Small area estimation of turnover of the Structural Business Survey

12

Sabine Krieg, Virginie Blaess, Marc Smeets

Discussion paper (201203)

Statistics Netherlands

The Hague/Heerlen, 2012

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 cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10–2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201203 X-10

Small area estimation of turnover of the Structural Business Survey

Summary: Since design-based methods like the generalized regression estimator have large design variances in the case of small sample sizes, model-based techniques can be considered as an alternative. In this paper a simulation study is carried out where small area estimation techniques, based on multilevel modelling, are applied to the variable turnover of the Structural Business Survey of Statistics Netherlands. By applying the EBLUP, the accuracy of the estimates can be substantially improved compared to the generalized regression estimator. The EBLUP estimates, however, are biased, which is partly caused by the skewed distribution of the variable turnover. It is found that by transforming the target variable both skewness and bias can be substantially reduced, whereas the variance increases. As a result, the accuracy is slightly improved compared to the EBLUP.

Keywords: small area estimation, multilevel models, skewness, transformation, bootstrapping

1 Introduction

Traditionally, statistical offices like Statistics Netherlands prefer design-based methods, for example the generalized regression (GREG) estimator, to produce estimates from survey samples. The advantage of these methods is that the estimates are always approximately design-unbiased. GREG estimators, however, have relatively large design variances in the case of small sample sizes. Therefore, a few years ago a research project was started to investigate how small area estimators can be applied to the surveys of Statistics Netherlands. These estimators rely explicitly on a statistical model that borrows information from other subpopulations or from preceding periods. Two approaches of small area estimation are explored at Statistics Netherlands. The first one is based on a linear mixed model and borrows information from other subpopulations, the second one is based on a structural time series model and borrows information from preceding periods. Both approaches have resulted in the implementation of small area estimators in the production process of the Labour Force Survey, first for the production of very detailed annual figures for municipalities (based on linear mixed models) and second for the production of monthly figures for six domains (based on structural time series models). Furthermore, the linear mixed model approach is applied to the National Safety Monitor (see Boonstra et al., 2008, van den Brakel and Krieg, 2009, Buelens and Benschop, 2009 for more information about these applications). Other national statistical institutes (NSIs) have also started to apply small area estimators in the production of official statistics. ESSnet (2011) provides an overview of applications to social statistics for many European and some non-European NSIs. Furthermore, this paper describes the research of the NSIs concerning SAE, which eventually will lead to more applications.

Note that the linear mixed model approach can be extended by a time component to borrow information from the past (Rao and Yu, 1994), whereas the time series approach can be extended by modelling correlations between domains to borrow information from these domains. For a comprehensive overview of SAE techniques see Rao (2003). An overview of new developments in SAE since 2003 can be found in Pfeffermann (2010).

In the above-mentioned applications of SAE at Statistics Netherlands, the target variables are categorial. In many other surveys of Statistics Netherlands, the target variable is continuous. Based on these surveys, many subpopulations with relatively small sample sizes are often considered and consistent estimates on different levels are needed. In this research project, SAE techniques are therefore applied to data with a continuous target variable; the variable turnover of the Structural Business Survey (SBS) is considered as an example. In this paper, the number of working persons is the most important auxiliary variable. It appears that the target variable is skewly distributed, as are the residuals of the linear mixed model with this auxiliary information. By transforming the target variable, the skewness can be reduced. A small area estimator based on such a transformation is developed in this paper. This new estimator, which is similar to one of the estimators developed in Chandra and Chambers (2011), the EBLUP (Rao, 2003) and the GREG estimator (Särndal et al., 1992) are applied to the SBS in a simulation study.

In another part of the research project, a situation with auxiliary information which is highly correlated to the target variable is investigated (Smeets et al., 2011). In that situation, the residuals of the linear mixed model are symmetrically distributed, which means that there is no need to transform the target variable.

In Section 2, the SBS and the sampling design of this survey are described. Then the different estimation methods are presented in Section 3. In Section 4, a simulation study is described where the methods are applied to the SBS. Finally, some conclusions are drawn and the next steps of the project are described in Section 5. Some background information, more detailed results and technical details are added in the appendix.

2 The Structural Business Survey

The SBS is an annual survey with the Dutch enterprises as target population. It measures the total production and describes the cost-benefit structure of the different sectors in the Netherlands. Almost all sectors are included in the survey. The most important exceptions are agriculture, the financial sector and government. The statistical unit is the enterprise.

The Standard Industrial Classification (SIC) provides a common statistical classification of economic activities of the enterprises. Based on the SIC the enterprises are divided into industries. Industries are combinations of SIC-codes and form the most important publication level for the SBS. There are also figures published at higher aggregation levels, e.g. the level of sectors. Another classification of the enterprises is given by size classes (SCs), which are defined by the number of working persons (WP). In the production process of the Dutch SBS, the Dutch classification of SCs is used. The same classification is used in this paper.

In this paper, the sector of the retail trade is considered. This sector consists of 20 industries. In the simulation of this paper, the variable tax-turnover is considered instead of the survey variable turnover. These variables are highly correlated, but whereas the last variable is only available for the sample, the first variable is available for almost the entire population. Enterprises with unknown tax-turnover are removed from the population for this project. Data of 2007 are considered. The methods are investigated in a situation where the following auxiliary information is known: Industry (categorial), SC (categorial), WP (continuous).

The sampling design used in this paper is similar to the design which is used for the SBS of 2009. It is a stratified design, where the strata are combinations of industries and SCs. In each stratum a simple random sample without replacement is drawn. The inclusion probability is generally larger for larger SCs and SC6 to SC9 (enterprises with more than 49 WP) are drawn with inclusion probability 1. Since estimates are needed only for the sample part, only SC1 - SC5 are considered in this paper. The total sample size is 5074, with sample sizes per industry between 21 and 769. In Appendix A the definition of the SCs, a short description of the industries of the retail trade, and information about the sampling design can be found.

2.1 Notation

The finite population with N elements is divided into m subpopulations (the industries) which are called domains in the small area estimation framework.

A stratified sample is drawn with H strata in all domains $j = 1, ..., m$, the strata are denoted by h. The observed value of the target variable for unit i in domain j is given by y_{ij}. The sample size in stratum h is denoted by n_h and the population size in stratum h by N_h. The total sample size and the population size in domain j are denoted by n_j and N_j, respectively. Similarly, n_{hj} and N_{hj} are the sample size and the population size in stratum h of domain j.

There are several explanatory variables for unit i in domain j given by the vector $\mathbf{x}_{ij} = (x_{ij}^1, ..., x_{ij}^p)^{\mathrm{t}}$. In order to add the constant to the model (weighting model for the GREG estimator, or the linear mixed model) it can be assumed that $x_{ij}^1 = 1$. Estimates for the population total $Y_j^{\mathrm{tot}} = \sum_{i=1}^{N_j} y_{ij}$ for target variable y for all domains $j = 1, ..., m$ have to be computed. In the example of the retail trade, we have $N = 63958$, $H = 5$, and $m = 20$.

3 Methods

3.1 Generalized regression estimator

The GREG estimator is a well-known design-based model-assisted estimator which is approximately design-unbiased (Särndal et al., 1992). The GREG estimator of the unknown population total Y_j^{tot} for domain j is given by:

$$\widehat{Y}_j^{\mathrm{GREG}} = \widehat{Y}_j^{\mathrm{HT}} + (\mathbf{X}_j^{\mathrm{tot}} - \widehat{\mathbf{X}}_j^{\mathrm{HT}})\widehat{\boldsymbol{\beta}}_j, \tag{1}$$

where the Horvitz-Thompson estimators are given by $\widehat{Y}_j^{\mathrm{HT}} = \sum_{h=1}^{H} \frac{N_{hj}}{n_{hj}} \sum_{i=1}^{n_{hj}} y_{ij}$, $\widehat{\mathbf{X}}_j^{\mathrm{HT}} = \sum_{h=1}^{H} \frac{N_{hj}}{n_{hj}} \sum_{i=1}^{n_{hj}} \mathbf{x}_{ij}^t$, and $\mathbf{X}_j^{\mathrm{tot}} = \sum_{i=1}^{N_j} \mathbf{x}_{ij}^t$ is the p-vector of population totals of the auxiliary information in domain j. Furthermore, $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^{\mathrm{t}}\mathbf{D}_j\mathbf{X}_j\right)^{-1}\left(\mathbf{X}_j^{\mathrm{t}}\mathbf{D}_j\mathbf{y}_j\right)$ with $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{n_jj})^{\mathrm{t}}$, $\mathbf{X}_j = (\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{n_jj})^{\mathrm{t}}$, $\mathbf{D}_j = \mathrm{blockdiag}(\frac{N_{1j}}{n_{1j}}\mathbf{I}_{n_{1j}}, \ldots, \frac{N_{Hj}}{n_{Hj}}\mathbf{I}_{n_{Hj}})$ with $\mathbf{I}_{n_{hj}}$ the identity matrix of size n_{hj}.

3.2 The EBLUP estimator

The EBLUP estimator (Rao, 2003) which is apllied in this paper, is based on linear mixed models of the following form:

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{t}}\boldsymbol{\beta} + \mathbf{x}_{ij1}^{\mathrm{t}}\boldsymbol{\vartheta}_j + e_{ij} \tag{2}$$

with $\boldsymbol{\vartheta}_j = (\vartheta_j^1, \vartheta_j^2, \ldots, \vartheta_j^{p_1})^{\mathrm{t}} \sim \mathcal{N}(0, \boldsymbol{\Theta})$, $e_{ij} \sim \mathcal{N}(0, k_{ij}^2\sigma_e^2)$ for $i = 1, \ldots, N_j$ and $j = 1, \ldots, m$. Here the vector of auxiliary information $\mathbf{x}_{ij} = (x_{ij}^1, ..., x_{ij}^p)^{\mathrm{t}}$ of length p is partitioned into two vectors $\mathbf{x}_{ij1} = (x_{ij}^1, ..., x_{ij}^{p_1})^{\mathrm{t}}$ and $\mathbf{x}_{ij2} = (x_{ij}^{p_1+1}, ..., x_{ij}^p)^{\mathrm{t}}$ of length p_1 and $p - p_1$ respectively. Furthermore $\boldsymbol{\beta}$ is a p-vector of fixed effects and $\boldsymbol{\vartheta}_j$ is a p_1-vector of random effects.

This model allows for differences across small areas between the intercepts as well as between the slopes which are selected in $\mathbf{x}_{ij1}$ (if the intercept is part of $\mathbf{x}_{ij1}$, the intercept is allowed to vary across small areas).

Since the population is not known in general, the model has to be estimated by means of a sample. The estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\vartheta}}$, $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\sigma_e^2}$ of the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\vartheta}_j$ as well as the unknown variance parameters $\boldsymbol{\Theta}$ and σ_e^2 are computed using the procedure lmer in the software package R (see R Development Core Team, 2009). Within this procedure, the restricted maximum likelihood (REML) method is used to estimate the parameters.

