DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 963–986, http://dx.doi.org/10.1515/JOS-2016-0051

Small-Area Estimation with Zero-Inflated Data – a Simulation Study

Sabine Krieg[1], *Harm Jan Boonstra*[1], *and Marc Smeets*[1]

Many target variables in official statistics follow a semicontinuous distribution with a mixture of zeros and continuously distributed positive values. Such variables are called zero inflated. When reliable estimates for subpopulations with small sample sizes are required, model-based small-area estimators can be used, which improve the accuracy of the estimates by borrowing information from other subpopulations. In this article, three small-area estimators are investigated. The first estimator is the EBLUP, which can be considered the most common small-area estimator and is based on a linear mixed model that assumes normal distributions. Therefore, the EBLUP is model misspecified in the case of zero-inflated variables. The other two small-area estimators are based on a model that takes zero inflation explicitly into account. Both the Bayesian and the frequentist approach are considered. These small-area estimators are compared with each other and with design-based estimation in a simulation study with zero-inflated target variables. Both a simulation with artificial data and a simulation with real data from the Dutch Household Budget Survey are carried out. It is found that the small-area estimators improve the accuracy compared to the design-based estimator. The amount of improvement strongly depends on the properties of the population and the subpopulations of interest.

Key words: Generalized linear mixed model; EBLUP; MCMC; Logit; Dutch Household Budget Survey.

1. Introduction

Traditionally, national statistical institutes (NSIs) such as Statistics Netherlands prefer design-based estimation methods, since these methods lead to approximately design-unbiased estimates. However, the demand for detailed estimates for subpopulations is increasing, while at the same time budgets are under continuous pressure. Therefore, several NSIs started to investigate the possibilities of small-area estimation (SAE), see, for example Eurarea (2004) and Boonstra et al. (2008). This model-based methodology is developed for situations where the sample sizes of the subpopulations (often called domains or areas in the SAE context) or time periods are too small to compute reliable estimates based on design-based methods. An SAE method borrows information from other domains or from other time periods to improve the accuracy of the domain estimates.

The most common SAE estimator is the Empirical Best Linear Unbiased Predictor (EBLUP) (Battese et al. 1988; Rao 2003). The EBLUP is based on a linear mixed model and assumes normal distributions. However, NSIs often have to deal with non-normally distributed data, for which the EBLUP may yield seriously biased estimates. For such

[1] Statistics Netherlands, Postbus 4481, 6401CZ Heerlen, The Netherlands. Emails: skrg@cbs.nl, hbta@cbs.nl, and mset@cbs.nl

situations, different adjustments of the EBLUP and some new SAE methods have been developed in recent years. For example, the robust EBLUP (Sinha and Rao 2009) reduces the influence of outliers in the data. Chandra and Chambers (2011b) developed an estimator for skewly distributed data, and the M-quantile estimator (Chambers and Tzavidis 2006) does not make any assumptions about the distribution.

This article deals with the estimation for variables that are zero for a substantial part of the population. This type of data is also called zero-inflated data. Pfeffermann et al. (2008) and Chandra and Sud (2012) developed an estimator for such kinds of data, the first using a Bayesian approach and the second a frequentist approach. The estimator is based on two models, the first being a linear mixed model for the nonzero values and the second a generalized linear mixed model for the binary zero indicator. Both the Bayesian and the frequentist approaches are used in this article, with a small simplification of the method used in Pfeffermann et al. (2008). The SAE method for zero-inflated data is compared with the EBLUP and with a design-based method (the survey regression estimator). In the first part of the article, a simulation with artificial data is carried out in which different populations are created to investigate the properties of the considered estimators in different situations. This simulation shows to what extent the model misspecification of the EBLUP increases the bias of the estimates and to what extent the accuracy of the estimates is improved when the estimators of Pfeffermann et al. (2008) and Chandra and Sud (2012) are applied instead. In a second simulation, the estimators are applied to real zero-inflated data of the Dutch Household Budget Survey (HBS). The HBS measures the consumption expenditures of Dutch households. Many target variables which describe the expenditures for different products, are zero inflated.

In Section 2 the considered methods are described. Then the results of the simulation with artificial populations are discussed in Section 3. The results of the simulation for the HBS follow in Section 4. In Section 5 the conclusions are given.

2. Methods

2.1. Notation

The finite population U with N elements is divided into m subpopulations or domains. A sample with n elements is drawn using simple random sampling without replacement. The observed value of the target variable for unit i in domain j is given by y_{ij}. The total sample and population size in domain j are denoted by n_j and N_j, respectively. The total sample is called S and the sample in domain j is called S_j.

The explanatory variables for unit i in domain j are given by the vector $\mathbf{x}_{ij} = \left(x_{ij}^1, \ldots, x_{ij}^p\right)^t$. An intercept is always included, that is, it can be assumed that $x_{ij}^1 = 1$. Population means $Y_j^{\text{mean}} = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}$ for target variable y for all domains $j = 1, \ldots, m$ have to be estimated.

The target variable y_{ij} is equal to zero for a substantial part of the population. We define

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad y_{ij} \neq 0 \\ 0 & \text{if} \quad y_{ij} = 0. \end{cases} \tag{1}$$

The subscript nz is used to denote the nonzero part of the population or sample.

2.2. Survey Regression

Survey regression (SR) is a design-based model-assisted estimator which is approximately design unbiased (Woodruff 1966; Battese et al. 1988; Särndal et al. 1992). In this article the SR is considered to be the reference estimator; the model-based methods are expected to be more accurate than the SR. The SR of the unknown population mean Y_j^{mean} for domain j is given by

$$\hat{Y}_j^{\text{SR}} = \hat{Y}_j^{\text{HT}} + \left(\mathbf{X}_j^{\text{mean}} - \hat{\mathbf{X}}_j^{\text{HT}}\right)^{\text{t}} \hat{\beta}, \tag{2}$$

where

$$\hat{\beta} = (\mathbf{X}^{\text{t}}\mathbf{X})^{-1}\mathbf{X}^{\text{t}}\mathbf{y}.$$

Here the Horvitz-Thompson estimators are given by $\hat{Y}_j^{\text{HT}} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ and $\hat{\mathbf{X}}_j^{\text{HT}} = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{x}_{ij}$. Furthermore, $\mathbf{X}_j^{\text{mean}} = \frac{1}{N_j}\sum_{i=1}^{N_j} \mathbf{x}_{ij}$ is the p-vector of population means of the auxiliary information in domain j, $\mathbf{y} = (y_{11}, \ldots, y_{n_1 1}, y_{12} \ldots, y_{n_m m})^{\text{t}}$ and $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{n_1 1}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{n_m m})^{\text{t}}$.

2.3. Empirical Best Linear Unbiased Predictor (EBLUP)

Consider the linear mixed model given by

$$y_{ij} = \mathbf{x}_{ij}^{\text{t}}\beta + \vartheta_j + e_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{3}$$

where

$$\vartheta_j \sim \mathcal{N}\left(0, \sigma_r^2\right), \;\; e_{ij} \sim \mathcal{N}\left(0, \sigma_e^2\right).$$

Here σ_e^2 is the within-area variance parameter, whereas σ_r^2 is the between-domain variance.

Based on Model (3), the EBLUP (Rao 2003) is considered to estimate the population means Y_j^{mean} for the domains $j = 1, \ldots, m$. The estimator for Y_j^{mean} is then given by

$$\hat{Y}_j^{\text{EBLUP}} = \mathbf{X}_j^{\text{mean}}\hat{\beta} + \hat{\vartheta}_j. \tag{4}$$

Expressions for $\hat{\beta}$ and $\hat{\vartheta}_j$ can be found in Rao 2003, sec. 7.2. The variance parameters σ_r^2 and σ_e^2 are estimated by the method of Restricted Maximum Likelihood (REML).

