In the present article we apply a bootstrap resampling method. We limit ourselves to classification errors in business statistics, but the method can be extended to other error types and is equally applicable to social statistics. We apply the method to a case study on quarterly turnover for the short-term business statistics (STS), where data for the statistical units (enterprises) underlying the largest businesses are directly observed through a census survey and the other units are observed in administrative data. Others have considered two-phase sampling (Demnati and Rao 2009) and the case of a sample survey overlapping with a selective register (Kuijvenhoven and Scholtus 2011). We limit the results to a simple-level estimator, but the methods described can also be applied to complex estimators or to temporal changes.

The rest of the article is organized as follows. In Section 2 we develop the theory to estimate the bias and variance due to classification errors. In Section 3 we present a case study, the results of which are shown in Section 4. We close with a discussion in Section 5.

2. Theory to Estimate the Bias and Variance Due to Classification Errors

Consider a population of N units that are classified into H strata (e.g., based on economic activity). Let y_i denote the turnover – or, more generally, any quantitative variable – of unit i, and s_i the (unknown) true stratum to which this unit should be assigned. Suppose we would like to know the total turnover in each stratum: $Y_h = \sum_{i=1}^{N} a_{hi} y_i$, 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 this article, we consider the relatively simple case that the true value of turnover is observed for all units. However, we do not observe the true stratum s_i but an approximation thereof, which may be affected by random classification errors. Denote the stratum to which unit i is actually assigned by $\hat{s}_i$, and let $\hat{a}_{hi} = I\{\hat{s}_i = h\}$. Then the estimated total turnover in stratum h is: $\hat{Y}_h = \sum_{i=1}^{N} \hat{a}_{hi} y_i$.

For simplicity, we suppose that random classification errors occur according to a known (or previously estimated) transition matrix $\mathbf{P} = (p_{gh})$, with $p_{gh} = \Pr(\hat{s}_i = h | s_i = g)$, where it is assumed that each unit in a given true stratum has the same probability of being misclassified in one of the other strata. (That is to say, each unit has the same transition matrix $\mathbf{P}$.) Moreover, we assume that classification errors are independent across units. Finally, we make the technical assumption that $p_{hh} > \max_{g \neq h} p_{gh}$ for all h.

In the application below, we will use a transition matrix of the following particular form:

$$\mathbf{P} = \begin{bmatrix} p & \frac{1-p}{H-1} & \cdots & \frac{1-p}{H-1} \\ \frac{1-p}{H-1} & p & \cdots & \frac{1-p}{H-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-p}{H-1} & \frac{1-p}{H-1} & \cdots & p \end{bmatrix} \tag{1}$$

In this special case, each unit is classified correctly with probability p and misclassified with probability $1 - p$. Moreover, the misclassified units are distributed uniformly over the other strata. This simple transition matrix is used to help in the exposition of the methodology, but possible extensions are indicated in the discussion. Note that for matrices that have the form (1), the above condition $p_{hh} > \max_{g \neq h} p_{gh}$ is equivalent to $p > 1/H$.

We would like to assess the bias and variance of $\hat{Y}_h$ as an estimator for Y_h, that is,

$$B(\hat{Y}_h) = E(\hat{Y}_h - Y_h) = \sum_{i=1}^{N} \{E(\hat{a}_{hi}) - a_{hi}\} y_i, \tag{2}$$

$$V(\hat{Y}_h) = \sum_{i=1}^{N} V(\hat{a}_{hi}) y_i^2, \tag{3}$$

where in (3) we used the assumption of independent classification errors across units.

In the relatively simple situation considered here, it is not too difficult to derive analytical expressions for (2) and (3); see the Appendix for more details. Note that the resulting expressions contain unknown quantities such as Y_h that need to be estimated. Moreover, in future applications we may want to consider situations that are more complex, where this analytical approach is not possible. Therefore, this article focuses on an alternative approach to estimate (2) and (3), based on bootstrap resampling, which can be generalized to more complex situations.

For each unit i, there is an infinite population of possible classification errors, modeled by the transition probabilities $\Pr(\hat{s}_i = h | s_i = g)$ in the matrix $\mathbf{P}$. The $\hat{s}_i$ actually observed is the result of one realization of this model. Under the resampling approach, we consider a new stratum assignment variable $\hat{s}_i^*$ that is obtained by applying the transition matrix $\mathbf{P}$ to the observed $\hat{s}_i$. That is to say, we consider realisations of the alternative classification error model given by

$$\Pr(\hat{s}_i^* = h | \hat{s}_i = g) \equiv \Pr(\hat{s}_i = h | s_i = g) = p_{gh}. \tag{4}$$

We also define: $\hat{a}_{hi}^* = I\{\hat{s}_i^* = h\}$. Finally, we define the so-called bootstrap replication of the estimated total turnover in stratum h: $\hat{Y}_h^* = \sum_{i=1}^{N} \hat{a}_{hi}^* y_i$.

In terms of these bootstrap replications, the bias and variance of $\hat{Y}_h$ as an estimator for Y_h may be estimated consistently by, respectively, the bias and variance of $\hat{Y}_h^*$ as an estimator for $\hat{Y}_h$ (e.g., Efron and Tibshirani 1993). In the particular situation considered here, it is possible to obtain the latter bias and variance analytically (see the Appendix). In general, they have to be estimated through Monte Carlo simulation. For this, we generate a large number (say, R) of random draws from the classification error model (4). Denote these draws by $\hat{s}_{i1}^*, \ldots, \hat{s}_{iR}^*$. From these $\hat{s}_{ir}^*$, we can compute $\hat{a}_{hir}^* = I\{\hat{s}_{ir}^* = h\}$ and subsequently $\hat{Y}_{hr}^* = \sum_{i=1}^{N} \hat{a}_{hir}^* y_i$. The bootstrap bias and variance are then estimated as follows (Efron and Tibshirani 1993):

$$\hat{B}_R^*(\hat{Y}_h) = m_R(\hat{Y}_h^*) - \hat{Y}_h, \tag{5}$$

$$\hat{V}_R^*(\hat{Y}_h) = \frac{1}{R-1} \sum_{r=1}^{R} \{\hat{Y}_{hr}^* - m_R(\hat{Y}_h^*)\}^2, \tag{6}$$

with

$$m_R(\hat{Y}_h^*) = \frac{1}{R}\sum_{r=1}^{R} \hat{Y}_{hr}^*,$$

the average value of the bootstrap replications. For sufficiently large values of R, $\hat{B}_R^*(\hat{Y}_h)$ and $\hat{V}_R^*(\hat{Y}_h)$ converge to the true bias and variance of $\hat{Y}_h^*$ as an estimator for $\hat{Y}_h$ and hence to consistent estimators of the bias and variance of $\hat{Y}_h$.