The population of size N_j is partitioned into sampled and non-sampled parts for each domain j, where the first n_j elements are in the sample. The information about the sampled part is directly used in the estimator, the non-sampled part is estimated based on the model. Thus the following estimator for the population total of domain j is found

$$\widehat{Y}_j^{\text{EBLUP}} = \sum_{i=1}^{n_j} y_{ij} + \widehat{\boldsymbol{\beta}}^{\text{t}} \sum_{i=n_j+1}^{N_j} \mathbf{x}_{ij} + \widehat{\boldsymbol{\vartheta}}_j^{\text{t}} \sum_{i=n_j+1}^{N_j} \mathbf{x}_{ij1}. \tag{3}$$

3.3 Small area estimation with transformed target variable

The standardized residuals $\frac{e_{ij}}{k_{ij}}$ in the linear mixed model are assumed to be normally distributed. In the model evaluation of some applications, however, it is found that they are skewed, which means that the model is misspecified. This problem can be approached by modelling a transformed value $\tau_{y_{ij}} = f(y_{ij})$ of the target variable y_{ij}. So a model

$$\tau_{y_{ij}} = \mathbf{x}_{ij}^{\text{t}}\boldsymbol{\beta}_\tau + \mathbf{x}_{ij1}^{\text{t}}\boldsymbol{\vartheta}_{j,\tau} + e_{ij,\tau}, \tag{4}$$

is estimated. The variables of this model are defined similarly as in Section 3.2, now with the subscript τ. Especially, $\boldsymbol{\vartheta}_{j,\tau} \sim \mathcal{N}(0, \boldsymbol{\Theta}_\tau)$ and $e_{ij,\tau} \sim \mathcal{N}(0, k_{ij,\tau}^2\sigma_{e,\tau}^2)$ for $i = 1, \ldots, N_j$ and $j = 1, \ldots, m$. For the fixed and random effects $\mathbf{x}_{ij}^{\text{t}}$ and $\mathbf{x}_{ij1}^{\text{t}}$ the same notation as in Section 3.2 is used. Suitable parameters have to be chosen in the specification of each model.

This model is estimated in the same way as the model in Section 3.2. Then, the value y_{ij} can be predicted by $y_{ij}^{\text{pred}} = f^{-1}(\mathbf{x}_{ij}^{\text{t}}\widehat{\boldsymbol{\beta}}_\tau + \mathbf{x}_{ij1}^{\text{t}}\widehat{\boldsymbol{\vartheta}}_{j,\tau} + e_{ij,\tau})$ with unknown $e_{ij,\tau} \sim \mathcal{N}(0, k_{ij,\tau}^2\sigma_{e,\tau}^2)$. In the situation of the EBLUP (Section 3.2), the residuals e_{ij} can be ignored since they are normally distributed with expectation 0. In the actual situation with a non-linear transformation f, ignoring is not appropriate since $E(f^{-1}(z + e_{ij,\tau})) \neq E(f^{-1}(z))$, despite $E(e_{ij,\tau}) = 0$.

Thus y_{ij}^{pred} is approximated by

$$\hat{y}_{ij}^{\text{pred}} = \frac{1}{r-1} \sum_{l=1}^{r-1} (f^{-1}(\mathbf{x}_{ij}^{\text{t}}\widehat{\boldsymbol{\beta}}_\tau + \mathbf{x}_{ij1}^{\text{t}}\widehat{\boldsymbol{\vartheta}}_{j,\tau} + g_{ij}^{-1}(\frac{l}{r}))), \tag{5}$$

with g_{ij} the distribution function of $\mathcal{N}(0, k_{ij,\tau}^2 \widehat{\sigma}_{e,\tau}^2)$ and $r = 5000$, which is large enough for an accurate approximation in this application.

Formula (5) is explicitly based on the assumption that the residuals are normally distributed with variance $k_{ij,\tau}^2 \widehat{\sigma}_{e,\tau}^2$. If this assumption is not correct, the predictions can be biased. In such a situation, it can be investigated whether the empirical distribution of the standardized residuals can be used instead of the normal distribution. See for example Duan (1983) for the use of the empirical distribution in another situation.

Thus the following estimator for the population total of domain j is found

$$\widehat{Y}_j^{\mathrm{SAE},\tau} = \sum_{i=1}^{n_j} y_{ij} + \sum_{i=n_j+1}^{N_j} \hat{y}_{ij}^{\mathrm{pred}}. \tag{6}$$

The estimator is called transformed SAE in this paper.

Alternatively to formula (5), it is also possible to compute the expected value of transformed normally distributed data, see for example Neyman and Scott (1960) or Bradu and Mundak (1970). For specific transformations as the logarithmic transformation, simple formulas can be found. The approximation (5) is, however, applicable for all transformations and easy to program, and therefore, preferred here.

In Chandra and Chambers (2011), a similar estimator is developed, using the expected value of transformed normally distributed data. They point out that both approaches yield biased estimates since $\boldsymbol{\beta}_\tau$, $\boldsymbol{\vartheta}_{j,\tau}$ and $\sigma_{e,\tau}^2$ are estimated, whereas formula (5) assumes these values to be known. In Chandra and Chambers (2011), a correction term for this bias is developed for the logarithmic transformation. In this paper, it is not corrected for this bias for different reasons; first and most importantly, since the bias is acceptable in the considered situation, as will be seen in Section 4. Second, a probably larger bias can be caused by model misspecification, especially when the transformation is not chosen correctly. Third, it is not directly obvious how this correction term can be adapted for a general transformation. Forth, the correction term does not account for the estimation of the $k_{ij,\tau}$. In a situation where the bias is not acceptable, a method which corrects for all bias simultaneously could be considered. Such methods are developed for other estimators, see for example Chambers et al. (2011), Gershunskaya and Lahiri (2011).

3.4 Methods to estimate the mean squared error

For the GREG estimator and for the EBLUP, analytic methods to estimate the mean squared error (MSE) are known, which are worked out for the situation of this paper in Appendix B.

For the transformed SAE, no analytic formula for the MSE is developed so far. As an alternative, bootstrapping can be applied (see Efron and Tibshirani, 1993 or Davidson and Hinkley, 1997 for a general introduction). Since the estimates for variance and MSE are more stable by parametric bootstrapping, especially in the case of small sample sizes and/or outliers in the sample, this method is often preferred in SAE, see for example Lahiri (2003), Butar and Lahiri (2003), Sinha and Rao (2009), Hall and Maiti (2006b). Parametric bootstrapping includes the assumption that the linear mixed model is true, since this model is used to create a population from which the bootstrap samples are drawn to estimate the MSE. Note that in this paper, the MSE is estimated from the distribution of the point estimates based on the bootstrap samples. In some other applications, for example Butar and Lahiri (2003), the bootstrap samples are used to correct the estimates based on a formula for bias.

To investigate the effects of serious model misspecification on bootstrapping, both parametric and non-parametric bootstrapping is applied on the EBLUP in a preliminary analysis. From this analysis it can be concluded that both parametric and non-parametric bootstrap estimates of variance and MSE can be unstable in a situation of model misspecification, and that the parametric bootstrap estimates of variance and MSE can be biased. Furthermore, non-parametric bootstrap estimates of the MSE of the GREG are compared with variance estimates for the GREG based on the analytic formula (B.1) in this preliminary analysis. For the GREG, both the non-parametric bootstrap estimates and the formula estimates are unstable due to the small sample sizes for the domains. See Appendix D for the results.

3.5 Consistency

Generally a reliable design-based estimate of the higher level (here the sector of the retail trade) is available, and it is desirable that the model-based estimates of the domains are consistent with this design-based estimate. In this paper, the sum of the GREG estimates for the m domains is used as sector estimate. There are different methods to achieve consistency. In this paper, the ratio adjustment

$$\widehat{Y}_j^{\text{adj,method}} = \frac{\widehat{Y}_j^{\text{method}} \sum_{d=1}^{m} \widehat{Y}_d^{\text{GREG}}}{\sum_{d=1}^{m} \widehat{Y}_d^{\text{method}}} \tag{7}$$

is used with EBLUP or transformed SAE as method. This simple formula does not take the variance of the model-based estimates into account. Alternatively, a Lagrange multiplier can be considered, where the adjustments depend on the variances. A disadvantage of this method is that these variances are estimated, which can result in an increase of the variances of the adjusted model-based estimates.

As another alternative, a small area estimation method could be developed where consistency is achieved directly. For some situations, such methods are described in the literature, see for example You and Rao (2002).

4 Simulation

In the simulation study of this paper the described methods are applied to the variable tax-turnover of the retail trade for 750 samples drawn from the population. Furthermore, for 250 of these samples, the variance and the MSE are estimated based on formulas (as far as available) and on bootstrapping. The bootstrap results are based on 200 bootstrap samples. First, the model specification is discussed in Section 4.1. Then the results are described, first the accuracy of the point estimates (Section 4.2), then some results about model evaluation (Section 4.3) and finally the estimation of variance and MSE are discussed (Section 4.4).

4.1 Model specification

The model is specified by describing the auxiliary information used in the fixed effects $\mathbf{X}$ and in the random effects $\mathbf{X}_1$, by specifying the parameters k_{ij}, and, if the transformed SAE is used, the transformation f.

In other applications, $k_{ij} = 1$ is often used. Preliminary analysis, however, has shown that the standardized residuals are heteroscedastic. This is not surprising, as large enterprises have a larger turnover, and it can be expected that the variation between large enterprises is also larger than the variation between small enterprises. Since there are also substantial differences between the domains, auxiliary information like WP is not sufficient to explain the heteroscedasticity. Therefore, the parameters k_{ij} and $k_{ij,\tau}$ are estimated based on the sample: the standard deviations of each stratum of the residuals of a regression model are used as parameters k_{ij} and $k_{ij,\tau}$ for all elements of this stratum. However, this standard deviation is based on the same sample and estimated with uncertainty. Since the estimated standard deviations are used explicitly in the estimation of the transformed SAE, the uncertainty of the regression model can strongly influence the model estimates. It can be expected that the accuracy of the transformed SAE can be improved in a situation where the estimation of the parameters $k_{ij,\tau}$ is not based on the sample. Therefore, standard deviations are also computed based on the entire population. Though this is not a realistic situation it will nevertheless show whether the accuracy of the estimates is improved substantially if the parameters $k_{ij,\tau}$ in the distribution function g_{ij} of formula (5) are exactly known. In order to compute the parameters k_{ij} and $k_{ij,\tau}$ in the regression model, the same auxiliary information is used as for the fixed effects in the linear mixed models.

To reduce positive skewness, a logarithmical transformation is often used. In this application, however, the standardized residuals are negatively skewed if this transformation is used. As an alternative, the transformation $f(x) = \sqrt[3]{x}$ is found as a simple transformation which reduces the skewness substantially.