A refined version of (4) would use predicted values only for the nonsampled part of the population, and the observed values for themselves. However, when sampling fractions are small, the difference is negligible and for that reason (4) is used in this article. The EBLUP estimator is computed with R (R Development Core Team 2009), where the function `lmer` of package `lme4` (Bates et al. 2015) is used to fit the linear mixed model.

2.4. A Small-Area Estimator for Zero-Inflated Data

In this section, an estimator is described that takes the zero inflation into account. There are two approaches to estimate the models: the frequentist approach (Subsection 2.4.1), described by Chandra and Sud (2012), and the Bayesian approach (Subsection 2.4.2), described by Pfeffermann et al. (2008). For both approaches we use the abbreviation

ZERO in the rest of the article, or ZERO-F or ZERO-B to make clear which approach is used. The theoretical properties of the estimators are discussed in Pfeffermann et al. (2008) and Chandra and Sud (2012).

Note that an important disadvantage of ZERO compared with the EBLUP is that ZERO can only be applied if the auxiliary information is known for all elements in the population.

2.4.1. The Frequentist Approach

The target variable y_{ij} is assumed to be the product of an underlying normally distributed variable y_{ij}^* and δ_{ij}, that is $y_{ij} = y_{ij}^* \delta_{ij}$. These two variables are modelled in two different (generalized) linear mixed models. The first model describes the distribution of y_{ij}^*:

$$y_{ij}^* = \mathbf{x}_{\text{nz},ij}^{\text{t}} \beta_{\text{nz}} + \vartheta_{\text{nz},j} + e_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{5}$$

where

$$\vartheta_{\text{nz},j} \sim \mathcal{N}\left(0, \sigma_{r,\text{nz}}^2\right), \quad e_{ij} \sim \mathcal{N}\left(0, \sigma_{e,\text{nz}}^2\right).$$

The second model describes the probabilities $p_{ij} = P(\delta_{ij} = 1) = P(y_{ij} \neq 0)$ of the target variable to be nonzero:

$$\text{logit}(p_{ij}) = \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mathbf{x}_{\text{z},ij}^{\text{t}} \beta_{\text{z}} + \vartheta_{\text{z},j}, \quad \text{for} \quad j = 1, \ldots, m \quad \text{and} \quad i = 1, \ldots, N_j, \tag{6}$$

with

$$\vartheta_{\text{z},j} \sim \mathcal{N}\left(0, \sigma_{r,\text{z}}^2\right).$$

Model (5) is estimated based on the nonzero part of the sample, Model (6) is estimated based on the complete sample, resulting in the estimates $\hat{\beta}_{\text{nz}}$, $\hat{\vartheta}_{\text{nz},j}$, $\hat{\beta}_{\text{z}}$, $\hat{\vartheta}_{\text{z},j}$ for the location parameters and in estimates $\hat{\sigma}_{r,\text{nz}}$, $\hat{\sigma}_{e,\text{nz}}$, $\hat{\sigma}_{r,\text{z}}$ for the variance parameters.

Based on these estimates, y_{ij}^* and p_{ij} are estimated for all elements in the population:

$$\hat{y}_{ij}^* = \mathbf{x}_{\text{nz},ij}^{\text{t}} \hat{\beta}_{\text{nz}} + \hat{\vartheta}_{\text{nz},j}, \tag{7}$$

$$\hat{p}_{ij} = \frac{\exp\left(\mathbf{x}_{\text{z},ij}^{\text{t}} \hat{\beta}_{\text{z}} + \hat{\vartheta}_{\text{z},j}\right)}{1 + \exp\left(\mathbf{x}_{\text{z},ij}^{\text{t}} \hat{\beta}_{\text{z}} + \hat{\vartheta}_{\text{z},j}\right)}. \tag{8}$$

The estimate for y_{ij} is then taken to be the product $\hat{y}_{ij} = \hat{y}_{ij}^* \hat{p}_{ij}$, and the mean for domain j can be estimated as

$$\hat{Y}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \hat{y}_{ij}^* \hat{p}_{ij}. \tag{9}$$

Note that the model for y^* can only be fitted using the nonzero observations, whereas it is applied to predict all population elements, zero or nonzero. In order to reduce the risk of

bias, it is therefore important to also include variables that predict δ_{ij} in the model for y^*. In this article we always use the same predictors $\mathbf{x}$ in both models.

Again, for convenience, the prediction in (9) is used for all population elements, including the ones observed. The mixed models can be estimated using the function `lmer` of R-package `lme4`. Within this function, the `family` parameter is taken to be `binomial(link = "logit")` for Model (6) and gaussian for Model (5).

Chandra and Sud (2012) proposed parametric bootstrapping for the estimation of the mean squared error.

2.4.2. The Bayesian Approach

The two Models (5) and (6) can also be estimated in a Bayesian fashion using a Markov Chain Monte Carlo (MCMC) simulation. Such a simulation results in a series of draws of parameters from their joint posterior distribution given the data. An important advantage of the Bayesian MCMC approach is that the draws can be used both for computing point estimates and for measures of accuracy, including interval estimates. Parametric bootstrapping, as proposed by Chandra and Sud (2012) for the frequentist approach, is less easily available in R software packages.

The MCMC simulation is carried out over R runs. The first part of the MCMC simulation (burnin) is not used, as it depends too strongly on the starting values. Moreover, only every lth run is retained to save memory and increase the effective number of independent draws. In the end, r runs are retained for further analysis. Both R and r have to be chosen sufficiently large so that the Markov chain can converge and explore the entire distribution. There is no reason that the number of retained runs r_z and r_{nz} has to be equal for the two Models (5) and (6) to achieve this goal. Equality $r = r_z = r_{nz}$ is necessary for the computation of model estimates for Y_j. In all MCMC simulations carried out for this article we have taken $R = 40{,}000$ runs with a burnin of 20,000 and thinning by retaining each 20th iteration, so that $r = 1{,}000$ draws are retained for posterior analysis. From inspection of trace plots and autocorrelations, these numbers were seen to be adequate.

From the parameter draws obtained for both MCMC simulations, posterior draws for the small-area quantities of interest can be computed by simulating from the posterior predictive distributions:

1. Draw residuals $e_{ij,\rho} \sim \mathcal{N}\left(0, \sigma^2_{e,\mathrm{nz},\rho}\right)$ independently for all population units i, j and for each MCMC iteration $\rho = 1, \ldots, r$, and form posterior predictions

$$y^*_{ij,\rho} = \mathbf{x}^{\mathrm{t}}_{\mathrm{nz},ij}\beta_{\mathrm{nz},\rho} + \vartheta_{\mathrm{nz},j,\rho} + e_{ij,\rho}.$$

   All parameter draws $\sigma_{e,\mathrm{nz},\rho}$, $\beta_{\mathrm{nz},\rho}$, $\vartheta_{\mathrm{nz},j,\rho}$ are part of the MCMC simulation output.
2. Similarly, draw zero indicators independently from the Bernoulli distribution according to

$$\delta^*_{ij,\rho} \sim \mathrm{Be}(p_{ij,\rho}),$$

with probability of a nonzero response value

$$p_{ij,\rho} = \frac{\exp\left(\mathbf{x}_{z,ij}^{t}\beta_{z,\rho} + \vartheta_{z,j,\rho}\right)}{1 + \exp\left(\mathbf{x}_{z,ij}^{t}\beta_{z,\rho} + \vartheta_{z,j,\rho}\right)}.$$

3. Combine the posterior predictive draws to obtain posterior draws for the small-area estimands,

$$Y_{j,\rho}^{*} = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij,\rho}^{*}\,\delta_{ij,\rho}^{*}. \tag{10}$$

Estimates for the domain means of interest are now obtained as MCMC approximations of the posterior means, that is,

$$\hat{Y}_{j,\text{mcmc}} = \frac{1}{r}\sum_{\rho=1}^{r} Y_{j,\rho}^{*}.$$

The mean squared error of $\hat{Y}_{j,\text{mcmc}}$ under the model, that is, the posterior variance, is approximated by

$$\text{mse}(\hat{Y}_{j,\text{mcmc}}) = \frac{1}{r}\sum_{\rho=1}^{r}\left(Y_{j,\rho}^{*} - \hat{Y}_{j,\text{mcmc}}\right)^2. \tag{11}$$

Credible intervals are also considered. In particular, highest posterior 95% intervals have been computed using the R package `coda` (Plummer et al. 2006).

