Estimating from mixed sources with incomplete coverage

Joep Burger[1,3], Arnout van Delden[2], Piet Daas[1], and Pieter Vlag[2]
[1] Statistics Netherlands, Heerlen, The Netherlands
[2] Statistics Netherlands, The Hague, The Netherlands
[3] Corresponding author: Joep Burger, e-mail: j.burger@cbs.nl

Abstract

For policymakers and alike, it is crucial to distinguish between socio-economic trends in official statistics and noise caused by various sources of error in the statistical process. This has become more difficult as official statistics are increasingly based on a mix of sources that typically do not involve probability sampling. We apply a resampling method to assess the sensitivity of mixed-source statistics to source-specific non-sampling errors. The method can be used to compare industries and releases, and can assist in deciding how to allocate resources in the statistical process. The example suggests that shifting classification resources from small and medium enterprises to large enterprises has virtually no net effect on accuracy, because the gain in precision is offset by the creation of bias.

Keywords: accuracy, bootstrap resampling, non-sampling error, register-based statistics, short-term business statistics, timeliness

1. Introduction

Official statistics provide statistical information on a country's social and economic development to policymakers, researchers and the general public. Traditionally, the information is collected through sample surveys, where probability theory can be applied to make inference about the whole population based on a random sample of the population. Although theoretically sound, sample surveys suffer from practical problems. Among others, they are expensive and they burden respondents with questionnaires.

Administrative registers are another source of information for official statistics. They provide a population frame from which samples can be drawn, and auxiliary information that can be used to correct for selective non-response in sample surveys. Moreover, statistics can be produced entirely based on administrative registers (UNECE 2007). Register-based or virtual censuses cost one or two orders of magnitude less per inhabitant than a traditional census (Chamberlain and Schulte Nordholt 2004) without any additional burden on respondents. On the other hand, administrative registers are not designed for statistical purposes. They may suffer from selective undercoverage, and the administrative units and variables may not match statistical definitions (Bakker and Daas 2012).

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 micro-integration processing. It is not clear, however, how to assess the accuracy of such mixed-source estimators. Since register data are not based on random sampling, no sampling errors are made and the theoretical framework from survey methodology does not apply. This does not imply, however, that the estimate is error-free. Other sources of non-sampling error remain (Zhang 2012).

In this contribution, we apply a bootstrap resampling method to assess the sensitivity of mixed-source statistics to source-specific non-sampling errors. We use a case study on quarterly turnover for the short-term business statistics (STS). We limit the results

to turnover level, but the methods described can also be applied to changes in turnover.

2. Methods

At Statistics Netherlands, quarterly turnover for STS is based on a mix of survey and administrative data. The turnover of most businesses is indirectly obtained from the administrative source VAT, whereas the statistical units underlying the largest businesses are directly observed through a census survey. The rationale of this design is that for larger businesses, the administrative unit is not uniquely linked to one statistical unit. Early estimates typically need to be produced before all survey and administrative data are available. The missing data are imputed using past information. Because no samples are drawn and missing data are imputed, no complicated design-based or model-based estimators are required to make inference 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.

We will focus on nine industries of economic activity (Fig. 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).