The sampling design is stratified with SC $\times$ industry as strata. This design is taken into account by using SC in the fixed and random effects and by allowing the parameters to vary over the industries due to the random effects.

In the model without transformation, it is obvious to assume a linear relationship between WP and turnover. In this model, the following auxiliary information is used in the fixed effects and the random effects

$$1 + \mathrm{SC} + \mathrm{WP}. \tag{8}$$

After transforming turnover, it is likely that a transformation of WP is preferable above using WP directly as auxiliary information. Since a more or less linear relationship between the variable $\sqrt[8]{\mathrm{WP}}$ and the target variable is found, the following auxiliary information is used in the fixed effects and in the random effects

$$1 + \mathrm{SC} + \sqrt[8]{\mathrm{WP}}. \tag{9}$$

In other applications, more parsimonious random intercept models are used. Preliminary analysis, however, has shown that such models are not sufficient in the situation of the SBS since there are also some large differences between the slopes of the domains. In Appendix C, simulation results for random intercept models and for models with $k_{ij} = 1$ are presented. It can be concluded that the accuracy cannot be improved with these models compared to the GREG.

With more complex models, for example including interaction effects or functions of WP as $\sqrt{\mathrm{WP}}$, only small improvements seem to be possible, which was also investigated in a preliminary analysis. The results of this analysis are not shown in this paper.

4.2 Point estimates

In Table 1, Table 2, and Table 3 the relative bias, relative standard error and relative root mean squared error (RMSE) of the different methods (GREG, EBLUP, transformed SAE noted as sae_τ, transformed SAE with $k_{ij,\tau}$ based on the population noted as sae_τ^{pop} and consistent estimates of these three model based estimators noted with subscript c) for all domains j are shown.

These values are computed as $100 * \frac{Y_j^{\text{tot}} - \text{mean}(\hat{Y}_j^{\text{tot}})}{Y_j^{\text{tot}}}$, $100 * \frac{\text{se}(\hat{Y}_j^{\text{tot}})}{Y_j^{\text{tot}}}$, and $100 * \frac{\sqrt{\text{var}(\hat{Y}_j^{\text{tot}}) + \text{bias}(\hat{Y}_j^{\text{tot}})^2}}{Y_j^{\text{tot}}}$ respectively, with $\hat{Y}_j^{\text{tot}}$ the estimate of domain j based on one of the methods. Furthermore, the mean of the relative biases over all domains and the mean of the absolute relative biases are computed in Table 1. The mean of the relative standard errors and the mean of the relative RMSE over all domains are computed in Table 2 and Table 3.

As expected, the GREG is not or only slightly biased (column 3 of Table 1). On contrast, column 4 of Table 1 shows that the EBLUP is biased by at least 3% (positive or negative) for 50% of the domains with a maximum of more than 11% in domain 52413. By applying the transformed SAE, the bias is reduced. The use of $k_{ij,\tau}$ based on the population only slightly reduces the bias. In cases where the bias of (almost) all domains has the same sign, making the domain estimates consistent with an unbiased estimate on a higher level can strongly reduces the bias. In this application however, the consistent estimates are only slightly less biased.

The precision of the estimates (Table 2) is substantially improved by applying a method based on linear mixed models, compared to the GREG. For one domain, the transformed SAE is much more precise than the EBLUP (52610), whereas for some other domains, the EBLUP is substantially more precise (52321, 52422). For most domains, both methods are comparable; overall, the EBLUP is slightly more precise than the transformed SAE. The precision could be substantially improved if the $k_{ij,\tau}$ were known, as is shown by sae_τ^{pop}. The consistent estimates, however, are slightly less precise than the non-consistent estimates. This can be explained by Table 4: in this table the bias, standard error and RMSE of the sector estimates are computed, where the sector estimates are computed as sum of the estimates for the domains. The table shows that the model-based methods are more precise than the GREG, also for the entire sector.

Similarly to the precision (Table 2) the accuracy of the estimates is substantially improved by applying a method based on linear mixed models (Table 3). Whereas the transformed SAE is more accurate than the GREG for all domains, the EBLUP is less accurate than the GREG for 7 domains due to the bias of this method. Overall, the EBLUP is slightly less accurate than the transformed SAE. Furthermore the accuracy could be substantially improved when the $k_{ij,\tau}$ were known. The consistent estimates are slightly less accurate than the non-consistent estimates.

Domain	Y_j^{tot}	*greg*	*eblup*	sae_τ	sae_τ^{pop}	$eblup_{\text{c}}$	$sae_{\tau,\text{c}}$	$sae_{\tau,\text{c}}^{\text{pop}}$
52110	3969	0.11	1.20	-0.33	-0.13	-1.09	0.14	-0.01
52120	112	-0.38	3.14	-1.05	0.01	0.91	-0.58	0.13
52200	3326	-0.42	1.74	-1.54	-1.07	-0.53	-1.07	-0.96
52310	1457	-0.02	2.87	0.14	0.50	0.63	0.61	0.61
52321	644	-0.74	6.13	0.38	0.71	3.97	0.83	0.82
52330	71	-0.33	0.91	-2.14	-0.71	-1.37	-1.66	-0.59
52413	85	-0.49	-11.77	-5.38	-5.37	-14.36	-4.90	-5.25
52422	2732	0.04	3.67	-1.23	-0.85	1.44	-0.79	-0.75
52431	673	-0.21	3.24	-0.97	-0.56	1.00	-0.51	-0.45
52440	1988	0.38	2.03	-1.68	-1.42	-0.24	-1.21	-1.31
52450	1281	-1.44	6.43	3.34	3.94	4.24	3.76	4.02
52460	2049	-0.17	4.74	-0.05	0.30	2.54	0.41	0.41
52470	760	0.23	3.89	-0.44	-0.09	1.67	0.02	0.02
52485	2558	0.57	0.91	-0.82	-0.62	-1.38	-0.36	-0.51
52491	2692	-0.12	1.52	-1.08	-0.77	-0.75	-0.62	-0.66
52500	279	-0.07	-0.10	-0.79	-0.12	-2.44	-0.34	-0.02
52610	384	-0.21	-2.05	8.93	9.36	-4.41	9.34	9.47
52620	913	-0.15	-2.03	-1.87	-1.38	-4.39	-1.40	-1.26
52630	593	-0.13	3.04	-0.15	0.30	0.79	0.31	0.41
52700	434	0.32	-3.75	-1.20	-0.85	-6.15	-0.73	-0.74
Mean abs.		0.33	3.26	1.68	1.45	2.71	1.48	1.42
Mean		-0.16	1.29	-0.40	0.06	-1.00	0.06	0.17

Table 1. Population totals (in 1,000,000 euro), relative biases (in %) for 20 domains, mean of the relative bias over the 20 domains and mean of the absolute values of the relative bias over the 20 domains for 7 methods, simulated with 750 runs.

	greg	*eblup*	sae_τ	sae_τ^{pop}	$eblup_{\text{c}}$	$sae_{\tau,\text{c}}$	$sae_{\tau,\text{c}}^{\text{pop}}$
52110	3.18	2.77	2.82	3.04	3.22	2.99	3.38
52120	11.06	7.63	8.79	7.33	7.81	8.73	7.36
52200	4.64	4.04	3.88	3.03	4.33	4.01	3.37
52310	2.52	2.31	2.24	2.63	2.66	2.33	2.89
52321	14.63	5.23	8.57	5.75	5.52	8.71	5.96
52330	13.60	10.30	12.05	11.19	10.54	12.00	11.19
52413	13.42	9.99	10.86	8.82	10.40	10.86	8.94
52422	11.06	3.98	6.70	4.10	4.56	7.13	4.57
52431	7.10	6.08	6.46	5.80	6.42	6.51	5.96
52440	6.50	5.25	5.61	5.01	5.50	5.68	5.21
52450	16.99	12.00	12.04	7.67	12.59	12.23	8.04
52460	4.37	3.43	3.79	3.05	3.69	3.88	3.37
52470	6.59	5.90	5.78	5.71	6.16	5.81	5.83
52485	9.29	5.33	6.16	4.57	5.66	6.42	4.87
52491	5.46	4.81	3.83	2.89	5.12	4.01	3.25
52500	23.25	11.98	13.39	8.26	12.57	13.54	8.54
52610	16.68	16.66	10.61	7.23	17.09	10.62	7.30
52620	4.39	4.14	3.91	3.50	4.45	3.97	3.75
52630	9.15	6.80	7.21	5.56	7.22	7.27	5.83
52700	6.39	6.19	5.62	4.80	6.50	5.65	4.99
Mean	9.51	6.74	7.02	5.50	7.10	7.12	5.73

Table 2. Relative standard error (in %) for 20 domains and mean of the relative standard errors over the 20 domains for 7 methods, simulated with 750 runs.

	greg	*eblup*	sae_τ	sae_τ^{pop}	$eblup_{\text{c}}$	$sae_{\tau,\text{c}}$	$sae_{\tau,\text{c}}^{\text{pop}}$
52110	3.19	3.02	2.84	3.04	3.40	2.99	3.38
52120	11.06	8.25	8.86	7.33	7.87	8.75	7.36
52200	4.66	4.40	4.18	3.22	4.36	4.15	3.50
52310	2.52	3.68	2.25	2.67	2.73	2.41	2.95
52321	14.65	8.06	8.58	5.79	6.80	8.75	6.02
52330	13.61	10.34	12.23	11.21	10.62	12.11	11.20
52413	13.43	15.43	12.13	10.32	17.73	11.91	10.37
52422	11.06	5.42	6.81	4.19	4.78	7.17	4.63
52431	7.10	6.89	6.54	5.82	6.50	6.53	5.98
52440	6.51	5.63	5.86	5.20	5.50	5.80	5.37
52450	17.05	13.61	12.49	8.63	13.29	12.79	8.99
52460	4.38	5.85	3.79	3.07	4.48	3.90	3.39
52470	6.60	7.07	5.79	5.71	6.38	5.81	5.83
52485	9.31	5.40	6.22	4.62	5.82	6.43	4.89
52491	5.46	5.04	3.98	3.00	5.18	4.06	3.32
52500	23.25	11.98	13.42	8.26	12.80	13.55	8.54
52610	16.68	16.79	13.86	11.83	17.65	14.14	11.96
52620	4.39	4.61	4.33	3.76	6.25	4.21	3.96
52630	9.15	7.45	7.22	5.56	7.27	7.27	5.84
52700	6.39	7.24	5.75	4.88	8.95	5.69	5.05
Mean	9.52	7.81	7.35	5.91	7.92	7.42	6.13

Table 3. Relative RMSE (in %) for 20 domains and mean of the relative RMSE over the 20 domains for 7 methods, simulated with 750 runs.

	greg	*eblup*	sae_τ	sae_τ^{pop}
Relative bias	-0.07	2.19	-0.53	-0.18
Relative standard error	2.11	1.37	1.58	1.38
Relative RMSE	2.11	2.59	1.67	1.39

Table 4. Relative bias, relative standard error and relative RMSE of sector estimates (as sum of estimates of the 20 domains), for 4 methods, simulated with 750 runs.