This is an example of a parametric bootstrap method. Using the observed stratum assignments as a starting point, we resample the classification errors from an explicit model, given by the transition matrix $\mathbf{P}$. Technically, resampling model (4) can be justified as a parametric bootstrap method provided that $\hat{s}_i$ is a Maximum Likelihood Estimator (MLE) for s_i. Under the condition $p_{hh} > \max_{g \neq h} p_{gh}$ introduced above, this is indeed the case (see the Appendix).

As discussed in the Appendix, the above bootstrap estimators $\hat{B}_R^*(\hat{Y}_h)$ and $\hat{V}_R^*(\hat{Y}_h)$ are consistent but not unbiased with respect to the true bias and variance of $\hat{Y}_h$. For the special case that $\mathbf{P}$ has the form (1), it is shown in the Appendix that improved, bias-corrected bootstrap estimators may be computed as follows:

$$\hat{B}_{R,BC}^*(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} \hat{B}_R^*(\hat{Y}_h), \tag{7}$$

$$\hat{V}_{R,BC}^*(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} \left[\hat{V}_R^*(\hat{Y}_h) - \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\right], \tag{8}$$

with $K = \sum_{i=1}^{N} y_i^2$. Note that, under the assumptions made here, all quantities on the right-hand sides of Expressions (7) and (8) are known. For more complex situations, analytical bias corrections for the bootstrap estimators are not readily available; we will return to this point in the discussion.

In the application below, the matrix $\mathbf{P}$ will be assumed to be known. In general, it would have to be estimated. This would require an 'audit sample' of units for which both s_i and $\hat{s}_i$ are observed. Having obtained an estimate $\hat{\mathbf{P}}$ of $\mathbf{P}$, we can apply the above bootstrap method by resampling from the classification error model (4) with $\mathbf{P}$ replaced by $\hat{\mathbf{P}}$.

3. Case Study

3.1. Data

At Statistics Netherlands, quarterly turnover for STS is based on a mix of primary and administrative data. The turnover estimates are published in four subsequent releases: 30 days, 60 days, 90 days, and one year after the end of the reference period. The turnover of most businesses is obtained from Value Added Tax (VAT) data, whereas the statistical units (enterprises) underlying the largest and most complex businesses are directly observed through a census survey. The rationale behind this design is that for larger and more complex businesses, it is not possible to make a one-to-one link between

administrative units and statistical units. Furthermore, early estimates typically need to be produced before the survey and administrative data are completely available. The missing data are imputed using ratio imputation, based on data from early respondents and historical information of the nonresponding units. Because no samples are drawn and missing data are imputed, no complicated design-based or model-based estimators are required to make inferences about the target population. The estimator for the total quarterly turnover in a given industry is simply the sum of observed and imputed values over all units in both strata. More information about the case study can be found in van Delden and de Wolf (2013) and the references therein.

The turnover estimates of subsequent quarters are not only used to publish turnover growth rates – stratified by economic activity – for the STS regulation, but are also used to compute yearly turnover levels. Those turnover levels are used to calibrate results of the Structural Business Statistics (SBS), which in turn are used as one of the sources to determine the gross domestic product. Thus, for both the turnover levels and the growth rates we would like to have precise and approximately unbiased results.

We will focus on nine industries of economic activity (Figure 1), defined by the Dutch particularization of NACE Rev. 2 within Division 45: "Wholesale and retail trade and repair of motor vehicles and motorcycles". In most of those industries, turnover estimates are based on a combination of survey and administrative data. In some industries, such as 45111 ("Import of new cars and light motor vehicles"), estimates are based mainly on survey data. In others, such as 45194 ("Wholesale and retail trade and repair of caravans") and 45402 ("Retail trade and repair of motorcycles and related parts and accessories"), estimates are completely based on administrative data. The proportion of values that are imputed instead of observed can be substantial for early estimates (30 days after the end of the reference period) but is almost negligible for final estimates (one year after the end of the reference period).

3.2. *Parameter Values and Scenarios*

In this article, we assess the sensitivity of these estimates to classification errors. According to an internal Service Level Agreement (SLA), the three-digit NACE code should be correct for at least 95% of large enterprises (survey data) and 65% of small and medium-sized enterprises (admin data). These values resemble those of an audit held in 2000 and 2003 on the quality of the three-digit NACE code in the Dutch Business Register, which reported that 97% of the NACE codes are correct for large units (20 employees or more) in Retail Trade and 69% of the NACE codes are correct for small units (up to 19 employees) averaged over industries. The proportion of correct NACE codes is higher for large units than for small units because more resources are invested in classifying a large unit's economic activity through profiling.

We applied the SLA figures at industry level to the survey/admin division of units, which roughly correlates with unit size. We assumed that the first two digits of the NACE code in our nine industries are correct and that the probability of moving from one industry to another is the same for all industries. We used this assumption for ease of computation, which aims to illustrate the procedure of the sensitivity analysis. Whether this assumption is valid needs to be verified by carrying out a detailed audit on classification errors within