The MCMC simulations have been carried out using the function `MCMCglmm` from the R package of the same name (Hadfield 2010), which supports both models by way of Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990). We use weakly informative default priors as implemented in `MCMCglmm` for the coefficients and variance parameters in both models. In particular, the regression coefficients in both models are assigned normal priors with zero mean and very large variance. Following Gelman (2006), we use parameter-expanded inverse-chi-squared priors for the random effect variances in both models, implying half-Cauchy priors on the standard-deviation parameters. The scales of the half-Cauchy priors are taken to be 25, larger than the scale of the response variable in both models. The half-Cauchy priors are more robust than inverse chi-squared priors and their parameter-expansion representation also improves convergence and mixing of the Gibbs sampler, especially in situations with relatively small random effect variances (Gelman et al. 2008). For the residual variance of Model (5), a default noninformative prior $p(\sigma_{e,\text{nz}}^2) \propto 1/\sigma_{e,\text{nz}}^2$ is used.

2.4.3. Correlated Random Effects

In Pfeffermann et al. (2008), a single two-part model is used that allows for correlations between the random effects of the two submodels. It is possible that such a model would better fit the data. For this article we have chosen to use the somewhat simpler model in which components are treated independently. The main reason for this simplification is

that the separate models can be fit using relatively fast and standard functions in R. In an example, Pfeffermann et al. (2008) showed that taking the correlation into account only slightly improved the accuracy of the estimates.

3. Simulation with Artificial Populations

3.1. Lay-Out of the Simulation

To investigate the properties of the ZERO and to compare it with the SR and the EBLUP, a simulation with artificial populations is carried out. From the artificial populations, samples are drawn repeatedly. Based on these samples, the SR, EBLUP, and ZERO are computed. In most cases, only the frequentist approach (ZERO-F) is used because the MCMC simulation (ZERO-B) takes much more computation time. This choice makes it possible to simulate many different situations. In a small part of the investigated situations, the MCMC approach is also applied and both approaches are compared.

We start with the description of the main part of the simulations with artificial populations. The artificial populations consist of $m = 50$ domains with $N = 60{,}000$ elements. The domains are not equally sized. The domain size increases from 30 for the first five domains up to 3,250 for the last domain.

The creation of the artificial populations starts with drawing an auxiliary variable x from the normal distribution $\mathcal{N}(2,2.25)$. The mean of the auxiliary variable is then more or less equal for all domains. This is not realistic. To get an idea of the consequences of unequal means of the auxiliary variable, the value of the 0.9-quantile of the vector x is added for one randomly chosen domain. This is not realistic either, but it makes it easier to analyze the effects of such a deviation. The random effects $\vartheta_{\mathrm{nz},j}$ and $\vartheta_{\mathrm{z},j}$ for the domains $j = 1, \ldots, m$ are independently distributed following $\mathcal{N}\left(0, \sigma^2_{r,\mathrm{nz}}\right)$ and $\mathcal{N}\left(0, \sigma^2_{r,\mathrm{z}}\right)$. The target variable is then computed as $y_{ij} = y^*_{ij}\delta_{ij}$, where y^* and δ are generated according to Models (5) and (6) and $\delta_{ij} \sim \mathrm{Be}(p_{ij})$ is Bernoulli distributed taking value 1 with probability p_{ij}. Model (6) is extended with residuals $e_{ij,\mathrm{z}} \sim \mathcal{N}\left(0, \sigma^2_{e,\mathrm{z}}\right)$. In both models the vector of covariates consists of two components, the intercept and the generated auxiliary variable x. The corresponding coefficients will be referred to as $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$ with subscripts 0 and 1 corresponding to the intercept and x, respectively.

With different choices for $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$, $\sigma^2_{r,\mathrm{nz}}$, $\sigma^2_{r,\mathrm{z}}$, $\sigma^2_{e,\mathrm{nz}}$, $\sigma^2_{e,\mathrm{z}}$ different types of populations can be created. In total, 48 situations based on different parameter sets are investigated. The parameters are chosen in such a way that populations with a wide range of properties are included in the study, with

- a small (around 0.1), medium (around 0.5), or large proportion (around 0.85) of nonzeros by an appropriate choice of $\beta_{0,\mathrm{nz}}$, $\beta_{1,\mathrm{nz}}$, $\beta_{0,\mathrm{z}}$, $\beta_{1,\mathrm{z}}$,
- a small (around 0.2) or large (around 0.7) correlation between the auxiliary variable x and p, by an appropriate choice of $\sigma^2_{e,\mathrm{z}}$,
- a small (around 0.3) or large (around 0.7) correlation between the auxiliary variable x and y^*, by an appropriate choice of $\sigma^2_{e,\mathrm{nz}}$,
- small or large random effects $\vartheta_{\mathrm{z},j}$ and $\vartheta_{\mathrm{nz},j}$ by an appropriate choice of $\sigma^2_{r,\mathrm{z}}$ and $\sigma^2_{r,\mathrm{nz}}$. In the case of small random effect variances, their frequentist estimates are often zero.

The considered sets of parameters and corresponding types of populations are shown in Table 1.

For each set of parameters, ten different populations are created, and with each population, a simulation with 500 runs is carried out. In each run, a sample of size $n = 2{,}000$ using simple random sampling without replacement is drawn. By creating different populations of each type, coincidences in the populations have less influence. The number of ten populations per set of parameters turns out to be adequate, for the generation of different sets of ten populations consistently gives almost the same properties. At the same time, with 500 runs for each population it is possible to analyze the results for different domains, for example domains with large random effects.

In addition to the simulation with 48 different parameter sets, a few special cases are investigated. First, the simulations with the first four parameter sets are repeated with population and sample sizes that are three times as large for all domains. Second, a correlation of 0.5 and 0.9 between the random effects of the two model parts is added. This is also investigated with the first four parameter sets, with the original population and sample sizes. Third, the simulations with the first four parameter sets are repeated using a Bayesian approach (with independent random effects). Here, only a single population is created, for which a simulation with 1,000 runs is carried out. The frequentist approach is applied to the same 1,000 samples.

In SAE it is sometimes useful to include the domain mean $\bar{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij}$ as auxiliary information (Bafumi and Gelman 2006; Neuhaus and McCulloch 2006). It appears that this is also the case for the EBLUP in this application, especially for the domain where the 0.9-quantile of the vector $\mathbf{x}$ is added. Therefore, for the EBLUP $\mathbf{x}_{ij} = (1, x_{ij}, \bar{x}_j)^t$. For the other estimators $\mathbf{x}_{ij} = (1, x_{ij})^t$ is used, as the additional area-level covariate would slightly deteriorate the accuracy of these estimates (results not presented).