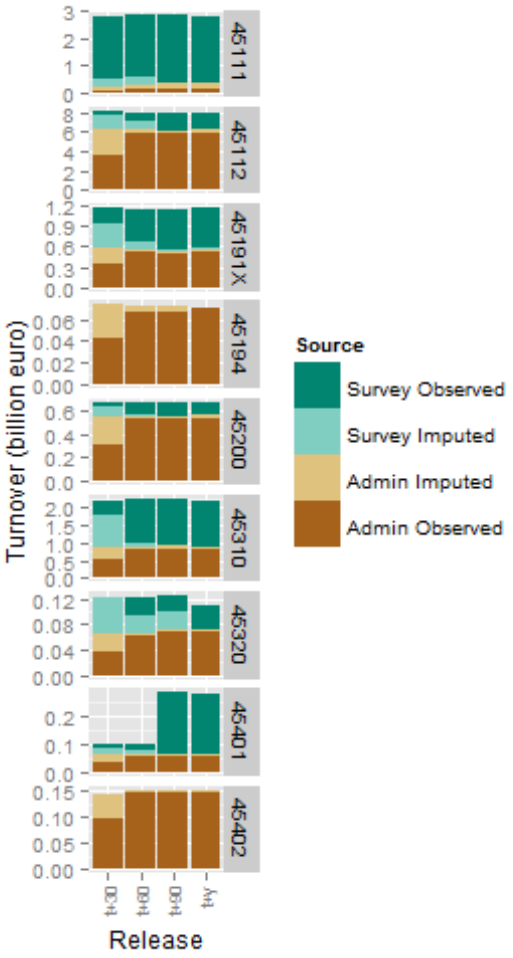

Figure 1 Mixed-source estimates of quarterly turnover as a function of time since end of reference period (third quarter of 2011) for nine industries within the Dutch particularization of NACE Rev. 2 within Division 45. Note that the y-axes are scaled independently between industries.

In this paper, we assess the sensitivity of these estimates to source-specific non-sampling errors, focusing on errors along the line of representation. Representation errors can be divided into coverage errors and classification errors. A coverage error occurs when a unit is unjustly included (overcoverage) or excluded (undercoverage) from the target population. A classification error occurs when the unit falls in the wrong category of a classification such as size class or economic activity. For a given category, misclassification may be regarded as a coverage error.

According to an internal Service Level Agreement, the three-digit NACE code should be correct for at least 95% of large enterprises (survey stratum) and 65% of small and medium enterprises (admin stratum). We applied these figures at industry level. 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 can then define two source-specific 9×9 transition matrices (scenario 1):

$$p^{\text{survey}} = \begin{pmatrix} 0.95 & 0.01 & \cdots \\ 0.01 & 0.95 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \text{ and}$$

$$p^{\text{admin}} = \begin{pmatrix} 0.65 & 0.04 & \cdots \\ 0.04 & 0.65 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

By switching the matrices between sources, we also studied what would happen if instead 65% of large enterprises (survey stratum) and 95% of small and medium enterprises (admin stratum) would be correctly classified for economic activity (scenario 2).

Using this input, we drew a new industry code for each unit from these transition matrices, recalculated the population parameter per industry, and repeated this a large number of times: 10,000 simulations per estimate, which seemed sufficient for confidence intervals to converge (not shown). The central limit theorem states that the distribution of the simulated replicates from this resampling method will tend towards a normal distribution.

The bias of our estimator caused by misclassification was estimated by the difference between the average of the simulated replicates and the total turnover estimated from observed and imputed data. The variance of our estimator caused by misclassification was estimated by the variance between the simulated replicates. The square of the bias and the variance were added, resulting in the mean square error (MSE) as a measure of accuracy. The square root (R) was taken to revert to the unit of the data (euro), and it was normalized (CV) to the total turnover estimated from observed and imputed data to make estimates comparable between releases and industries.

3. Results

Simulations under scenario 1 show that source-specific misclassification can result in strongly biased estimates (Fig. 2). Our dataset contains one large industry (45112), which is overestimated if some units in the small industries are misclassified. The small industries are underestimated if some units in the large industry are misclassified. In industry 45401 late estimates are more accurate than early estimates because they are based on more units with a likely correct NACE code (Fig. 1). In the other industries there is no effect of release on accuracy because the ratio between survey and administrative data remains fairly constant.