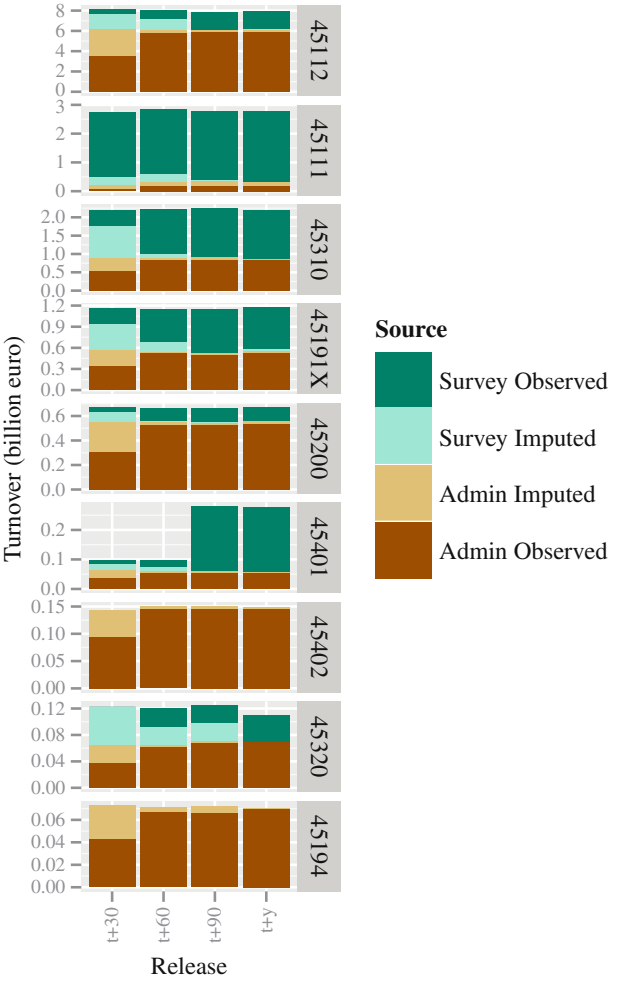

Fig. 1. Mixed-source estimates of quarterly turnover at 30 days, 60 days, 90 days and one year after the end of the reference period (third quarter of 2011) for nine industries within the Dutch particularization of NACE Rev. 2 within Division 45. Industries are ordered from large to small. Note that the y-axes are scaled independently between industries.

Division 45. The results of such an audit may lead to extensions, which are mentioned in the discussion.

We can then define two source-specific 9 × 9 transition matrices (Scenario 1):

$$\mathbf{P}^{\text{survey}} = \begin{bmatrix} \frac{19}{20} & \frac{1}{160} & \cdots & \frac{1}{160} \\ \frac{1}{160} & \frac{19}{20} & \cdots & \frac{1}{160} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{160} & \frac{1}{160} & \cdots & \frac{19}{20} \end{bmatrix}$$

and

$$\mathbf{P}^{\text{admin}} = \begin{bmatrix} \frac{13}{20} & \frac{7}{160} & \cdots & \frac{7}{160} \\ \frac{7}{160} & \frac{13}{20} & \cdots & \frac{7}{160} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{7}{160} & \frac{7}{160} & \cdots & \frac{13}{20} \end{bmatrix}$$

Note that both matrices are special cases of the matrix $\mathbf{P}$ in (1).

Although it makes intuitive sense to allocate more resources to large units that have a large impact on the statistical outcome, one could also argue that many small units may still have a considerable impact and should not be ignored altogether. In order to study the relative importance of resource allocation, we introduce a second scenario. By switching the matrices between sources, we studied what would happen if instead 65% of large enterprises (survey data) and 95% of small and medium-sized enterprises (admin data) were correctly classified for economic activity (Scenario 2). In summary, we are comparing a scenario where classification resources are mainly allocated to large units receiving a questionnaire with a scenario where classification resources are mainly allocated to small units whose information is derived from administrative sources.

3.3. Resampling

Using this input, we first drew a new industry code for each unit from these transition matrices. For instance, a unit that receives a survey and is classified in industry 45111 has a probability of $19/20$ of remaining in 45111 and a probability of $1/160$ of ending up in one of the other eight industries. A unit for which the data come from the admin source and that is classified in industry 45111 has a probability of $13/20$ of remaining in 45111 and a probability of $7/160$ of ending up in one of the other eight industries. We then recalculated the population parameter per (new) industry. Next, we repeated this a large number of times: $R = 10,000$ simulations per estimate, which seemed sufficient for confidence intervals to converge (Burger et al. 2013). From these replications, the bias and variance due to classification errors were estimated using the bias-corrected expressions (7) and (8). In summary, we assumed a stochastic error process and we used resampling to quantify the effects of this error process on the turnover estimates.

4. Results

Each turnover estimate is compared with the distribution of bootstrap replications in Figure 2a. The estimated variance and the square of the bias were added together, resulting in the mean square error (MSE) as a measure of accuracy. The square root (RMSE) was taken to revert to the unit of the data (euro), and was normalized (relative root mean squared error; RRMSE) to the total turnover estimated from observed and imputed data to make estimates comparable between releases and industries (Figure 2b).

The RRMSE can be alarmingly high: over 900% (Figure 2). We would like to stress, however, that we have estimated not the true accuracy of the turnover estimates, but their