4.3 Some remarks about model evaluation and model selection

By the linear mixed model, it is assumed that the residuals are normally distributed with variance $k_{ij}^2\sigma_e^2$ or $k_{ij,\tau}^2\sigma_{e,\tau}^2$ respectively. Furthermore, normality is assumed for the random effects. As mentioned before, the standardized residuals of the EBLUP are skewed to the right. This is the case for all samples in the simulation (skewness between 2.04 and 6.94, mean over all runs in the simulation 4.04). The skewness is larger in the smaller SCs than in the larger SCs (mean skewness of 5.47 in SC 1 up to 0.89 in SC 5 over all runs in the simulation). There are also substantial differences in skewness between the industries (mean skewness between 0.22 in domain 52310 and 8.24 in 52610). By the transformation, the skewness is substantially reduced (between -0.08 and 1.17, mean over all runs 0.36). In SC 1 the mean skewness is still positive (0.68), whereas in SC 5 it is now negative (-0.37).

Theoretically, it is possible to improve the model by choosing another transformation, which would further reduce the skewness and possibly improve the accuracy of the model estimates. It is, however, impossible to select the best transformation in a practical situation when only one sample is known. Therefore, a simulation with the selected transformation can give realistic results.

A formal test for normality is not carried out in this paper. However, symmetrical but not normally distributed standardized residuals can cause biased model estimates since formula (5) is based explicitly on the normality assumption.

The variances of the standardized residuals are quite similar for all SCs and for all industries in most of the simulations. This shows that the model accounts well for heteroscedasticity by using the k_{ij} in the variance structure of the residuals.

4.4 Estimation of variance and mean squared error

In this section, the estimation of the variance and the MSE is evaluated. Table 5 shows the relative RMSEs and the relative standard errors respectively, computed as $\frac{\sqrt{\overline{\text{mse}}}}{\text{rmse}_{\text{method}}}$ and $\frac{\sqrt{\overline{\text{var}}}}{\text{se}_{\text{method}}}$. Here $\overline{\text{mse}}$ and $\overline{\text{var}}$ are the mean of the MSE and variance estimates over 250 runs in the simulation (except for the GREG variance estimate based on the formula, which is applied in all 750 runs), $\text{rmse}_{\text{method}}$ is the empirical RMSE and $\text{se}_{\text{method}}$ is the empirical standard error (based on 750 runs of the simulation), where method is one of the methods GREG, EBLUP, or transformed SAE. The variance of the GREG is estimated by formula (B.1), the MSE of the EBLUP by formula (B.2), and the MSE and variance of the transformed SAE by parametric bootstrapping.

Values around 1 in Table 5 mean that the MSE and the variance are estimated (approximately) unbiased. As expected, the variance estimates for the GREG estimates are not (or only slightly) biased. The RMSE of the EBLUP is generally overestimated by a factor of maximal 1.81. In a few domains and for a few cases, the formula results in negative estimates of the MSE, these results are ignored. By parametric bootstrapping, the RMSE of the transformed SAE is overestimated in almost all domains by a factor of maximal 1.41. On the other hand, the standard error is underestimated in almost all domains, by a factor of minimal 0.66.

Domain	$greg^{\text{var}}$	$eblup^{\text{mse}}$	sae_{τ}^{mse}	sae_{τ}^{var}
52110	1.00	1.39	1.29	0.86
52120	0.91	1.09	0.97	0.66
52200	0.99	1.42	1.14	0.85
52310	1.03	0.79	1.26	0.86
52321	0.99	1.06	1.41	0.91
52330	0.91	1.09	1.28	0.90
52413	0.94	1.05	1.22	0.94
52422	0.99	1.30	1.23	0.96
52431	1.01	1.30	1.28	0.92
52440	1.03	1.55	1.34	1.00
52450	0.98	0.59	1.04	0.72
52460	0.99	0.88	1.19	0.85
52470	0.96	1.16	1.30	0.92
52485	0.97	1.50	1.38	0.96
52491	1.04	1.24	1.15	0.86
52500	0.98	1.39	1.29	0.94
52610	0.99	0.61	0.79	0.73
52620	0.96	1.81	1.15	0.88
52630	0.98	1.22	1.24	0.85
52700	1.03	1.37	1.35	0.98

Table 5. Root of the mean of the estimates for the variance of the GREG, for the MSE of the EBLUP and for the MSE and variance of the transformed SAE, divided by empirical standard error and RMSE, respectively.

For a more detailed analysis, boxplots are provided for variance, MSE and bias estimates of the transformed SAE, and for the MSE of the EBLUP. In the boxplots, the variance and MSE estimates are divided by the empirical variance and MSE respectively, which means that values around 1 again coincide with accurate estimates. The bias estimates based on bootstrapping are divided by the population totals. This boxplot can be compared with the empirical bias from Table 1. Figure 1 shows the MSE estimates of the EBLUP. The negative estimates of the MSE which are found in some cases are ignored in the figure. The figure shows that the MSE is overestimated by a factor around 2 for more than 50% of the runs in the simulation in around 50% of the domains. In some domains, the MSE is underestimated in most of the simulation runs. Only in a few runs, the MSE is overestimated by a factor larger than 5.

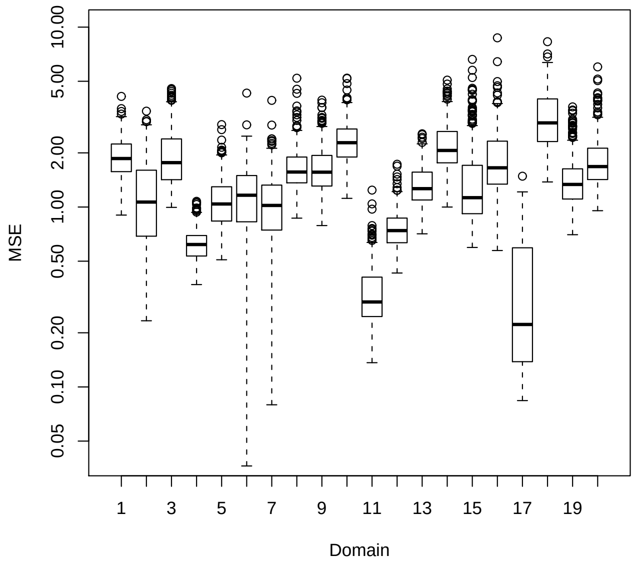

Figure 1. Estimation results MSE EBLUP by formula.

Variance, bias and MSE are estimated separately for the transformed SAE by parametric bootstrapping. First, variance and bias are discussed as components of the MSE. To get an impression of the influence of the number of bootstraps, results based on less bootstraps are also discussed. Figure 2 shows the relative variance estimates of the transformed SAE. As is already seen in Table 5, the variance is slightly underestimated. Furthermore, the variance estimates are relatively stable, compared to the MSE estimates of the EBLUP (Figure 1).

The variance estimates based on 50, 100, 150 and 200 bootstraps are shown side by side (most left: 50 bootstraps, most right: 200 bootstraps) for each of the domains 1 - 10 in Figure 3. It can be concluded that the precision of the variance estimates is improved by increasing the number of bootstraps. The results for domain 11-20 are similar and omitted in this paper. More than 200 bootstraps could further improve the precision, which is not investigated in the simulation study since the computation time is too large.

The right plan of Figure 4 shows the relative bias estimates of the transformed SAE. For all domains, the bias is estimated around zero with a large variance.

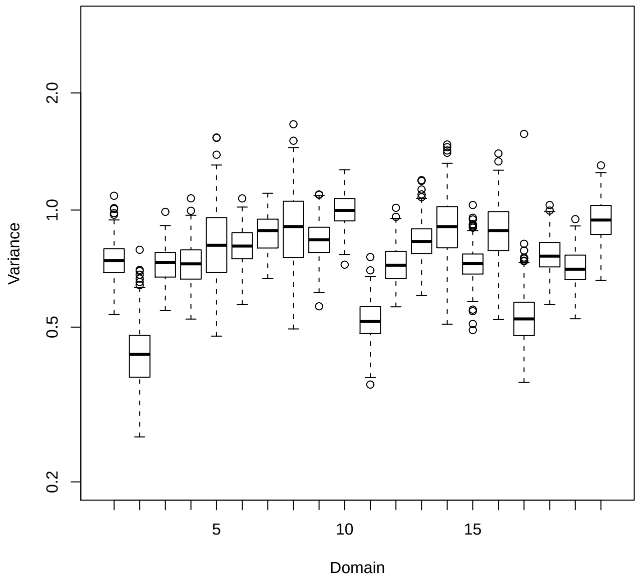

Figure 2. Estimation results variance transformed SAE by parametric bootstrapping, based on 200 bootstraps.

Also the substantial bias in domain 52413 and 52610 (number 7 and 17 in the figure, see Table 1) is not recognized. Comparison with the left plan of Figure 4 shows that the precision of the bias estimates is only slightly improved by increasing the number of bootstraps from 50 to 200. Efron and Tibshirani (1993) point out that for estimation of the bias, more bootstraps can be necessary than for estimation of the variance, which is, however, not investigated in a simulation for this situation because of the computation time. Experiments with substantially more than 200 bootstraps in a few examples lead to the preliminary conclusion that parametric bootstrapping is not the suitable method for stable bias estimation in this situation.

The right plan of Figure 5 shows the relative MSE estimates of the transformed SAE. These MSE estimates are less precise than the MSE estimates of the EBLUP (Figure 1), which is caused by the unprecise bias estimates.