3.2. *Evaluation Measures*

The most important quality measure of the estimators is the accuracy measured by the mean squared error (mse). We use the root mse (rmse), computed as

$$\text{rmse}_j = \sqrt{\sum_{q=1}^{\nu} \left(\hat{Y}_{j,q} - Y_j^{\text{mean}}\right)^2 / \nu}, \tag{12}$$

with Y_j^{mean} the population mean of domain j for the target variable y, $\hat{Y}_{j,q}$ the estimate for this population mean based on one of the methods in the qth run of the simulation, and ν the number of runs in the simulation. In the simulation with artificial populations, $\nu = 500$.

The mse is the sum of the variance and the squared bias. In order to further analyze the accuracy of the methods, the standard deviation (root of the variance, sd) and the bias are also discussed. These measures are computed as

$$\text{sd}_j = \sqrt{\sum_{q=1}^{\nu} (\hat{Y}_{j,q} - \bar{Y}_j)^2 / \nu}, \quad \text{bias}_j = \sum_{q=1}^{\nu} \left(Y_j^{\text{mean}} - \hat{Y}_{j,q}\right) / \nu \tag{13}$$

where $\bar{Y}_j = \sum_{q=1}^{\nu} \hat{Y}_{j,q} / \nu$ is the mean of the estimates.

Table 1. Description of the types of populations, model parameters, fractions nonzeros and population means.

No.	$\beta_{0,z}$	$\beta_{1,z}$	$\beta_{0,nz}$	$\beta_{1,nz}$	$\sigma_{r,z}$	$\sigma_{r,nz}$	$\sigma_{e,z}$	$\sigma_{e,nz}$	Fraction nonzeros	Popmean
1	−4	2.0	10	1	0.2	0.08	1.0	1	0.51	6.70
2	−4	2.0	10	1	2.0	0.08	1.0	1	0.50	6.63
3	−4	2.0	10	1	0.2	0.80	1.0	1	0.51	6.70
4	−4	2.0	10	1	2.0	0.80	1.0	1	0.50	6.64
5	−4	2.0	30	1	0.2	0.08	1.0	5	0.51	16.87
6	−4	2.0	30	1	2.0	0.08	1.0	5	0.51	16.86
7	−4	2.0	30	1	0.2	0.80	1.0	5	0.51	16.78
8	−4	2.0	30	1	2.0	0.80	1.0	5	0.51	16.80
9	−1	0.5	10	1	0.2	0.08	2.0	1	0.50	6.31
10	−1	0.5	10	1	2.0	0.08	2.0	1	0.52	6.26
11	−1	0.5	10	1	0.2	0.80	2.0	1	0.51	6.27
12	−1	0.5	10	1	2.0	0.80	2.0	1	0.50	6.21
13	−1	0.5	30	1	0.2	0.08	2.0	5	0.51	16.39
14	−1	0.5	30	1	2.0	0.08	2.0	5	0.51	16.35
15	−1	0.5	30	1	0.2	0.80	2.0	5	0.50	16.25
16	−1	0.5	30	1	2.0	0.80	2.0	5	0.50	16.27
17	0	2.0	10	1	0.2	0.08	0.2	1	0.88	10.86
18	0	2.0	10	1	2.0	0.08	0.2	1	0.83	10.45
19	0	2.0	10	1	0.2	0.80	0.2	1	0.88	10.87
20	0	2.0	10	1	2.0	0.80	0.2	1	0.85	10.47
21	0	2.0	30	1	0.2	0.08	0.2	5	0.88	28.33
22	0	2.0	30	1	2.0	0.08	0.2	5	0.85	27.34
23	0	2.0	30	1	0.2	0.80	0.2	5	0.88	28.42
24	0	2.0	30	1	2.0	0.80	0.2	5	0.84	27.41
25	2	0.5	10	1	0.2	0.08	2.0	1	0.86	10.50
26	2	0.5	10	1	2.0	0.08	2.0	1	0.81	9.90
27	2	0.5	10	1	0.2	0.80	2.0	1	0.86	10.51
28	2	0.5	10	1	2.0	0.80	2.0	1	0.82	9.99
29	2	0.5	30	1	0.2	0.08	2.0	5	0.86	27.83
30	2	0.5	30	1	2.0	0.08	2.0	5	0.80	26.17
31	2	0.5	30	1	0.2	0.80	2.0	5	0.86	27.72
32	2	0.5	30	1	2.0	0.80	2.0	5	0.81	26.05
33	−9	2.0	10	1	0.3	0.10	0.5	1	0.09	1.36
34	−9	2.0	10	1	3.0	0.10	0.5	1	0.16	2.15
35	−9	2.0	10	1	0.3	1.00	0.5	1	0.10	1.37
36	−9	2.0	10	1	3.0	1.00	0.5	1	0.15	2.04
37	−9	2.0	30	1	0.3	0.10	0.5	5	0.09	3.24
38	−9	2.0	30	1	3.0	0.10	0.5	5	0.15	5.09
39	−9	2.0	30	1	0.3	3.00	0.5	5	0.09	3.27
40	−9	2.0	30	1	3.0	3.00	0.5	5	0.16	5.09
41	−6	0.6	10	1	0.3	0.10	2.5	1	0.07	0.95
42	−6	0.6	10	1	3.0	0.10	2.5	1	0.14	1.80
43	−6	0.6	10	1	0.3	1.00	2.5	1	0.07	0.94
44	−6	0.6	10	1	3.0	1.00	2.5	1	0.15	1.83
45	−6	0.6	30	1	0.3	0.10	2.5	5	0.07	2.34
46	−6	0.6	30	1	3.0	0.10	2.5	5	0.14	4.52
47	−6	0.6	30	1	0.3	3.00	2.5	5	0.07	2.37
48	−6	0.6	30	1	3.0	3.00	2.5	5	0.14	4.57

Since the population size for the first five domains is only 30 and the inclusion probability is $\frac{1}{30}$, empty samples occur regularly for these domains in the simulation. In these runs the SR cannot be computed. In the comparison of the accuracy of the SR with the EBLUP and ZERO-F, the first five domains are therefore ignored. In the other domains, empty samples are very rare but not impossible in the simulation. These runs are ignored in the computation of the abovementioned measures rmse_j, bias_j, and sd_j for the SR. Since these cases are very rare, this does not disturb the results.

In the simulation with ten populations with the same parameters, the mean of these measures over the ten populations is computed.