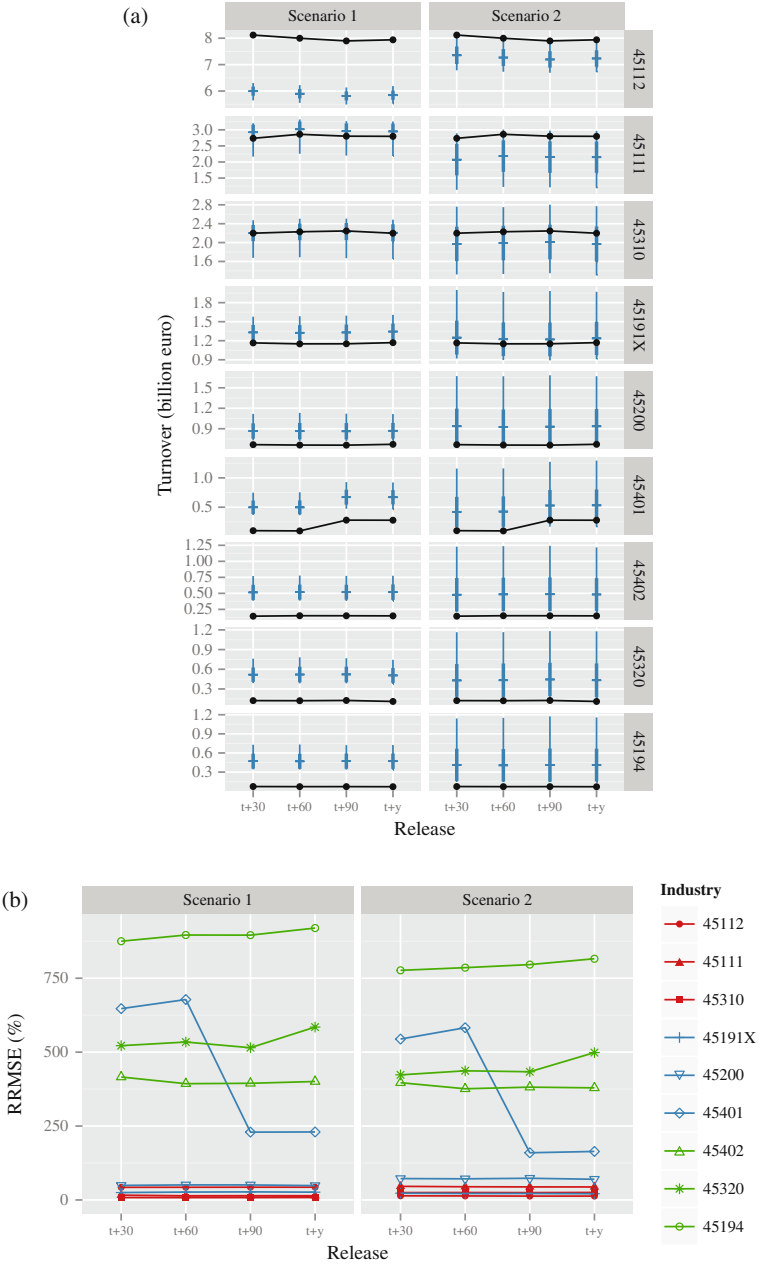

Fig. 2. Sensitivity of mixed-source estimates to source-specific classification error. (a) Quarterly turnover per industry and release estimated from observed and imputed data (black dots and lines), and simulated mean (blue horizontal dashes) ± SD (blue thick bars), and 2.5th and 97.5th percentiles (blue thin bars) using 10,000 simulations per estimate. Note that the y-axes are scaled independently between industries. (b) Root mean square error normalized to the quarterly turnover estimated from observed and imputed data. Classification error is assumed largest in admin stratum (Scenario 1) or survey stratum (Scenario 2). Industries are ordered from large to small.

relative sensitivity to classification errors. In particular, having uniform transition probabilities between strata may not be a realistic assumption. Moreover, the RRMSE correlates negatively with the turnover estimates, that is, only small industries (a few hundred million euros or less) have such a high RRMSE.

Simulations under Scenario 1 show that source-specific misclassification can result in strongly biased estimates (Figure 2). Our dataset contains one high-turnover industry (45112). In Figure 2, the simulated total turnover for this industry lies consistently below the original estimate. According to Expressions (5) and (7), this means that the total turnover of this industry is underestimated relative to the unknown true value. This bias may be explained as follows. First, the turnover in 45112 is substantially based on units using the admin data (Figure 1), which have a fair chance of being misclassified. Second, misclassified units from other industries that are classified erroneously in 45112 typically have low turnover. Similarly, the total turnover of low-turnover industries such as 45194 is overestimated relative to the unknown true value, because many small units are erroneously replaced by units from higher-turnover industries. This confirms the analytical solution showing that the absolute bias increases the more the turnover of an industry deviates from the average turnover of the other industries (see the Appendix). In industry 45401, late estimates are more accurate than early estimates because they are based on more units with a likely correct industry code (survey data, see also Figure 1). In the other industries we do not observe an effect of release on accuracy because the ratio between survey and administrative data remains fairly constant and the imputed values were held fixed (see the discussion).

When we assume that the economic activity is more reliable for small and medium-sized enterprises than for large enterprises (Scenario 2), our estimates are indeed less precise, but also less biased (Figure 2). This suggests that shifting the focus of editing the industry classification from small and medium-sized enterprises to large enterprises can result in more biased estimates. Such a shift in resources has virtually no net effect on accuracy of the level estimates (see Figure 2b), because the gain in precision is offset by the creation of bias.

For the simple scenarios used here, it is possible to derive analytical expressions for the bias-corrected bootstrap estimators of bias and variance; see Expressions (17) and (18) in the Appendix. Note that we can apply these expressions separately to survey and admin data, as there is no interaction between the two data sources in this study. For Scenario 1, working out Expressions (17) and (18) with $H = 9$ and $p = \frac{19}{20}$ (survey data) or $p = \frac{13}{20}$ (admin data), we find:

$$\hat{B}^{*}_{\infty,BC}(\hat{Y}_h) = \frac{8}{151}\left\{\overline{\hat{Y}}^{(-h),\text{survey}} - \hat{Y}_h^{\text{survey}}\right\} + \frac{56}{97}\left\{\overline{\hat{Y}}^{(-h),\text{admin}} - \hat{Y}_h^{\text{admin}}\right\},$$

and

$$\begin{aligned}\hat{V}^{*}_{\infty,BC}(\hat{Y}_h) = {} & \frac{38}{755}\hat{K}_h^{\text{survey}} + \frac{159}{24160}\sum_{g\neq h}\hat{K}_g^{\text{survey}} - \frac{311}{483200}K^{\text{survey}} \\ & + \frac{182}{485}\hat{K}_h^{\text{admin}} + \frac{1071}{15520}\sum_{g\neq h}\hat{K}_g^{\text{admin}} - \frac{12593}{310400}K^{\text{admin}}.\end{aligned}$$

In these expressions, $\hat{Y}_h^X = \sum \hat{a}_{hi} y_i$, $\hat{K}_h^X = \sum \hat{a}_{hi} y_i^2$, and $K^X = \sum y_i^2$ where the sums are over all units in source X, with $X \in \{\text{survey}, \text{admin}\}$; in addition, $\overline{\hat{Y}}^{(-h),X} = \frac{1}{H-1} \sum_{g \neq h} \hat{Y}_g^X$. Analogous expressions are obtained for Scenario 2 by interchanging the coefficients for survey data and admin data.

The numerical solution for the bias and standard deviation closely resembles the analytical solution derived in the Appendix (Figure 3). The mean difference in bias between the numerical and analytical solution is zero euro with the maximum absolute difference being merely twelve million euros (eleven percent of the analytical solution). The mean relative difference in standard deviation is 0.6% with the maximum relative absolute difference being 7.4% of the analytical solution. This confirms that 10,000 simulations are sufficient to approximate the analytical solution.