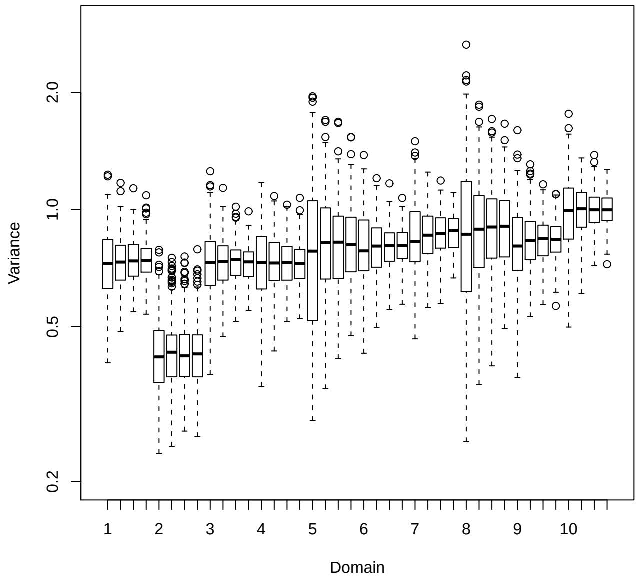

Figure 3. Estimation results variance transformed SAE by parametric bootstrapping, based on 50, 100, 150 and 200 bootstraps (from left to right), for domain 1-10

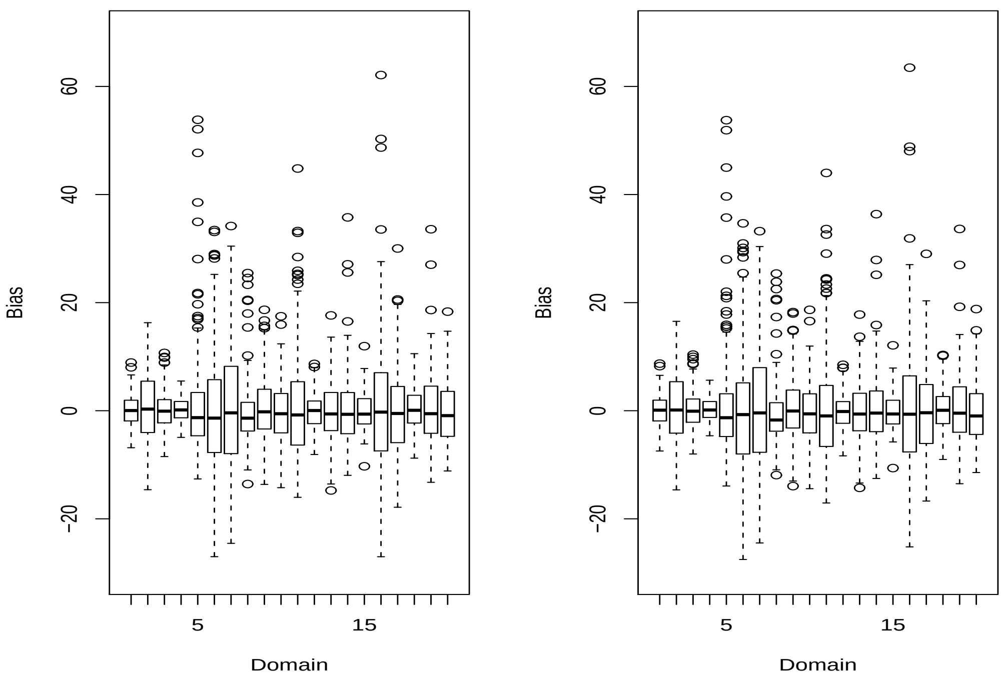

Figure 4. Estimation results bias transformed SAE by parametric bootstrapping, relative to population total, based on 50 (left plan) and 200 (right plan) bootstraps.

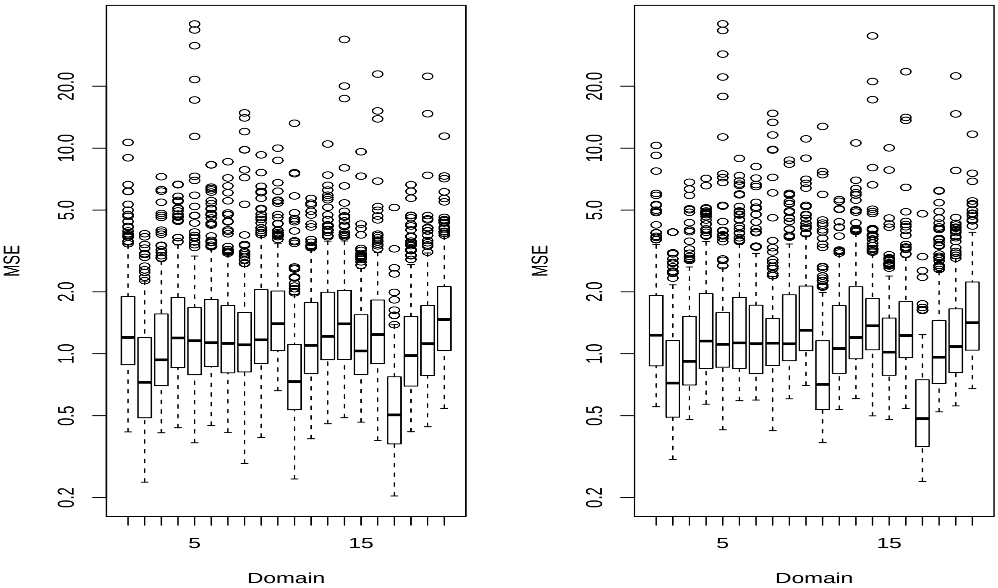

Figure 5. Estimation results MSE transformed SAE by parametric bootstrapping, based on 50 and 200 bootstraps.

5 Conclusion and discussion

By applying small area estimators, the accuracy of the estimates of the variable turnover of the Dutch Structural Business Survey can be improved substantially, compared to the generalized regression estimator. The applied small area estimators are based on a linear mixed model; this model is used to borrow information from other subpopulations to improve the accuracy of the estimates.

One of the applied methods is the EBLUP (Rao, 2003). In this application with heterogeneous data, a careful model specification with random slopes is required. Furthermore, it is necessary to take the heteroscedasticity of the residuals into account. The model estimates are biased for some subpopulations, which is probably caused by model misspecification (skewed residuals). Nevertheless, the accuracy of the estimates is substantially improved for many subpopulations.

To reduce the skewness of the residuals in this situation, the target variable is transformed. Based on this transformation, a new estimator is developed (called transformed SAE in this paper), similar to one of the estimator described in Chandra and Chambers (2011). The model estimates based on this model are substantially less biased, but also less precise, than the EBLUP estimates. All together the transformation results in a small improvement of the accuracy, compared to the EBLUP.

No analytic formula is developed yet to estimate the variance and MSE of the transformed SAE. In this paper, parametric bootstrapping is applied, which yields sufficiently stable and slightly biased estimates for the variance. On the other hand, the bias estimates are very unstable, which results in unstable MSE estimates. Though the variance estimates may be sufficient as an indication of the precision of the estimates in practical applications, more research is advisable to find more stable methods to estimate the MSE, where two approaches are possible. First, it can be investigated whether the bootstrap approach can be improved, where more complex methods can be considered. For example the double bootstrap of Hall and Maiti (2006a), Hall and Maiti (2006b) (both for parametric and non-parametric bootstrapping) is worth considering. The computation time of a double bootstrap is, however, a factor which has to be taken into account. Another possible approach is the development of an analytic formula, which would be attractive since the computation time is much smaller.

By implementing a small area estimator in the production process of the Dutch Structural Business Survey, the accuracy of the estimates can be improved substantially. However, more research is necessary before implementation can take place. First, the methods have to be investigated in the other sectors of the survey. Second, in this paper the variable tax-turnover is considered, whereas in the production process the survey variable turnover is relevant. It is likely that the results for this variable will be similar to the results described in this paper. Third, in the survey many target variables are considered, which ideally are estimated simultaneously. For some of these variables not very strong auxiliary information is available, this situation is comparable to the situation considered in this paper. For other variables very strong auxiliary information is available, as investigated in Smeets et al. (2011). There are also variables which are zero for a substantial part of the population, such a situation is not investigated yet.

Of course, the described methods are general, i.e. they can also be applied to other surveys with a continuous target variable, where a similar gain in accuracy may be possible.

Whereas the methods considered in this paper are sufficient to improve the accuracy compared to the generalized regression estimator, it is attractive to also investigate other small area estimators, which can possibly further improve the accuracy or which require less complex model specifications. A cooperation project with the University of Southampton has started where M-quantile estimators of Chambers and Tzavidis (2006) and robust small area estimators of Sinha and Rao (2009) and Gershunskaya and Lahiri (2011) are under consideration.

6 Acknowledgements

The authors wish to thank Harm Jan Boonstra (Statistics Netherlands), Geert Molenberghs (University of Hasselt and Leuven), Paul Smith (Office of National Statistics), and Nikos Tzavidis (University of Southampton) for their useful comments about SAE and mixed models and Martijn Tennekes (Statistics Netherlands) for his help using R. Furthermore the authors thank Harm Jan Boonstra, Jan van den Brakel and Frank van de Pol for sharing some computer capacity and making the simulation of this paper possible.

References

Boonstra, H., van den Brakel, J., Buelens, B., Krieg, S., and Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *Metron* LXVI (1), 21–49.

Bradu, D. and Mundak, Y. (1970). Estimation in Lognormal Linear Models. *Journal of the American Statistical Association* 65 - 329, 198–211.

Buelens, B. and Benschop, T. (2009). *Small area estimation of violent crime victim rates in the Netherlands.* Technical Report DMH-2009-02-03-BBUS, Statistics Netherlands.

Butar, F. and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small area estimators. *Journal of Statistical Planning and Inference* 112, 63–76.

Chambers, R., Chandra, H., Salvati, N., and Tzavidis, N. (2011). Outlier robust small area estimation. *Journal of Royal Statistical Society B* Invited to review and resubmit.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika* 93, 255–268.

Chandra, H. and Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology* 37, 39–51.

Davidson, A. and Hinkley, D. (1997). *Bootstrap Methods and their applications.* Cambridge University Press.

Duan, N. (1983). Smearing estimate: A nonparametric Retransformation Method. *Journal of the American Statistical Association* 78 - 383, 605–610.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

ESSnet (2011). *Report on Work Package 2 - State of the art.* Technical Report in preparation, ESSnet on Small Area Estimation, EISTAT, DESTATIS, GUS, INE, INSEE, ONS, SSB, SFSO, EUROSTAT.

Gershunskaya, J. and Lahiri, P. (2011). Robust Small Area Estimation. VII Conference Survey Sampling in Economic and Social Research, Katowice, 18-20 September 2011.

Gross, S. (1980). Median estimation in sample surveys. American Statistical Assosiation, Procedings Joint Statistical Meetings, Section on Survey Research Methods 181-184.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics* 34, 1733–1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society* 68, 221–238.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small area estimation. *Statistical Science* 18, 199–210.

McCarthy, P. and Snowden, C. (1985). *The bootstrap and finite population sampling.* Technical Report Public Health Service Publication 85-1369, U.S.Government printing office, Washington, In Vital and Health Statistics 2-95.

Moura, F. and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology* 25 (1), 73–80.

Neyman, J. and Scott, E. (1960). Correction for bias introduced by a transformation of variables. *The Annals of Mathematical Statistics* 31 - 3, 643–655.

Pfeffermann, D. (2010). *New important developments in small area estimation.* Technical Report Working Paper M10/11, University of Southampton, Methodology.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association* 85, 163–171.

R Development Core Team (2009). *R: A language and environment for statistical computing.* Technical Report http://www.R-project.org, R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics* 22, 511–528.

Rao, J. N. K. (2003). *Small Area Estimation.* John Wiley, New York.