3.3. Results

Table 2 shows the mean absolute bias and mean rmse over the domains and over the ten created populations for the SR, the EBLUP and ZERO-F. In the first six columns of the table, where the SR is compared with the EBLUP and ZERO-F, only domains 6−50 are included, as mentioned in the end of Subsection 3.2. The table shows that in all cases considered, both SAE methods are more accurate than the SR, and ZERO-F is more accurate than the EBLUP. The gain in accuracy strongly depends on the properties of the population. The following points are noticed:

- the SR is generally approximately design unbiased. Small nonzero values are due to the approximate nature of SR's design unbiasedness and to the finite number of simulation runs.
- Both model-based SAE methods are biased. The bias of the EBLUP is generally only slightly larger than the bias of ZERO-F. The model misspecification does not cause a serious bias of the EBLUP.
- Generally, the improvement in accuracy of both SAE methods with respect to the SR is very large in the cases with small $\sigma_{r,\text{z}}$ (odd numbers). In those cases, the rmse is often more than halved by the SAE methods. In the case of large $\sigma_{r,\text{z}}$, the rmse of the SAE methods is usually around ten percent smaller than the rmse of the SR.
- In some cases, the gain in accuracy of ZERO-F with respect to the EBLUP in the five smallest domains is substantially larger than in the other domains. Therefore, it is important to compare the EBLUP and ZERO-F with and without these domains included.
- In many cases, the additional gain in accuracy by using the ZERO-F instead of the EBLUP is only five percent to ten percent.
- Larger gains with ZERO-F instead of the EBLUP are possible in the case of large $\sigma_{r,\text{nz}}$, small $\sigma_{r,\text{z}}$ and a small residual variance $\sigma^2_{e,\text{nz}}$, especially if the nonzero fraction is around 0.5 or 0.85 (number 3,11, 19, 27).
- Larger gains with ZERO-F instead of the EBLUP are also possible in the case of a small residual variance $\sigma^2_{e,\text{z}}$ if the nonzero fraction is around 0.1 or 0.85 (number 17−24, 33−40). This is not surprising as small $\sigma_{e,\text{z}}$ means that Model (6) is almost the true model used to simulate the data. The gain is somewhat larger if the nonzero fraction is around 0.1 than if it is 0.85.
- Altogether, the possible gain with ZERO-F instead of the EBLUP depends only slightly on the nonzero fraction.

Table 2. Mean absolute bias and mean rmse.

	Domains							
No.	bias 6–50 SR	bias 6–50 EBLUP	bias 6–50 ZERO-F	rmse 6–50 SR	rmse 6–50 EBLUP	rmse 6–50 ZERO-F	rmse 1–50 EBLUP	rmse 1–50 ZERO-F
1	0.03	0.21	0.20	0.78	0.29	0.27	0.38	0.33
2	0.03	0.18	0.16	0.74	0.70	0.65	0.82	0.77
3	0.03	0.28	0.20	0.78	0.41	0.30	0.49	0.36
4	0.03	0.18	0.16	0.74	0.70	0.65	0.83	0.78
5	0.08	0.54	0.53	2.14	0.79	0.74	0.97	0.83
6	0.07	0.52	0.44	2.04	1.92	1.73	2.25	2.06
7	0.07	0.64	0.62	2.13	0.91	0.85	1.08	0.93
8	0.07	0.53	0.50	2.05	1.92	1.76	2.39	2.11
9	0.04	0.26	0.26	1.01	0.38	0.36	0.43	0.40
10	0.03	0.23	0.22	0.89	0.85	0.83	1.03	0.99
11	0.04	0.34	0.27	1.01	0.49	0.38	0.57	0.44
12	0.03	0.25	0.24	0.90	0.86	0.85	1.07	1.02
13	0.09	0.73	0.74	2.73	1.03	0.99	1.17	1.12
14	0.09	0.60	0.55	2.38	2.28	2.20	2.69	2.63
15	0.09	0.73	0.72	2.72	1.06	0.99	1.22	1.16
16	0.09	0.58	0.58	2.42	2.29	2.24	2.68	2.63
17	0.02	0.13	0.12	0.51	0.18	0.16	0.22	0.18
18	0.02	0.14	0.12	0.53	0.48	0.41	0.54	0.46
19	0.02	0.19	0.11	0.51	0.40	0.22	0.43	0.26
20	0.02	0.13	0.12	0.53	0.49	0.44	0.61	0.51
21	0.06	0.32	0.31	1.64	0.50	0.44	0.65	0.51
22	0.06	0.49	0.38	1.68	1.50	1.19	1.75	1.35
23	0.06	0.50	0.42	1.63	0.74	0.65	0.86	0.72
24	0.06	0.48	0.45	1.67	1.48	1.26	1.64	1.42
25	0.02	0.17	0.17	0.68	0.25	0.24	0.31	0.28
26	0.02	0.20	0.19	0.67	0.63	0.61	0.75	0.70
27	0.03	0.25	0.16	0.68	0.46	0.27	0.49	0.31
28	0.02	0.19	0.20	0.67	0.63	0.62	0.74	0.71
29	0.07	0.45	0.46	1.98	0.67	0.64	0.77	0.74
30	0.07	0.56	0.52	1.94	1.80	1.64	2.27	1.88
31	0.07	0.61	0.54	1.98	0.86	0.79	0.99	0.92
32	0.07	0.55	0.57	1.96	1.82	1.72	2.12	1.95
33	0.02	0.16	0.14	0.59	0.23	0.19	0.32	0.21
34	0.02	0.16	0.12	0.63	0.59	0.47	0.69	0.56
35	0.02	0.17	0.14	0.59	0.24	0.20	0.29	0.22
36	0.02	0.16	0.13	0.62	0.58	0.47	0.70	0.56
37	0.05	0.36	0.32	1.43	0.52	0.44	0.66	0.52
38	0.05	0.38	0.31	1.54	1.44	1.17	1.77	1.43
39	0.06	0.40	0.35	1.45	0.57	0.49	0.70	0.54
40	0.05	0.38	0.30	1.53	1.43	1.17	1.78	1.40
41	0.02	0.13	0.13	0.55	0.20	0.18	0.26	0.22
42	0.02	0.14	0.13	0.56	0.54	0.51	0.74	0.62
43	0.02	0.14	0.14	0.56	0.21	0.19	0.27	0.22
44	0.02	0.15	0.13	0.59	0.56	0.53	0.68	0.60
45	0.05	0.35	0.34	1.41	0.52	0.49	0.59	0.56

Table 2. Continued.

	Domains							
No.	bias 6–50 SR	bias 6–50 EBLUP	bias 6–50 ZERO-F	rmse 6–50 SR	rmse 6–50 EBLUP	rmse 6–50 ZERO-F	rmse 1–50 EBLUP	rmse 1–50 ZERO-F
46	0.06	0.43	0.38	1.52	1.43	1.36	1.77	1.63
47	0.05	0.35	0.34	1.43	0.52	0.49	0.63	0.60
48	0.05	0.36	0.37	1.48	1.42	1.37	1.69	1.62

Another way to summarize the results about the rmse is to compute the ratios $rmse_{\text{EBLUP}}/rmse_{\text{ZERO}-\text{F}}$ for all domains and the ten populations and compute quantiles of these ratios. The results are shown in Table 3. Since the focus of this article is the comparison of the EBLUP and ZERO-F, such a comparison is not carried out between SR and SAE methods. We see the following results:

- In all cases there are at least some domains where the EBLUP is more accurate than ZERO-F.
- In almost all cases, the 35% quantile is larger than 1, so ZERO-F is more accurate than the EBLUP in at least 65% of the domains.
- In the cases with large $\sigma_{r,z}$ (even numbers) and a nonzero fraction of around 0.5, the differences between the domains are relatively small with a ten percent quantile of between 0.96 and 1.03 and a 90% quantile between 1.05 and 1.2.
- In the cases of large residual variances $\sigma^2_{e,z}$ and $\sigma^2_{e,nz}$ and a nonzero fraction of around 0.5 (number 13 and 15), the differences between the domains are also relatively small.
- For a nonzero fraction around 0.85 or 0.1, the differences between the domains are generally larger, with two exceptions (small random effects $\vartheta_{z,j}$ and $\vartheta_{nz,j}$ and large residual variance $\sigma^2_{e,z}$, nonzero fraction of around 0.85 (number 25 and 29).
- In many cases with small $\sigma_{r,z}$ (odd numbers), the EBLUP is substantially more accurate than ZERO-F for quite a large fraction of the domains (10% quantile smaller than 0.9). These are often the cases where the mean gain of ZERO-F with respect to the EBLUP over all domains is relatively large. This means that the gain in accuracy in many domains has to be paid for with some substantial loss in accuracy in some other domains.