5. Discussion

For policymakers and other users of official statistics, it is crucial to distinguish real differences between statistical outcomes from noise caused by various error sources in the statistical process. This has become more difficult as official statistics are now increasingly based upon a mix of sources that typically do not involve probability sampling. We have described a case study where statistical units (enterprises) underlying large and complex businesses are directly observed through a census survey and the turnover of smaller and less complex enterprises is obtained from tax data.

The resampling method described in the current article provides insight into the sensitivity of mixed-source statistics to a source-specific nonsampling error. Results can be used to compare industries and releases, and can assist in deciding where to invest resources into the statistical process. Our results show that bias occurs especially in those strata that deviate strongly from the mean value in other strata. The example we have shown also suggests that shifting classification resources from small and medium-sized enterprises to large enterprises has virtually no net effect on the accuracy of the level estimates, because the gain in precision is offset by the creation of bias. On the other hand, this resource allocation might improve the accuracy of temporal turnover changes, because the creation of bias in both time points is annihilated, whereas the gain in precision is not. Results indicate that level estimates will become less biased when NSIs

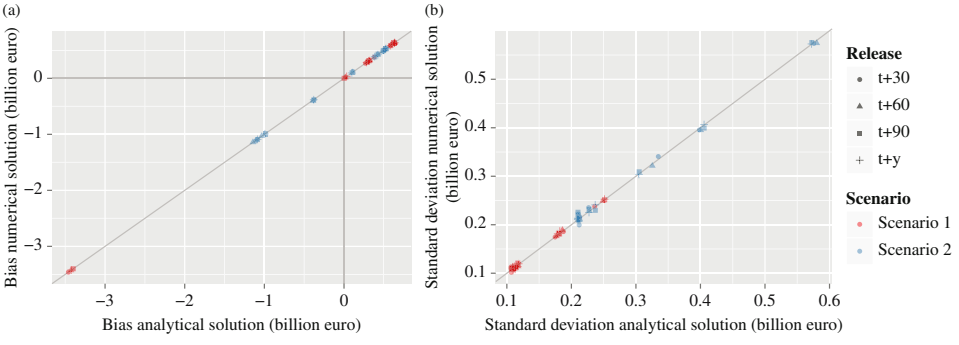

Fig. 3. Comparison between analytical and numerical solution for (a) bias and (b) standard deviation. Diagonal shows $y = x$.

find ways to improve the correctness of the industry codes of small enterprises, while maintaining the industry code quality of large enterprises. Because manual coding will be too expensive in practice, other approaches are needed. One possible future direction is to automatically collect data on products and services from business websites combined with text-mining techniques to translate the results into reliable industry codes.

The resampling method that we have presented can be used not only for sensitivity analyses but also to estimate the accuracy of outcomes. A major prerequisite to achieving this is to find cost-effective ways that can be used by NSIs to obtain a sound estimation of the error distribution. In our case study, we used reasonable parameter values for the probability that the observed industry code is correct. Nonetheless, we simply assumed that the probability of moving from one industry to another is the same, whereas in reality we expect those probabilities to vary, both between pairs of strata and between units. With our current parameter settings, we found extremely high RRMSEs in some industries. These results underline that our parameterization needs to be refined before drawing final conclusions about the data. We encourage other NSIs to run similar simulations with their own parameter settings of the transition matrix.

We see two steps to improve estimating a transition matrix. First, we need to understand which variables determine the correctness of an industry code for a specific unit, for instance its (observed) size class, its three-digit NACE code and the occurrence of an event (birth, merger, take-over etc.). Second, we need to estimate the error distribution. Possibilities for estimating the desired input would be to compare different sources, to derive estimates from the editing process, to apply audit sampling, and/or to model the true economic activity as a latent class. Note that accuracy estimates can also be extended to account for uncertainty in knowledge about those parameter values. Zhang (2011) used bootstrap resampling to account for that issue. In a Bayesian approach, uncertainty about the parameter values would be modeled by a prior distribution.

NSIs typically develop new estimators as new data sources become available or the statistical process is redesigned. The resampling method can also be applied to compare different estimators and to test which estimator is the least sensitive to the error process. It could also be used to decide about the line of demarcation between the survey and the admin data.

Note that we have assumed that the imputed turnover values are independent of the industry code. In reality, the industry code is used as auxiliary information in the imputation process. It would therefore be more realistic to impute missing values after resampling instead of assuming fixed imputed values (Shao and Sitter 1996). This would affect early releases where a substantial proportion of the estimate is based on imputed values. We expect that, when variation due to imputation is accounted for, classification errors will affect early releases more than late ones.

A theoretical difficulty that remains to be solved is that the direct bootstrap estimators of bias and variance may be biased in practice. In the above simplified application, we could correct this bias analytically. However, we also want to be able to use the bootstrap method in more realistic situations (as discussed above) where analytical derivations are no longer feasible, and we have no reason to assume that the bootstrap estimators will be less biased in these applications. It may be possible to obtain bias corrections to the bootstrap estimators numerically, for example, by applying a nested version of the

bootstrap in which the bootstrap resamples are resampled themselves; see Efron and Tibshirani (1993). Another, computationally more attractive possibility could be to work with so-called bias-corrected bootstrap confidence intervals (Efron and Tibshirani 1993; DiCiccio and Efron 1996) instead of bias and variance estimates. This remains to be investigated.

The resampling method could be adapted to specific situations or needs. First of all, we could extend the method to account for overcoverage and undercoverage of units in the population frame. To that end, we could introduce an exclusion stratum, 'outside the population', and for each industry code estimate the overcoverage (true value is 'outside the population') and the undercoverage: the proportion that is unjustly missing. Furthermore, we could extend the method to study measurement errors, a combination of (interacting) nonsampling errors or errors due to nonprobability sampling (see for instance de Munnik et al. 2013). Another extension could be to assess the effect on accuracy of changes over time rather than of levels.