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

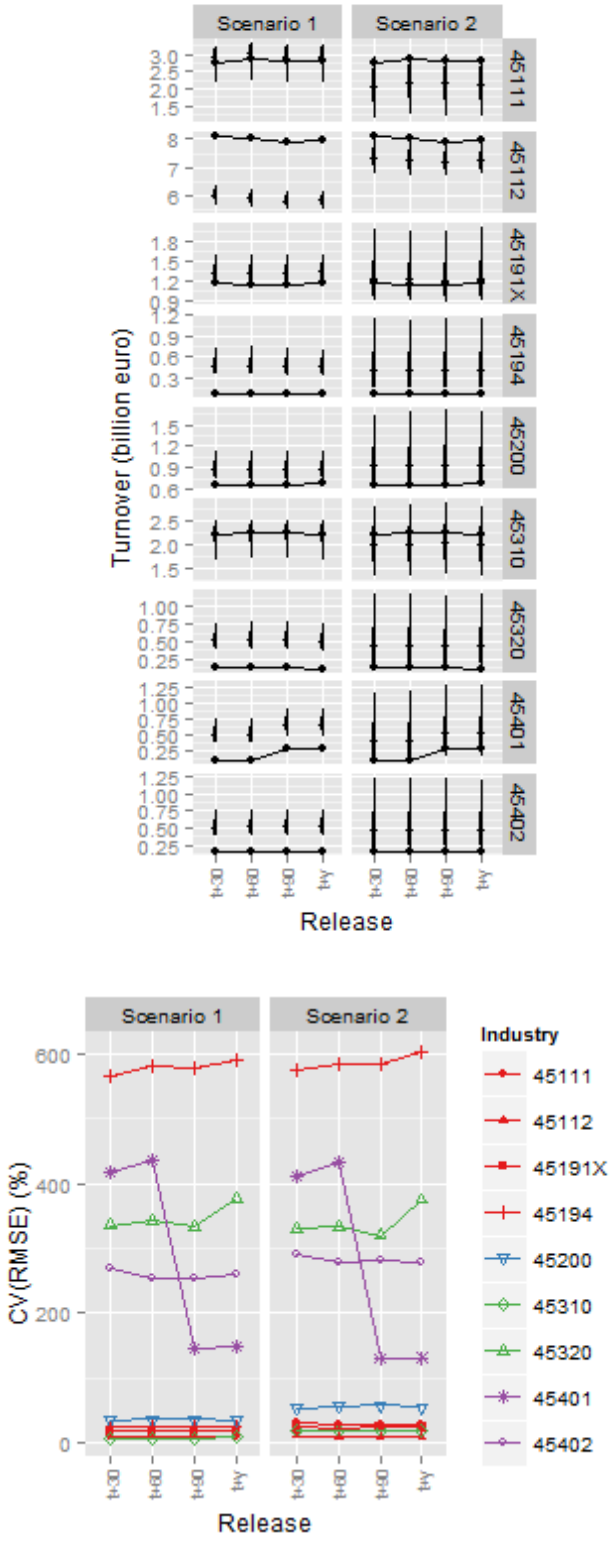

Figure 2 Sensitivity of mixed-source estimates to source-specific classification error. Top: quarterly turnover per industry and release estimated from observed and imputed data (dots and lines), and simulated mean (dashes) ± SD (thick bars), and 2.5th and 97.5th percentiles (thin bars) using 10,000 simulations per estimate. Note that the y-axes are scaled independently between industries. Bottom: 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).

4. Discussion

For policymakers and alike, it is crucial to distinguish between socio-economic trends in official statistics and noise caused by various sources of error in the statistical process. This has become more difficult as official statistics are increasingly based on a mix of sources that typically do not involve probability sampling. We have described such a case study where statistical units underlying large enterprises are directly observed through a census survey and turnover of small and medium enterprises is indirectly obtained from the tax register.

The resampling method described here provides insight into the sensitivity of mixed-source statistics to source-specific non-sampling error. It does not provide absolute estimates of accuracy, but can be used to compare industries and releases, and can assist in deciding where to invest resources into the statistical process. The example we have shown suggests that shifting classification resources from small and medium enterprises to large enterprises has virtually no net effect on accuracy, 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 changes in turnover, because the creation of bias in both time points is annihilated, whereas the gain in precision is not.

The resampling method could be adapted to specific situations or needs. For example, the transition matrices could be parameterized according to quality standards. The method could also be used to study the sensitivity of estimates to other sources of non-sampling error, such as measurement errors, or to a combination of (interacting) non-sampling errors. Another extension could be to assess the effect on the relative accuracy of changes over time.

Acknowledgments

We thank Arjen de Boer for providing the raw data, and Marjan Jongen, Willem Heijnen, Ton Bonné and Jeffrey Hoogland for discussion. This work was supported by the ESSnet on the Use of Administrative and Accounts Data for Business Statistics.

References

Bakker, B.F.M. and P.J.H. Daas (2012). Methodological challenges of register-based research. *Statistica Neerlandica*, 66, 2-7.

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:* E. Schulte Nordholt, M. Hartgers and R. Gircour (eds.). The Dutch Virtual Census of 2001. Analysis and Methodology. Statistics Netherlands, Voorburg/Heerlen.

UNECE (2007). *Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics.* United Nations, New York.

Zhang, L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.