3.4. Results for Domains

Table 3 shows that the gain in accuracy of ZERO-F with respect to the EBLUP sometimes differs strongly between the domains. An analysis of the results for the domains shows that in the situations with large $\sigma_{r,z}$ (even numbered rows), the gain in accuracy of ZERO-F generally depends strongly on the size of the random effects $\vartheta_{z,j}$. The gain is larger in the domains with the smallest (most negative) and/or the largest random effects. This gain is

Table 3. Quantiles, minimum and maximum of ratios rmse EBLUP and ZERO-F.

No.	Min	10%	25%	35%	50%	65%	75%	90%	Max
1	0.34	0.85	0.97	1.01	1.06	1.11	1.15	1.34	27.79
2	0.78	1.02	1.04	1.05	1.06	1.08	1.10	1.16	4.36
3	0.34	0.78	1.07	1.19	1.33	1.55	1.76	2.25	21.54
4	0.76	1.01	1.03	1.04	1.06	1.07	1.09	1.15	2.88
5	0.41	0.87	0.97	1.01	1.06	1.11	1.15	1.33	34.43
6	0.77	1.03	1.06	1.07	1.09	1.11	1.13	1.20	5.18
7	0.26	0.86	0.97	1.00	1.05	1.09	1.14	1.29	45.14
8	0.72	0.99	1.03	1.05	1.07	1.09	1.12	1.18	56.46
9	0.75	0.98	1.00	1.01	1.02	1.03	1.04	1.08	6.32
10	0.85	0.98	1.00	1.00	1.01	1.02	1.03	1.06	11.52
11	0.42	0.81	0.98	1.06	1.20	1.39	1.57	1.97	8.67
12	0.55	0.98	0.99	1.00	1.01	1.01	1.02	1.05	4.93
13	0.85	0.99	1.01	1.01	1.02	1.03	1.04	1.06	3.33
14	0.83	0.99	1.01	1.01	1.02	1.04	1.05	1.10	1.31
15	0.77	0.94	0.98	1.00	1.02	1.05	1.07	1.12	4.68
16	0.81	0.96	1.00	1.00	1.01	1.03	1.04	1.08	1.47
17	0.28	0.84	0.96	1.01	1.08	1.14	1.23	1.44	43.97
18	0.53	1.03	1.08	1.11	1.15	1.22	1.31	1.56	6.15
19	0.36	1.21	1.50	1.65	1.84	1.98	2.11	2.49	5.42
20	0.62	1.03	1.08	1.09	1.12	1.16	1.19	1.28	19.79
21	0.18	0.85	0.97	1.04	1.12	1.19	1.25	1.56	40.88
22	0.50	1.06	1.14	1.19	1.25	1.33	1.44	1.69	62.60
23	0.12	0.82	0.96	1.04	1.13	1.21	1.28	1.49	31.22
24	0.50	0.98	1.07	1.11	1.15	1.21	1.29	1.51	5.04
25	0.91	0.97	0.99	1.01	1.02	1.05	1.06	1.10	17.24
26	0.64	0.95	0.99	1.00	1.04	1.08	1.13	1.28	11.82
27	0.40	0.94	1.27	1.42	1.68	1.93	2.10	2.56	5.27
28	0.65	0.94	0.98	1.00	1.02	1.05	1.07	1.14	12.87
29	0.71	0.97	0.99	1.01	1.03	1.05	1.06	1.09	6.46
30	0.61	0.98	1.02	1.04	1.10	1.19	1.25	1.37	16.10
31	0.64	0.88	0.97	1.00	1.06	1.11	1.16	1.30	3.43
32	0.62	0.94	1.00	1.02	1.05	1.10	1.16	1.33	12.37
33	0.39	0.79	0.99	1.10	1.21	1.35	1.49	1.99	18.16
34	0.48	1.09	1.16	1.18	1.24	1.33	1.46	1.80	8.12
35	0.16	0.80	0.96	1.08	1.20	1.34	1.48	1.97	18.84
36	0.52	1.07	1.14	1.17	1.23	1.33	1.42	1.75	4.21
37	0.13	0.76	0.95	1.03	1.14	1.25	1.37	1.90	11.24
38	0.51	1.07	1.13	1.16	1.23	1.34	1.51	1.85	5.12
39	0.11	0.76	0.94	1.02	1.13	1.27	1.38	1.85	9.72
40	0.43	1.06	1.13	1.15	1.21	1.34	1.52	1.76	5.97
41	0.45	0.88	0.94	0.98	1.01	1.06	1.09	1.22	6.51
42	0.60	0.95	1.00	1.02	1.08	1.17	1.22	1.35	6.02
43	0.59	0.85	0.94	1.00	1.06	1.11	1.18	1.37	8.00
44	0.61	0.96	1.00	1.02	1.08	1.15	1.21	1.34	4.13
45	0.64	0.90	0.96	0.98	1.01	1.05	1.08	1.17	3.00
46	0.55	0.95	0.99	1.01	1.06	1.12	1.18	1.31	4.81
47	0.61	0.85	0.92	0.96	1.01	1.07	1.12	1.27	2.26
48	0.54	0.95	0.99	1.01	1.05	1.11	1.16	1.27	3.16

Table 4. Ratio mean rmse EBLUP and ZERO-F over the ten created populations and over groups of domains, ordered by size of random effects $\vartheta_{z,j}$.

No.	Fraction nonzeros	Ratio domains 1–10	Ratio domains 11–40	Ratio domains 41–50
2	0.50	1.10	1.05	1.08
18	0.83	1.09	1.15	1.43
34	0.16	1.60	1.23	1.12

caused by both a smaller bias and a smaller standard deviation of ZERO-F in these domains. In situations with around 50% nonzero target variables, the gain in accuracy is similar in the domains with the smallest and the largest random effects. In situations with around 85% nonzero target variables, this gain is larger in the domains with the largest random effects. In situations with around ten percent nonzero target variables, it is the opposite. This is demonstrated for three situations in Table 4. There, for three groups of domains, the mean rmse is computed over the selected domains and over the ten created populations for each situation. This is done for both the EBLUP and the ZERO-F. The column 'Ratio domains 1–10' shows the ratio of both values for the ten domains with the smallest (most negative) random effects $\vartheta_{z,j}$. The same ratio for the ten domains with the largest random effects is given in the column 'Ratio domains 41–50' and the ratio for the other 30 domains is computed in column 'Ratio domains 11–40'. For the other situations with large $\sigma_{r,z}$, similar results are found. However, sometimes the pattern is disturbed due to coincidences in the domains.

In the situations with small $\sigma_{r,z}$ (odd numbers), there is no visible influence of the size of the random effects $\vartheta_{z,j}$ on the gain in accuracy in the domains of ZERO-F with respect to the EBLUP. In a few cases, a similar dependency on the size of the random effects $\vartheta_{\mathrm{nz},j}$ is visible. The gain in accuracy of ZERO-F with respect to the EBLUP does not depend strongly on the domain size. The gain in accuracy of both SAE methods with respect to the design-based SR decreases with increasing sample size, a rather general phenomenon in small-area estimation.

In many situations with small $\sigma_{r,z}$, the differences between the domains cannot be explained by domain size or the size of the random effects.