Appendix

The Observed Industry Code As An MLE for the True Industry Code

Recall from Section 2 that the resampling model (4) can be justified as a parametric bootstrap method provided that $\hat{s}_i$ is a Maximum Likelihood Estimator (MLE) for s_i. Below we will prove that this is the case.

Let $\boldsymbol{s} = (s_1, \ldots, s_N)'$ and $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_N)'$ denote vectors of true and observed industry codes, respectively. Since classification errors are assumed to be independent across units, the joint parametric model for the observed industry codes is given by:

$$\Pr\left(\hat{\boldsymbol{s}} = (h_1, \ldots, h_N)' | \boldsymbol{s} = (g_1, \ldots, g_N)'\right) = \prod_{i=1}^{N} \Pr\left(\hat{s}_i = h_i | s_i = g_i\right) = \prod_{i=1}^{N} p_{g_i h_i}.$$

Consider the log-likelihood function of the unknown parameter vector s, given the observed industry codes $\hat{s}$. By definition, it holds that:

$$\log L\left(\boldsymbol{s} = (g_1, \ldots, g_N)' | \hat{\boldsymbol{s}} = (h_1, \ldots, h_N)'\right) = \sum_{i=1}^{N} \log p_{g_i h_i}.$$

Since we assumed independence across units, we can maximize this sum by maximizing each term separately. Under the condition that $p_{hh} > \max_{g \neq h} p_{gh}$ for all h, it follows that the i^{th} term is maximized by choosing $s_i = g_i = h_i = \hat{s}_i$. We conclude that the MLE of $\boldsymbol{s}$ is given by $\hat{\boldsymbol{s}}$. As noted in Section 2, this justifies the use of resampling model (4) as an application of the parametric bootstrap. In addition, it follows that $\hat{Y}_h$ is a so-called 'plug-in estimator' of Y_h, which justifies Expression (5) (Efron and Tibshirani 1993).

While $\hat{s}_i$ is the MLE of s_i here, it will be shown below that the direct bootstrap estimators (5) and (6) are biased with respect to (2) and (3). This may be explained by the fact that we are using a sample of size N to estimate the N unknown parameters $s_1, \ldots, s_N$ of the parametric model. It is well known that MLEs – and, by extension, bootstrap estimators – are usually biased in situations where the effective sample size is small.

On the other hand, the bootstrap estimators *are* asymptotically consistent, because we are *not* in a situation where the number of unknown parameters increases with the sample size. Given our fixed population of N units, we could – in theory – obtain m independently assigned industry codes $\hat{s}_{i1}, \ldots, \hat{s}_{im}$ for each unit, thereby drawing a sample of size mN from the parametric model. The bias in the corresponding bootstrap estimators – with Model (4) applied to the MLE of s_i based on $\hat{s}_{i1}, \ldots, \hat{s}_{im}$ – would then vanish as $m \to \infty$.

Derivation of Bias and Variance

For the highly simplified situation considered in Section 2, we can derive analytical expressions for the bias and variance of $\hat{Y}_h$.

Let $\boldsymbol{a}_i = \big(a_{1i}, \ldots, a_{Hi}\big)'$ and $\hat{\boldsymbol{a}}_i = (\hat{a}_{1i}, \ldots, \hat{a}_{Hi})'$. Given that classification errors are described by a transition matrix $\mathbf{P} = (p_{gh})$, we observe that:

$$E(\hat{a}_{hi}) = \sum_{g=1}^{H} a_{gi} E\big(\hat{a}_{hi} | s_i = g\big) = \sum_{g=1}^{H} a_{gi} \Pr\big(\hat{s}_i = h | s_i = g\big) = \sum_{g=1}^{H} a_{gi} p_{gh},$$

and hence that $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}'\boldsymbol{a}_i$. Here we used that $a_{gi} = 1$ for exactly one $g \in \{1, \ldots, H\}$. Now let $\boldsymbol{y} = \big(Y_1, \ldots, Y_H\big)'$ and $\hat{\boldsymbol{y}} = (\hat{Y}_1, \ldots, \hat{Y}_H)'$ denote vectors of (estimated) stratum totals. By definition, $\boldsymbol{y} = \sum_{i=1}^{N} \boldsymbol{a}_i y_i$ and $\hat{\boldsymbol{y}} = \sum_{i=1}^{N} \hat{\boldsymbol{a}}_i y_i$. Noting that $E\big(\hat{\boldsymbol{y}}\big) = \sum_{i=1}^{N} E(\hat{\boldsymbol{a}}_i) y_i = \mathbf{P}'\boldsymbol{y}$, we obtain for the bias of $\hat{\boldsymbol{y}}$:

$$B\big(\hat{\boldsymbol{y}}\big) = E\big(\hat{\boldsymbol{y}}\big) - \boldsymbol{y} = \big(\mathbf{P}' - \mathbf{I}\big)\boldsymbol{y}, \tag{9}$$

with $\mathbf{I}$ denoting the $H \times H$ identity matrix. In particular, this yields the following expression for the bias of a single stratum total (2):

$$B\big(\hat{Y}_h\big) = \big(p_{hh} - 1\big) Y_h + \sum_{g \neq h} p_{gh} Y_g.$$

In the special case that $\mathbf{P}$ has the Form (1), this expression can be simplified to:

$$B\big(\hat{Y}_h\big) = \big(p - 1\big) Y_h + \frac{1-p}{H-1} \sum_{g \neq h} Y_g = \big(1 - p\big)\big\{\bar{Y}^{(-h)} - Y_h\big\}, \tag{10}$$

where $\bar{Y}^{(-h)} = \frac{1}{H-1} \sum_{g \neq h} Y_g$ is the average stratum total over all strata *except* stratum h. This formula shows that the (absolute) bias decreases with p, as expected. It also shows that the (absolute) bias increases the further Y_h deviates from $\bar{Y}^{(-h)}$. In other words, bias occurs especially for those strata that deviate strongly from the mean value in other strata.