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

Sinha, S. K. and Rao, J. N. K. (2009). Robust Methods for Small Area Estimation. *Canadian Journal of Statistics* 37 (3), 381–399.

Smeets, M., Blaess, V., and Krieg, S. (2011). *Small area estimation for business surveys with strong auxiliary information.* Technical Report DMH-2012-1-25-mset, Statistics Netherlands.

van den Brakel, J. and Krieg, S. (2009). Estimation of the Monthly Unemployment Rate through Structural Time Series Modelling in Rotating Panel Design. *Survey Methodology* 16, 177–190.

Appendix A Description of the size classes and the industries of the retail trade

The size classes are given by: SC 1 (1 WP), SC 2 (2-4 WP), SC 3 (5-9 WP), SC 4 (10-19 WP), SC 5 (20-49 WP), SC 6 (50-99 WP), SC 7 (100-199 WP), SC 8 (200-499 WP), SC 9 (500 WP and more).

Short descriptions of the industries are provided below. Between brackets first the SIC-code and then information about the sample design: total sample size per industry, total population size per industry, then sample size per industry per SC with between brackets population size per industry per SC).

- Food, non-specialized (52110, 230, 1898, 11(385), 65(680), 35(275), 43(201), 76(357)),
- Non-food, non-specialized (52120, 41, 110, 15(38), 15(37), 3(7), 5(18), 3(10)),
- Food, specialized (52200, 620, 7684, 211(1913), 259(4169), 83(1312), 50(249), 17(41)),
- Dispensing chemists (52310, 187, 498, 17(38), 15(48), 30(139), 65(209), 60(64)),
- Chemists, medical and orthopaedic goods (52321, 50, 1029, 5(226), 16(513), 16(217), 7(64), 6(9)),
- Perfume, cosmetics (52330, 21, 199, 5(99), 5(69), 4(20), 5(9), 2(2)),
- Textiles (52413, 35, 562, 7(331), 16(201), 9(27), 1(1), 2(2)),
- Clothing (52422, 202, 6612, 23(2314), 77(3404), 23(657), 49(189), 30(48)),
- Footwear, leather (52431, 65, 1247, 6(270), 29(728), 14(188), 7(50), 9(11)),
- Furniture, lighting, household articles (52440, 129, 4017, 13(1311), 45(1950), 21(605), 32(122), 18(29)),
- Electric (52450, 147, 2391, 22(1021), 69(1028), 31(272), 16(58), 9(12)),
- Hardware, paints and glass (52460, 343, 2887, 49(886), 118(1297), 68(491), 70(169), 38(44)),
- Books (52470, 83, 1331, 13(349), 37(692), 22(232), 7(47) 4(11)),
- Photo, jewelery, art, bicycles, sport, camping (52485, 204, 6671, 32(2518), 79(3259), 34(746), 41(121), 18(27)),
- Garden, animals, computers, toys, baby articles, carpeting (52491, 769, 8407, 65(3566), 511(3744), 84(841), 68(196), 41(60)),
- Second-hand stores (52500, 114, 2545, 52(1730), 44(727), 6(63), 5(18), 7(7)),
- Mail order houses (52610, 392, 1904, 288(1441), 86(431), 9(21), 5(7) 4(4)),
- Stalls and market (52620, 598, 6914, 358(4470), 175(2180), 47(234), 17(29), 1(1)),

- Other non-store (52630, 639, 3643, 511(2556), 110(971), 10(95), 5(18), 3(3)),
- Repair (52700, 205, 3409, 74(2378), 91(880), 26(130), 10(17), 4(4)).

Appendix B Variance estimation of the GREG and MSE estimation of the EBLUP

To compute the variance of the GREG (Särndal et al., 1992), first the GREG estimator is expressed with weights w_{ij}:

$$\widehat{Y}_j^{\mathrm{GREG}} = \sum_{h=1}^{H_j}\sum_{i=1}^{n_{hj}} w_{ij} y_{ij},$$

with

$$w_{ij} = \frac{N_{hj}}{n_{hj}} \left(1 + \mathbf{x}_{ij}^{\mathrm{t}} \left(\sum_{h=1}^{H} \frac{N_{hj}}{n_{hj}} \sum_{i=1}^{n_{hj}} \mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{t}} \right)^{-1} (\mathbf{X}_j^{\mathrm{tot}} - \widehat{\mathbf{X}}_j^{\mathrm{HT}})^t \right).$$

Then the variance of the GREG estimator is estimated by the following formula:

$$\mathrm{var}(\widehat{Y}_j^{\mathrm{GREG}}) = \sum_{h=1}^{H} \frac{n_{hj}(1-\frac{n_{hj}}{N_{hj}})}{n_{hj}-1} \left(\sum_{i=1}^{n_{hj}} (w_{ij}e_{ij})^2 - \frac{1}{n_{hj}} \left(\sum_{i=1}^{n_{hj}} w_{ij}e_{ij} \right)^2 \right), \quad \text{(B.1)}$$

with $e_{ij} = y_{ij} - \mathbf{x}_{ij}^{\mathrm{t}}\widehat{\boldsymbol{\beta}}_j$. Since the GREG estimator is approximately design unbiased, formula (B.1) can be considered as a formula for the MSE.

For the EBLUP, an analytic formula for the MSE is developed by Prasad and Rao (1990), see also Rao (2003), which is worked out in Moura and Holt (1999) for a two-level model where the hyperparameters are estimated by REML. The MSE of the EBLUP for domain j based on formula (3) where $\boldsymbol{\Theta}$ and σ_e^2 are estimated by REML can be estimated by

$$\mathrm{mse}_j = \widehat{g}_1 + \widehat{g}_2 + 2\widehat{g}_3, \quad \text{(B.2)}$$

with

$$\begin{aligned} \widehat{g}_1 &= \mathbf{X}_{j1}^{\mathrm{R}}(\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{V}}_j^{-1}\mathbf{X}_{j1}\widehat{\boldsymbol{\Theta}})(\mathbf{X}_{j1}^{\mathrm{R}})^{\mathrm{t}}, \\ \widehat{g}_2 &= (\mathbf{X}_j^{\mathrm{R}} - \mathbf{X}_{j1}^{\mathrm{R}}\widehat{\boldsymbol{\Theta}}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{V}}_j^{-1}\mathbf{X}_j) \\ &\quad \times (\sum_{k=1}^{m} \mathbf{X}_k^{\mathrm{t}}\widehat{\mathbf{V}}_k^{-1}\mathbf{X}_k)^{-1}(\mathbf{X}_j^{\mathrm{R}} - \mathbf{X}_{j1}^{\mathrm{R}}\widehat{\boldsymbol{\Theta}}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{V}}_j^{-1}\mathbf{X}_j)^{\mathrm{t}} \end{aligned}$$

with $\widehat{\mathbf{V}}_j = \widehat{\mathbf{R}}_j + \mathbf{X}_{j1}\widehat{\boldsymbol{\Theta}}\mathbf{X}_{j1}^{\mathrm{t}}$, $\mathbf{X}_j^{\mathrm{R}}$ and $\mathbf{X}_{j1}^{\mathrm{R}}$ the totals of the non-sampled part in industry j of the covariates $\mathbf{x}_{ij}^{\mathrm{t}}$ and $\mathbf{x}_{ij1}^{\mathrm{t}}$. Furthermore, $\mathbf{X}_j = (\mathbf{x}_{1j}, \ldots, \mathbf{x}_{n_j j})^{\mathrm{t}}$ and $\mathbf{X}_{j1} = (\mathbf{x}_{1j1}, \ldots, \mathbf{x}_{n_j j1})^{\mathrm{t}}$.

Furthermore

$$\widehat{g}_3 = \mathbf{X}_{j1}^{\mathrm{R}}(\widehat{\mathbf{G}}_j^{-1})^{\mathrm{t}} \left[\sum_{k=1}^{s-1}\sum_{l=1}^{s-1} b_{kl}\boldsymbol{\Delta}_k\widehat{\mathbf{C}}_j\boldsymbol{\Delta}_l^{\mathrm{t}} \right] \widehat{\mathbf{G}}_j^{-1}(\mathbf{X}_{j1}^{\mathrm{R}})^{\mathrm{t}}$$

$$-2\mathbf{X}_{j1}^{\mathrm{R}}(\widehat{\mathbf{G}}_j^{-1})^{\mathrm{t}}\left[\sum_{k=1}^{s-1} b_{ks}\boldsymbol{\Delta}_k\right]\widehat{\mathbf{T}}_j\widehat{\boldsymbol{\Theta}}(\mathbf{X}_{j1}^{\mathrm{R}})^{\mathrm{t}}$$
$$+b_{ss}\mathbf{X}_{j1}^{\mathrm{R}}\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{S}}_j\widehat{\boldsymbol{\Theta}}(\mathbf{X}_{j1}^{\mathrm{R}})^{\mathrm{t}},$$

where

$$\begin{aligned}
\widehat{\mathbf{G}}_j &= \mathbf{I}_q + \mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{R}}_j^{-1}\mathbf{X}_{j1}\widehat{\boldsymbol{\Theta}},\\
\widehat{\mathbf{C}}_j &= \widehat{\mathbf{G}}_j^{-1}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{R}}_j^{-1}\mathbf{X}_{j1},\\
\widehat{\mathbf{T}}_j &= \widehat{\mathbf{G}}_j^{-2}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{R}}_j^{-2}\mathbf{X}_{j1},\\
\widehat{\mathbf{S}}_j &= \widehat{\mathbf{G}}_j^{-3}\mathbf{X}_{j1}^{\mathrm{t}}\widehat{\mathbf{R}}_j^{-3}\mathbf{X}_{j1},\\
\boldsymbol{\Delta}_k &= \frac{\partial\boldsymbol{\Theta}}{\partial\delta_k}, \text{ with } \delta_1,\ldots,\delta_{s-1} \text{ the } s-1=\tfrac{q(q+1)}{2} \text{ distinct elements of } \boldsymbol{\Theta}.
\end{aligned}$$

Furthermore, b_{kl} with $k,l = 1,\ldots,s$ are the elements of the $s\times s$-matrix $\widehat{\mathbf{B}}$. This matrix is an approximation of $\mathrm{E}[(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta})(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta})^{\mathrm{t}}]$, with $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_s)$ and $\delta_s=\sigma_e^2$. The matrix $\mathbf{B}$ is defined by its inverse $\mathbf{B}^{-1}$ in the following way:

$$\mathbf{B}_{kl}^{-1} = \sum_{j=1}^{m}\tau\left(\mathbf{P}_j\frac{\partial\mathbf{V}_j}{\partial\delta_k}\mathbf{P}_j\frac{\partial\mathbf{V}_j}{\partial\delta_l}\right), \text{ for } k,l=1,\ldots,s,$$

where

$$\mathbf{P}_j = \mathbf{V}_j^{-1} - \mathbf{V}_j^{-1}\mathbf{X}_j(\sum_{k=1}^{m}\mathbf{X}_k^{\mathrm{t}}\mathbf{V}_k^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_j^{\mathrm{t}}\mathbf{V}_j^{-1}.$$

Appendix C Simulation results for models with random intercept and the homoscedasticity assumption

A simulation study is carried out with 500 runs to investigate whether more parsimonious models (with random intercept or based on the homoscedasticity assumption) are sufficient to improve the accuracy of the GREG estimates for the variable tax-turnover of the retail-trade of the SBS. In this simulation the following estimators are applied:

1. The EBLUP with $1+\mathrm{SC}+\mathrm{WP}$ used as fixed effects, 1 used as random effects and $k_{ij}=1$.

2. The EBLUP with $1+\mathrm{SC}+\mathrm{WP}$ used as fixed effects, 1 used as random effects and k_{ij} computed as standard errors of each stratum of the residuals of the regression model with $1+SC+WP$ as auxiliary information.