The results for the domain where the 0.9 quantile of the vector $\mathbf{x}$ is added are special in many cases. There, the rmse of the EBLUP and the SR are similar, and the rmse of the ZERO-F is smaller, whereas in most of the other domains, the rmse of the EBLUP is smaller than the one of the SR.

3.5. Results for Larger Populations and for Correlated Random Effects

The simulations for the first four situations are repeated for larger populations ($N = 180{,}000$) and larger samples ($n = 6{,}000$). The results for these simulations are similar to those for the smaller populations and samples discussed in the previous subsection and are therefore not included in detail. As expected, the possible gain in accuracy by using SAE methods instead of the SR is smaller when the sample size increases. The gain in accuracy of ZERO-F with respect to the EBLUP is more or less equal to that with smaller sample sizes.

Table 5. Mean absolute bias and mean rmse, correlation 0.5.

	Domains							
No.	bias 6–50 SR	bias 6–50 EBLUP	bias 6–50 ZERO-F	rmse 6–50 SR	rmse 6–50 EBLUP	rmse 6–50 ZERO-F	rmse 1–50 EBLUP	rmse 1–50 ZERO-F
1	0.03	0.21	0.21	0.78	0.30	0.28	0.35	0.32
2	0.03	0.19	0.18	0.74	0.70	0.65	0.80	0.76
3	0.03	0.29	0.22	0.78	0.47	0.31	0.53	0.36
4	0.03	0.18	0.19	0.75	0.71	0.67	0.89	0.81

Furthermore, the simulations for the first four situations are repeated with correlated random effects $\vartheta_{z,j}$ and $\vartheta_{nz,j}$. The results are shown in Table 5 (for correlation 0.5) and Table 6 (for correlation 0.9). The accuracy of the SR is, as expected, not affected by this correlation. The effect on the accuracy of the EBLUP and ZERO-F is also small. Only for the EBLUP in Situation 3 is there some loss in accuracy, compared with the situation with uncorrelated random effects (Table 2). Despite the model misspecification of ZERO-F (by ignoring the correlation), the improvement of the accuracy by ZERO-F instead of the SR is of the same order as in the situation where the correlation is zero. Nevertheless, it can be useful to investigate ZERO with modelling the correlation in order to achieve an additional gain in accuracy. However, in the example of Pfeffermann et al. (2008) the improvement in accuracy by using this more complex model is very small.

3.6. Results for Bayesian Approach

Finally, the simulations for the first four situations are repeated with ZERO-B. For these simulations, a single population is created for each situation, the number of runs in the simulations being 1,000. The mean absolute bias, mean sd and mean rmse over the domains of ZERO-F and ZERO-B are shown in Table 7. The general conclusion is that the accuracy of both approaches is very similar. The bias is slightly reduced with the Bayesian approach, whereas the sd is slightly increased.

Figures 1 and 2 show boxplots of the model-based rmse based on the MCMC simulations for all 1,000 runs of the simulation for Situations 1 and 2. For Situations 3 and 4 similar results were obtained, so these results are omitted. The simulation rmses,

Table 6. Mean absolute bias and mean rmse, correlation 0.9.

	Domains							
No.	bias 6–50 SR	bias 6–50 EBLUP	bias 6–50 ZERO-F	rmse 6–50 SR	rmse 6–50 EBLUP	rmse 6–50 ZERO-F	rmse 1–50 EBLUP	rmse 1–50 ZERO-F
1	0.03	0.24	0.24	0.78	0.33	0.30	0.40	0.35
2	0.03	0.18	0.19	0.74	0.70	0.66	0.86	0.77
3	0.03	0.29	0.22	0.78	0.48	0.31	0.52	0.35
4	0.03	0.16	0.19	0.75	0.71	0.67	0.91	0.84

Table 7. Bias, sd and rmse of ZERO-F and ZERO-B, mean over the domains, for bias mean of absolute values.

No.	freq bias	mcmc bias	freq sd	mcmc sd	freq rmse	mcmc rmse
1	0.239	0.231	0.162	0.172	0.307	0.308
2	0.222	0.217	0.644	0.648	0.714	0.715
3	0.255	0.248	0.208	0.215	0.350	0.351
4	0.234	0.229	0.660	0.664	0.735	0.735

computed with (12), are added to the figures. Again, there is a large difference between the situations with large and small $\sigma_{r,z}$. For large $\sigma_{r,z}$ (Situation 2, Figure 2), the model-based rmse tracks the simulation rmse very well. In those cases, the variation of the model-based rmse is quite small (except for the smallest domains) and the bulk of the distribution is positioned closely around the simulation rmse. If $\sigma_{r,z}$ is small (Situation 1, Figure 1), the bulk of the distribution of the model-based rmse often deviates from the simulation rmse. The model-based rmses do not vary much over the domains in these cases, whereas the simulation rmses do. Nevertheless, the model-based rmses are of the same order of magnitude as the simulation rmses and can therefore be useful as an indication for the accuracy of the estimates, even in a repeated sampling sense.

4. Simulation with Dutch HBS Data

4.1. Design of HBS

The aim of the Dutch Household Budget Survey (HBS) is to measure the expenditures of households. Some of these expenses are on a regular basis, for example often the same amount of money is paid every month for rent and insurance premiums. Other expenses are quite regular, although with varying amounts of money spent. This often concerns cheaper products; for example, food is bought almost every week. Finally, there are also

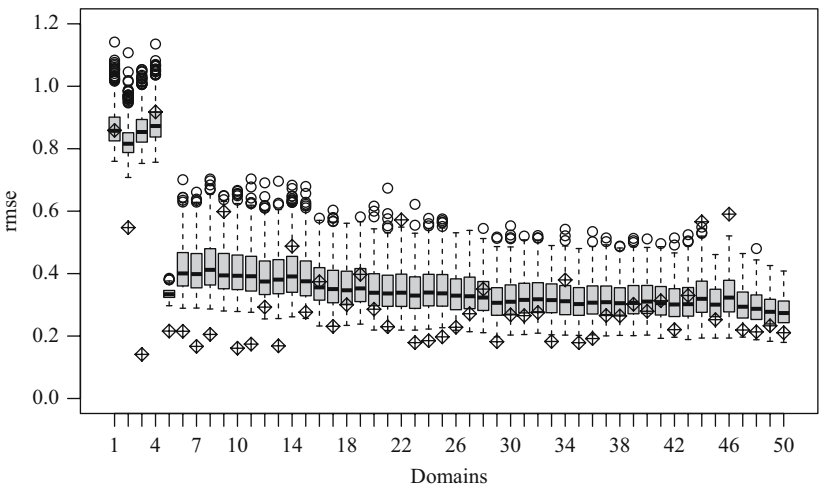

Fig. 1. Boxplots of MCMC estimates for rmse of ZERO-B from 1,000 simulation runs and rmse based on simulation (diamonds), Situation 1.