Next, we consider the variance of $\hat{\boldsymbol{y}}$. Since $\hat{\boldsymbol{a}}_i$ contains binary values, it holds that $\hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i{}' = \mathrm{diag}(\hat{\boldsymbol{a}}_i)$, where $\mathrm{diag}(\boldsymbol{x})$ denotes the diagonal matrix with $\boldsymbol{x}$ on the main diagonal. Similarly, $\boldsymbol{a}_i \boldsymbol{a}_i{}' = \mathrm{diag}(\boldsymbol{a}_i)$. Therefore, the variance-covariance matrix of $\hat{\boldsymbol{a}}_i$ may be written as follows:

$$V(\hat{\boldsymbol{a}}_i) = E(\hat{\boldsymbol{a}}_i \hat{\boldsymbol{a}}_i{}') - E(\hat{\boldsymbol{a}}_i) E(\hat{\boldsymbol{a}}_i{}') = \mathrm{diag}(E(\hat{\boldsymbol{a}}_i)) - \mathbf{P}'\boldsymbol{a}_i \boldsymbol{a}_i{}'\mathbf{P} = \mathrm{diag}(\mathbf{P}'\boldsymbol{a}_i) - \mathbf{P}'\mathrm{diag}(\boldsymbol{a}_i)\mathbf{P},$$

where we used $E(\hat{\boldsymbol{a}}_i) = \mathbf{P}'\boldsymbol{a}_i$ as derived above. Now using the fact that the variance-covariance matrix $V(\hat{\boldsymbol{y}})$ can be written as $V(\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} V(\hat{\boldsymbol{a}}_i) y_i^2$ [cf. Expression (3)],

we obtain:

$$V(\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} \left\{ \text{diag}(\mathbf{P}'\boldsymbol{a}_i y_i^2) - \mathbf{P}'\text{diag}(\boldsymbol{a}_i y_i^2)\mathbf{P} \right\} = \text{diag}(\mathbf{P}'\boldsymbol{k}) - \mathbf{P}'\text{diag}(\boldsymbol{k})\mathbf{P}. \tag{11}$$

Here, $\boldsymbol{k} = \left(K_1, \ldots, K_H\right)'$, with K_h denoting the sum of squared values for variable y_i in stratum h; that is, $K_h = \sum_{i=1}^{N} a_{hi} y_i^2$ and $\boldsymbol{k} = \sum_{i=1}^{N} \boldsymbol{a}_i y_i^2$. In particular, the main diagonal of $V\left(\hat{\boldsymbol{y}}\right)$ contains the following elements:

$$V\left(\hat{Y}_h\right) = \sum_{g=1}^{H} p_{gh} K_g - \sum_{g=1}^{H} p_{gh}^2 K_g = \sum_{g=1}^{H} p_{gh}(1 - p_{gh}) K_g.$$

In the special case that $\mathbf{P}$ has the Form (1), this formula simplifies to:

$$V(\hat{Y}_h) = p\left(1 - p\right)K_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right)\sum_{g \neq h} K_g. \tag{12}$$

Application to the Bootstrap Estimators and Derivation of (7) and (8)

Since the bootstrap replications $\hat{Y}_h^*$ are obtained by resampling from the classification error model (4), analogous analytical expressions to (9) and (11) may be derived for the bias and variance-covariance matrix of the bootstrap replications: $B(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = (\mathbf{P}' - \mathbf{I})\hat{\boldsymbol{y}}$ and $V(\hat{\boldsymbol{y}}^* | \hat{\boldsymbol{y}}) = \text{diag}(\mathbf{P}'\hat{\boldsymbol{k}}) - \mathbf{P}'\text{diag}(\hat{\boldsymbol{k}})\mathbf{P}$. Thus, for the case study in Section 3, it was possible to obtain bootstrap estimates of the bias and variance of the original estimators *without* resorting to Monte Carlo simulations. We denote these analytical estimates by $\hat{B}_{\infty}^*\left(\hat{Y}_h\right)$ and $\hat{V}_{\infty}^*\left(\hat{Y}_h\right)$, to indicate that the same estimates would also be obtained by taking the limit $R \rightarrow \infty$ in (5) and (6). In particular, for the special case that $\mathbf{P}$ has the Form (1), we obtain [cf. (10) and (12)]:

$$\hat{B}_{\infty}^*\left(\hat{Y}_h\right) = \left(1 - p\right)\left\{\overline{\hat{Y}}^{(-h)} - \hat{Y}_h\right\}, \tag{13}$$

$$\hat{V}_{\infty}^*\left(\hat{Y}_h\right) = p\left(1 - p\right)\hat{K}_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right)\sum_{g \neq h} \hat{K}_g, \tag{14}$$

in obvious notation.

It is not difficult to show that the above bootstrap estimators are biased with respect to the true bias and variance of $\hat{Y}_h$. In fact, we have:

$$E\{\hat{B}_{\infty}^*\left(\hat{\boldsymbol{y}}\right)\} = (\mathbf{P}' - \mathbf{I})E(\hat{\boldsymbol{y}}) = (\mathbf{P}' - \mathbf{I})\mathbf{P}'\boldsymbol{y} = \mathbf{P}'(\mathbf{P}' - \mathbf{I})\boldsymbol{y} = \mathbf{P}'B(\hat{\boldsymbol{y}})$$

according to Expression (9). Similarly,

$$E\left\{\hat{V}^*_\infty(\hat{\boldsymbol{y}})\right\} = \text{diag}(\mathbf{P}'E(\hat{\boldsymbol{k}})) - \mathbf{P}'\text{diag}(E(\hat{\boldsymbol{k}}))\mathbf{P}$$
$$= \text{diag}(\mathbf{P}'B(\hat{\boldsymbol{k}})) + \text{diag}(\mathbf{P}'\boldsymbol{k}) - \mathbf{P}'\text{diag}(B(\hat{\boldsymbol{k}}))\mathbf{P} - \mathbf{P}'\text{diag}(\boldsymbol{k})\mathbf{P}$$
$$= V(\hat{\boldsymbol{y}}) + \text{diag}\left(\mathbf{P}'\left(\mathbf{P}' - \mathbf{I}\right)\boldsymbol{k}\right) - \mathbf{P}'\text{diag}((\mathbf{P}' - \mathbf{I})\boldsymbol{k})\mathbf{P}.$$

In the last line, we used Expression (11). We also used the fact that $B\left(\hat{\boldsymbol{k}}\right) = \left(\mathbf{P}' - \mathbf{I}\right)\boldsymbol{k}$, by analogy with Expression (9). This shows that, in the presence of classification errors, $E\left\{\hat{B}^*_\infty(\hat{\boldsymbol{y}})\right\} \neq B(\hat{\boldsymbol{y}})$ and $E\left\{\hat{V}^*_\infty(\hat{\boldsymbol{y}})\right\} \neq V(\hat{\boldsymbol{y}})$.