3. The EBLUP with $1+\mathrm{SC}+\mathrm{WP}$ used as fixed effects and as random effects and $k_{ij}=1$.

4. The transformed SAE with $1+\mathrm{SC}+\sqrt[8]{\mathrm{WP}}$ used as fixed effects, 1 used as random effects and $k_{ij,\tau}=1$.

5. The transformed SAE with $1 + \mathrm{SC} + \sqrt[8]{\mathrm{WP}}$ used as fixed effects, 1 used as random effects and $k_{ij,\tau}$ computed as standard errors of each stratum of the residuals of the regression model with $1 + \mathrm{SC} + \sqrt[8]{\mathrm{WP}}$ as auxiliary information.

6. The transformed SAE with $1 + \mathrm{SC} + \sqrt[8]{\mathrm{WP}}$ used as fixed effects and as random effects and $k_{ij,\tau} = 1$.

7. The GREG estimator with $1 + \mathrm{SC} + \mathrm{WP}$ as auxiliary information.

Table 6 shows that the model estimates based on these models are strongly biased for some domains. In Table 7 the relative root mean squared errors are shown. The accuracy is not improved by the models, compared to the GREG estimator.

Method	EBLUP	EBLUP	EBLUP	tr. SAE	tr. SAE	tr. SAE	GREG
k_{ij}	1	stdev	1	1	stdev	1	
random ef.	interc.	interc.	slope	interc.	interc.	slope	
52110	-7.22	6.60	0.89	3.57	3.14	1.99	0.11
52120	-6.54	-3.95	2.55	-0.46	1.06	-1.30	-0.26
52200	5.61	-5.27	4.51	-15.73	-11.77	-3.84	-0.67
52310	0.62	7.80	3.99	5.23	-0.03	2.52	-0.05
52321	12.78	5.82	6.21	-4.19	-0.84	-3.69	-0.32
52330	61.44	21.02	6.74	6.00	47.28	-4.39	-0.53
52413	24.97	4.07	-29.76	-25.68	-4.68	-25.42	-0.07
52422	31.43	6.56	1.60	-5.57	-2.90	-5.55	-0.65
52431	11.82	3.17	1.44	-5.67	-1.65	-6.33	-0.17
52440	0.34	1.20	0.28	-5.26	-5.35	-2.78	0.23
52450	-36.79	12.66	-2.06	7.98	3.34	7.83	-1.27
52460	-5.31	-0.01	6.00	-1.11	-2.90	0.89	-0.31
52470	-2.15	3.39	0.66	-4.99	-1.83	-3.75	0.05
52485	-18.29	1.16	-7.91	-7.93	-3.72	-3.88	0.67
52491	20.96	2.27	-2.43	-6.08	-2.74	-4.52	-0.31
52500	198.72	16.50	9.70	-0.05	4.12	-9.16	-0.58
52610	-83.40	7.14	-2.25	20.06	9.21	20.69	-0.68
52620	18.23	-3.68	-10.68	-17.23	-2.78	-15.10	-0.13
52630	2.64	7.45	-18.88	6.37	6.00	1.44	-0.24
52700	80.27	6.58	-5.19	-16.59	-1.55	-17.95	0.10

Table 6. Relative bias (in %) of 3 EBLUPs, 3 transformed SAEs and the regression estimator, based on a simulation with 500 runs.

Method	EBLUP	EBLUP	EBLUP	tr. SAE	tr. SAE	tr. SAE	GREG
k_{ij}	1	stdev	1	1	stdev	1	
random ef.	interc.	interc.	slope	interc.	interc.	slope	
52110	8.41	7.67	3.19	4.51	4.51	3.63	3.28
52120	10.67	8.37	8.77	7.42	9.41	8.08	11.09
52200	9.87	6.88	8.07	16.08	12.53	5.00	4.78
52310	2.17	8.33	4.79	5.65	2.09	3.73	2.60
52321	15.61	8.25	8.01	7.28	8.69	6.82	13.43
52330	64.66	23.91	12.44	11.08	47.72	11.14	14.22
52413	30.40	12.82	30.54	27.40	12.29	27.04	13.10
52422	34.77	7.83	10.21	6.97	7.86	7.01	12.06
52431	16.62	6.64	8.25	8.24	6.96	8.73	7.24
52440	12.89	6.64	6.29	7.39	8.15	5.77	6.57
52450	59.74	14.92	20.62	11.54	12.23	11.57	16.76
52460	8.06	4.78	7.18	3.22	5.01	3.20	4.51
52470	13.05	6.50	10.25	7.84	6.26	6.97	6.59
52485	24.94	5.86	11.32	9.32	7.47	5.99	8.85
52491	22.77	4.53	5.65	6.77	4.77	5.45	5.58
52500	200.29	21.37	21.37	8.05	13.09	12.75	24.97
52610	85.43	18.30	19.96	20.89	13.27	21.84	16.85
52620	19.18	5.29	13.12	17.67	4.81	15.63	4.44
52630	7.75	8.54	20.19	7.42	8.23	5.75	9.18
52700	81.35	9.25	12.16	17.49	6.11	18.78	6.45
Mean	36.43	9.83	12.12	10.61	10.07	9.74	9.63

Table 7. Relative root mean squared error (in %) of 3 EBLUPs, 3 transformed SAEs and the regression estimator, based on a simulation with 500 runs.

Appendix D More bootstrap results

To investigate the effects of serious model misspecification on bootstrapping, both parametric and non-parametric bootstrapping is applied to the EBLUP. Furthermore, non-parametric bootstrap estimates of the MSE of the GREG are compared with variance estimates for the GREG based on the analytic formula (B.1). EBLUP and GREG are used to estimate tax-turnover for the 20 industries of the retail trade of the SBS.

Here, the basic concept of non-parametric bootstrapping is applied: a population of the size of the real population is created from the sample by simple random sampling with replacement in each stratum, taking care that each element i is at most $[w_i]$ times in the sample with w_i the inclusion weight and $[x]$ the smallest integer larger than or equal to x. This method, proposed by Gross (1980) and called without-replacement bootstrap by McCarthy and Snowden (1985) and Lahiri (2003), takes the finite population into account. It is assumed that this method is sufficient in the application of the SBS where the inclusion probability is small in most strata.

In the figures for the EBLUP, relative variances, MSE and bias are shown, which are computed in the same way as in Section 4.4. Figure 6 shows the results for the GREG, where the variance estimates based on formula (B.1) are compared with the MSE estimates based on non-parametric bootstrapping. Both methods yield unstable estimates for variance and MSE respectively due to small sample sizes.

In Figure 7 the variance estimates for the EBLUP by non-parametric bootstrapping are shown. For most of the runs and most of the domains, the variance is estimated quite well. For two domains, however (number 11 and 17), the variance estimate is unstable. Furthermore, for all domains there is a substantial part of the runs where the variance estimate is substantially too small or too large.

Figure 8 shows the results for the bias estimates of the EBLUP by non-parametric bootstrapping. The substantial bias in domain 52413 (number 7 in the figure) which is found in Table 1, is recognized well, leaving aside a large variance of this estimate. However, for a considerable part of the runs a substantial bias is found for the other domains by bootstrapping which is not recognized in the simulation (Table 1).

In Figure 9 the MSE estimates for the EBLUP by non-parametric bootstrapping are shown. As for the variances, for all domains there is a substantial part of the runs where the variance estimate is substantially too small or too large.

In Figure 10 the variance estimates for the EBLUP by parametric bootstrapping

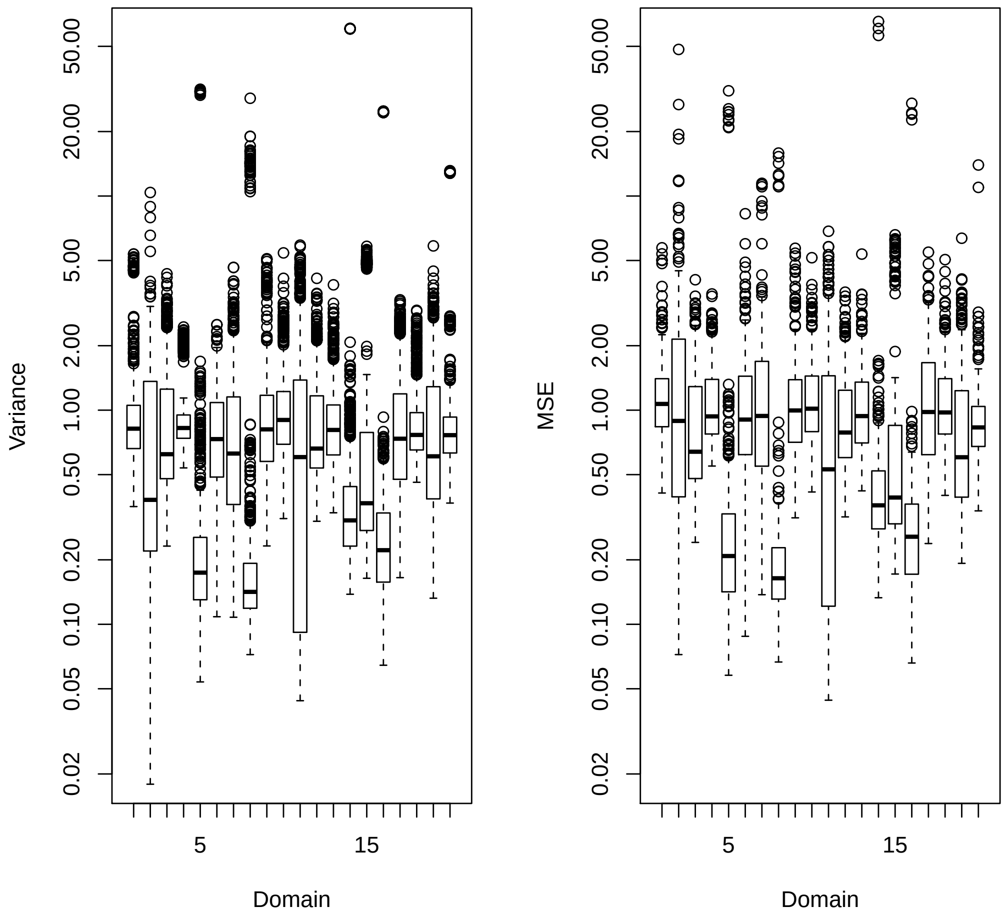

Figure 6. Estimation results variance GREG by formula (left plan) and MSE GREG by non-parametric bootstrapping (right plan).

are given. The stability of the estimates is comparable to the results based on non-parametric bootstrapping. Whereas by non-parametric bootstrapping the estimates are around the true variance (based on the simulation), the variance is substantially overestimated by parametric bootstrapping. It is likely that this is caused by model misspecification.