For the special case that $\mathbf{P}$ has the Form (1), we can simplify the above expression for $E\left\{\hat{B}^*_\infty(\hat{\boldsymbol{y}})\right\}$ to:

$$E\left\{\hat{B}^*_\infty(\hat{Y}_h)\right\} = pB(\hat{Y}_h) + \frac{1-p}{H-1}\sum_{g\neq h} B(\hat{Y}_g) = \left(p - \frac{1-p}{H-1}\right)B(\hat{Y}_h). \tag{15}$$

Here, we used the fact that the overall total turnover $Y = \sum_{h=1}^{H} Y_h = \sum_{i=1}^{N} y_i$ is not affected by classification errors; hence, $\sum_{h=1}^{H} \hat{Y}_h = Y$ and $\sum_{g\neq h} B(\hat{Y}_g) = -B(\hat{Y}_h)$. A similar, slightly more tedious derivation shows that, in this special case:

$$E\{\hat{V}^*_\infty(\hat{Y}_h)\} = \left(p - \frac{1-p}{H-1}\right)V(\hat{Y}_h) + \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K, \tag{16}$$

with $K = \sum_{h=1}^{H} K_h = \sum_{i=1}^{N} y_i^2$.

To derive the bias-corrected bootstrap estimators (7) and (8), we rearrange Expressions (15) and (16) as follows:

$$B(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1} E\{\hat{B}^*_\infty(\hat{Y}_h)\}$$

and

$$V(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}\left[E\left\{\hat{V}^*_\infty(\hat{Y}_h)\right\} - \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\right].$$

Replacing $E\{\hat{B}^*_\infty(\hat{Y}_h)\}$ and $E\{\hat{V}^*_\infty(\hat{Y}_h)\}$ in the right-hand sides by their respective (unbiased) estimators $\hat{B}^*_R(\hat{Y}_h)$ and $\hat{V}^*_R(\hat{Y}_h)$, we obtain Expressions (7) and (8). We can also obtain analytical versions of these bias-corrected bootstrap estimators by using (13) and (14):

$$\hat{B}^{*}_{\infty,BC}(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}(1-p)\left\{\overline{\hat{Y}}^{(-h)} - \hat{Y}_h\right\}, \tag{17}$$

$$\begin{aligned}\hat{V}^{*}_{\infty,BC}(\hat{Y}_h) = \left(p - \frac{1-p}{H-1}\right)^{-1}\Bigg[&p(1-p)\hat{K}_h + \frac{1-p}{H-1}\left(1 - \frac{1-p}{H-1}\right)\sum_{g\neq h}\hat{K}_g \\ &- \frac{(1-p)^2}{H-1}\left(1 + p - \frac{1-p}{H-1}\right)K\Bigg].\end{aligned} \tag{18}$$

6. References

Bakker, B.F.M. and P.J.H. Daas. 2012. "Methodological Challenges of Register-Based Research." *Statistica Neerlandica* 66: 2–7. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00505.x.

Bethlehem, J. 2009. *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: Wiley.

Bryant, J. and P. Graham. 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., J. Davies, D. Lewis, A. Van Delden, P. Daas, and J.-M. Frost. 2013. *Guidance on the Accuracy of Mixed-Source Statistics. Deliverable 6.3/2011 of ESSnet Admin Data*. Available at: http://essnet.admindata.eu/WorkPackage/ShowAllDocuments?objectId=4257 (accessed December 2013).

Chamberlain, J. and E. Schulte Nordholt. 2004. "The Results of the 2001 Census in the Netherlands, the United Kingdom and Some Other European Countries." In *The Dutch Virtual Census of 2001, Analysis and Methodology*, edited by E. Schulte Nordholt, M. Hartgers and R. Gircour, 225–241. Statistics Netherlands: Voorburg/Heerlen.

Delden, A. van and P.P. de Wolf. 2013. "A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data." In Proceedings of the Conferences on New Techniques and Technologies for Statistics, March 5–7 2013. Brussels. Available at: http://www.cros-portal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (accessed December 2013).

Demnati, A. and J.N.K. Rao. 2009. "Linearization Variance Estimation and Allocation for Two-Phase Sampling under Mass Imputation." Paper for the Federal Committee on Statistical Methodology Research Conference, November 2–4, Washington, DC. Available at: http://www.fcsm.gov/09papers/Demnati_VI-C.pdf (accessed December 2013).

De Munnik, D., M. Illing, and D. Dupuis. 2013. "Assessing the Accuracy of Non-Random Business Conditions Surveys: a Novel Approach." *Journal of the Royal Statistical Society, Series A*, 176: 371–388. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01035.x.

DiCiccio, T.J. and B. Efron. 1996. "Bootstrap Confidence Intervals." *Statistical Science* 11: 189–228. Doi: http://dx.doi.org/10.1214/ss/1032280214.

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

Kuijvenhoven, L., and S. Scholtus. 2011. *Bootstrapping Combined Estimator Based on Register and Sample Survey Data*. Discussion paper 201123. The Hague/Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/06202B2A-B6C1-40CC-B25B-4022B7712E59/0/2011x1023.pdf (accessed December 2013).

Särndal, C.-E., B. Swensson, and J.H. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J. and R.R. Sitter. 1996. "Bootstrap for Imputed Survey Data." *Journal of the American Statistical Association* 91: 1278–1288. Doi: http://dx.doi.org/10.1080/01621459.1996.10476997.

UNECE. 2007. *Register-Based Statistics in the Nordic Countries, Review of Best Practices with Focus on Population and Social Statistics*. New York: United Nations. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (accessed August 2015).

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

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

Zhang, L.-C. 2012b. "On the Accuracy of Register-Based Census Employment Statistics." Paper presented at the European Conference on Quality in Official Statistics (Q2012), May 30–June 1, Athens. Available at: http://www.q2012.gr/articlefiles/sessions/23.4_Zhang_AaccuracyRegisterStatistics.pdf (accessed December 2013).