Figure 11 shows the bias estimates for the EBLUP by parametric bootstrapping. As by non-parametric bootstrapping, the bias in domain 52413 (number 7 in the figure, see Table 1), is recognized well, again with a large variance. Furthermore, by parametric bootstrapping a substantial bias is estimated for most of the runs in domain 52500, 52620 and 52700 (number 16, 18 and 20 in the figure), which is not found in the simulation (Table 1).

From Figure 12, where the MSE estimates for the EBLUP by parametric bootstrapping are given, it can be concluded that the MSE is overestimated by

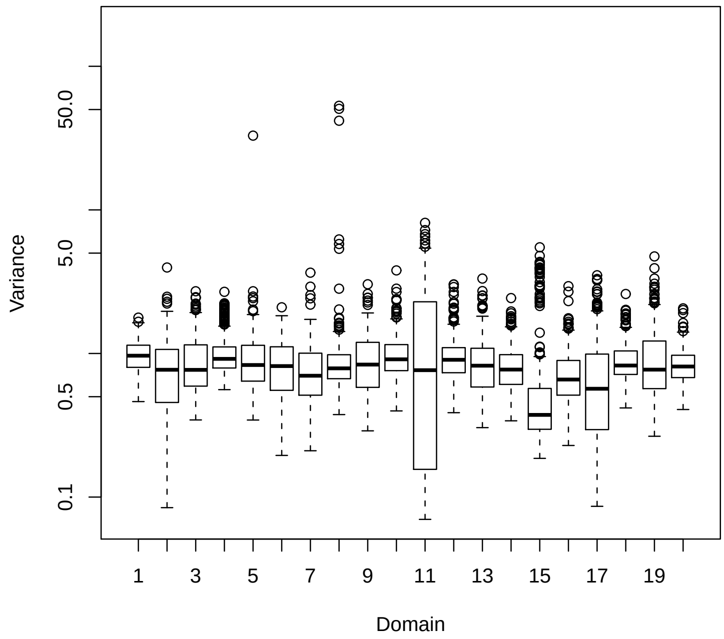

Figure 7. Estimation results variance EBLUP by non-parametric bootstrapping.

parametric bootstrapping, and that the estimates are unstable. This is a consequence of the unstable and overestimated variance estimates, see Figure 10.

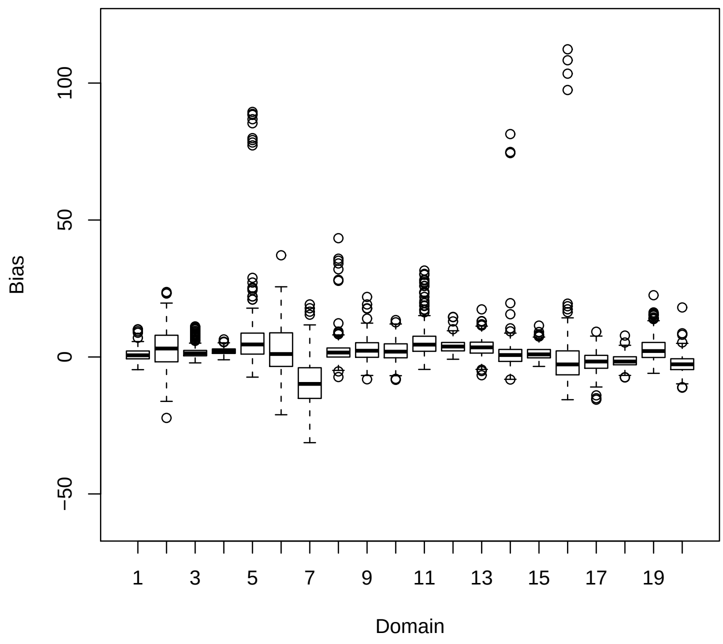

Figure 8. Estimation results bias EBLUP by non-parametric bootstrapping.

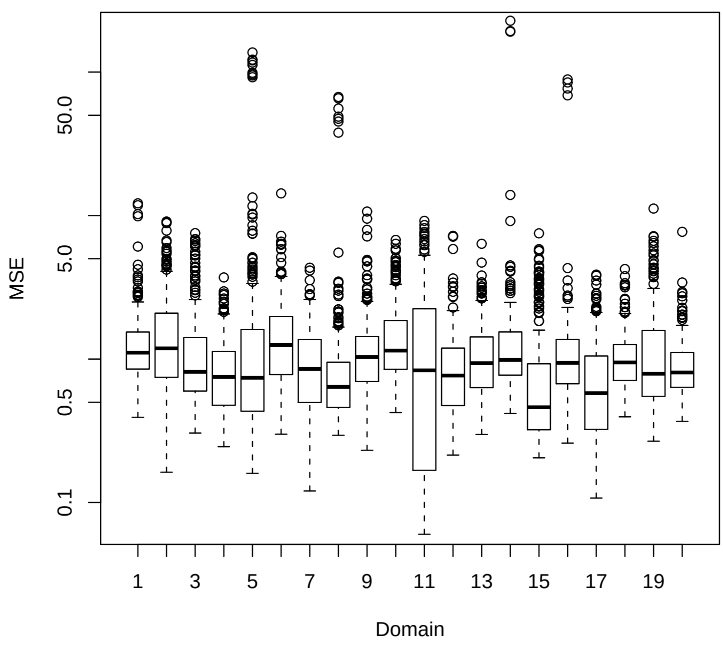

Figure 9. Estimation results MSE EBLUP by non-parametric bootstrapping.

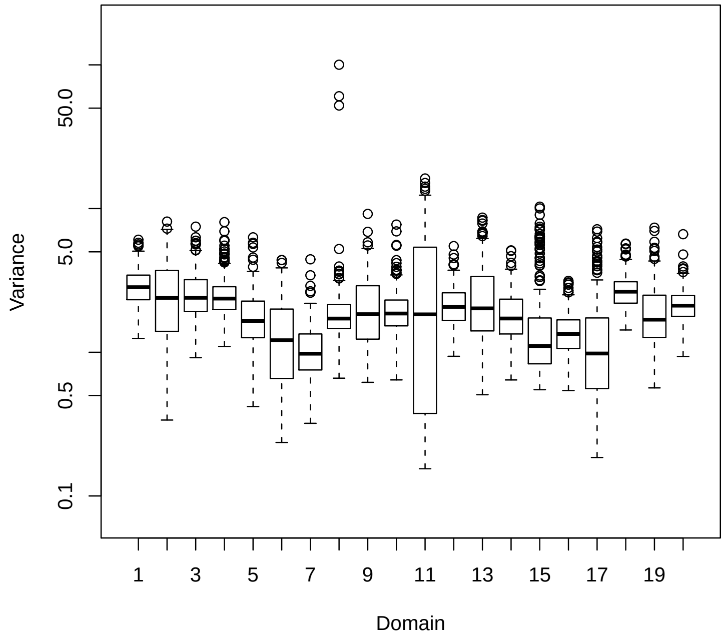

Figure 10. Estimation results variance EBLUP by parametric bootstrapping.

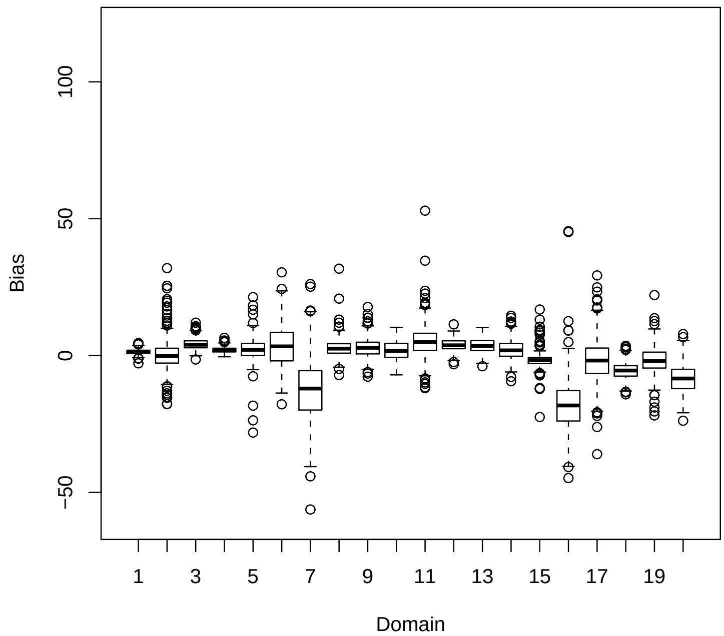

Figure 11. Estimation results bias EBLUP by parametric bootstrapping.

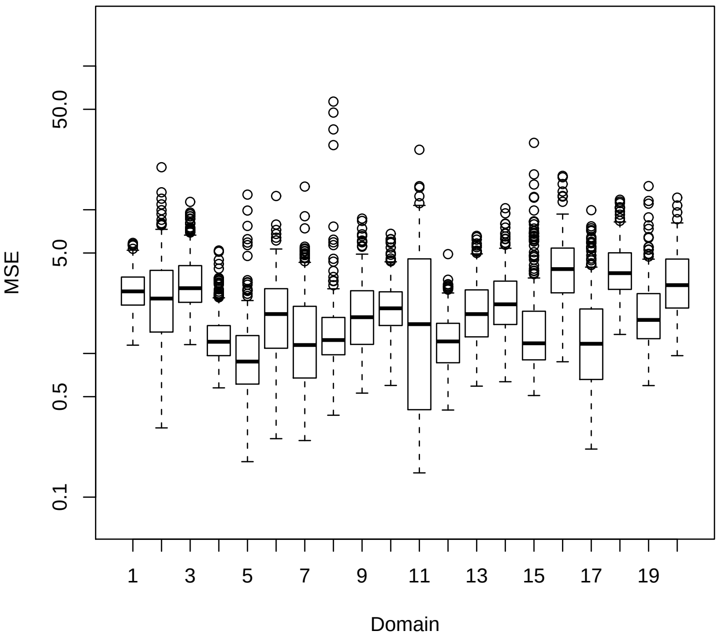

Figure 12. Estimation results MSE EBLUP by parametric bootstrapping.