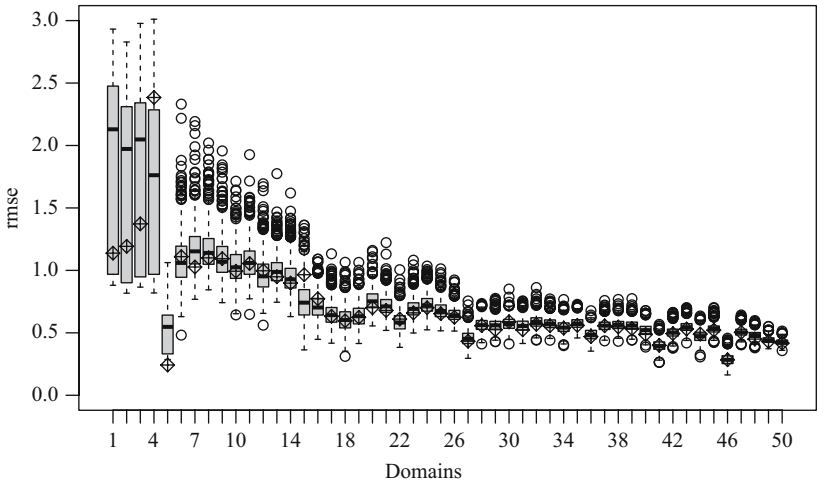

Fig. 2. Boxplots of MCMC estimates for rmse of ZERO-B from 1,000 simulation runs and rmse based on simulation (diamonds), Situation 2.

expenses that are more rare, for example furniture or clothes. These products are often, but not always, relatively expensive. Therefore, the HBS considers three kinds of expenditures which are measured in different parts of the survey. In this simulation we consider three kinds of expenditures, mainly measured in the part "large expenditures". In this part, the responding households keep a diary of their expenditures over €20.

The HBS has been redesigned repeatedly with the aim to increase response rates and decrease costs. Since 2012, the diary for large expenditures has been kept for a period of four weeks. For the simulation, data from the period 2005–2010 is used. In those years, the diary for large expenditures was kept for three months. To approximate the current design as far as possible, we use periods of one month in the simulation, in which each original sample household with expenditures over three months is considered as three independent sample households with expenditures over one month.

Data from 2005–2010 are combined in a single dataset of $N = 100{,}000$ households with expenditures for one month. The expenditures are corrected for inflation to have comparable prices over the years. This artificial population can be considered a representative sample from the population of Dutch households. The complete Dutch population consists of more than seven million households. The artificial population is chosen to be smaller for computational reasons.

Based on the HBS, household expenditures are published for the entire country and for different classifications in subpopulations. In this article, we consider a classification in $m = 11$ types of households. Table 8 shows these domains and their sizes in the artificial population. In the simulation, samples of size $n = 5{,}000$ are drawn by simple random sampling without replacement. Complications caused by different response probabilities which occur in practice are avoided. In the simulation 3,000 samples are drawn.

In the simulation, the expenditures for clothes, men's clothes and motor fuel are used as target variables. All three variables contain substantial amounts of zeros. This is partly because the households had no expenditures of this kind in the considered month, and

Table 8. Population size per type of household in artificial population of 100,000 households.

No.	Description	Population size
1	single man, younger than 65 years	12,976
2	single man, 65 years or older	2,985
3	single woman, younger than 65 years	11,176
4	single woman, 65 years or older	8,141
5	couple, main wage earner younger than 65 years	18,781
6	couple, main wage earner 65 years or older	10,514
7	couple with child(ren), all children younger than 18 years	19,803
8	couple with child(ren), at least one child 18 years or older	8,020
9	one-parent family, all children younger than 18 years	4,006
10	one-parent family, at least one child 18 years or older	2,291
11	other households	1,307

partly because they do not have expenditures of this kind at all. For example, households with only female members generally do not buy men's clothes and households without a car or a motorcycle do not buy motor fuel.

Table 9 shows the percentages of nonzero expenditures, the means of the nonzero expenditures, and the overall expenditure means for the three target variables and the eleven household types. There are substantial differences between the domains. These differences suggest that substantial random effects can be expected. However, part of the differences may be explained by other auxiliary variables used in the models.

For the considered estimators (SR, EBLUP, and ZERO), the same auxiliary information is used. This is a combination of different socio-economic variables. Income is the only continuous auxiliary variable; furthermore, categorical variables about the source of income of the main wage earner, the housing situation (owner or tenant) are used. Since the expenditures vary over the course of the year, quarter is also added.

Table 9. Percentage nonzero expenditures, mean of nonzero expenditures and overall mean for three target variables and eleven household types.

	Percentage			Mean of nonzeros			Mean expenditure		
No.	Clothes	Motor fuel	Men's clothes	Clothes	Motor fuel	Men's clothes	Clothes	Motor fuel	Men's clothes
1	23	46	22	134	99	136	31	45	29
2	21	57	20	111	71	111	24	40	22
3	40	41	0.7	109	75	89	43	30	0.6
4	35	28	0.6	114	55	75	39	15	0.5
5	47	73	22	164	109	136	76	79	30
6	40	70	17	142	78	111	57	54	19
7	56	73	20	160	113	130	89	82	26
8	56	76	27	164	121	126	92	92	33
9	44	53	3.5	105	86	94	46	45	3.3
10	42	61	9.5	124	89	110	52	54	10
11	44	64	18	158	110	128	69	71	23

The accuracy of the estimates slightly depends on the auxiliary information, as was investigated in a preliminary study (details not presented). However, the main results do not change, as long as the model is not overfitted.

As in Section 3, the point estimates of ZERO-F and ZERO-B are very similar. Therefore, only the results under ZERO-B are presented. In Subsection 4.2 the point estimates are discussed, and in Subsection 4.3 results for the mse estimates as well as credible intervals are presented.

4.2. Point Estimates

The rmses for the different estimators for the target variable clothes are shown for the eleven domains in Figure 3. The results are mixed: for each estimator, there is at least one domain where this estimator has the largest rmse. On the other hand, the SAE methods are more accurate than the SR for a majority of the domains. This is most clear for ZERO-B which is more accurate than the SR for all but one domain. The Domains 1 and 2 (single men, younger than 65 years/65 years and older) are special for clothes, since these households do not buy many clothes (compare Table 9). ZERO-B can handle these domains better than the EBLUP. The results for men's clothes and motor fuel are similar and therefore not shown in figures. There, other domains are special due to very small expenditures in these domains. For men's clothes, these are Domains 3, 4, and 9 (single women, younger than 65 years/65 years and older, one-parent family, all children younger than 18 years), where again ZERO is more accurate than the EBLUP. For motor fuel, it is

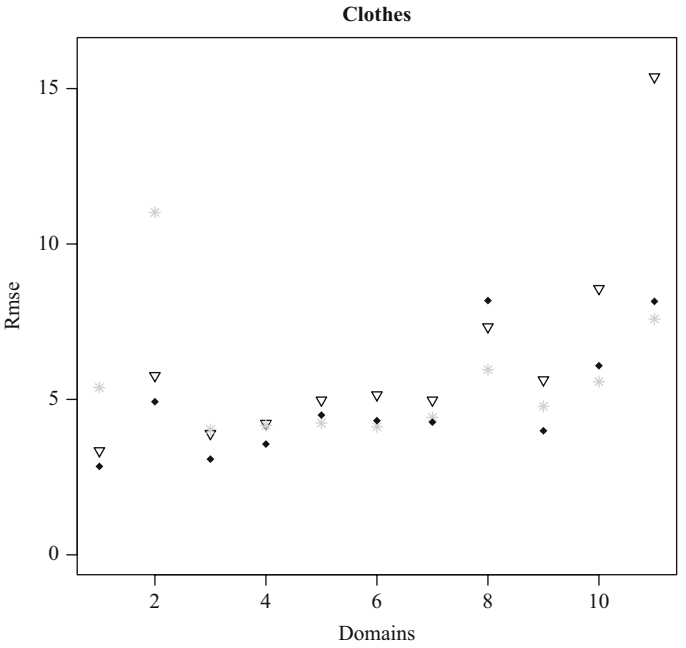

Fig. 3. Root mse of three estimators with four different fixed effects for clothes (triangle: SR, star: EBLUP, diamond: ZERO-B).

Table 10. Mean rmse (first three columns), mean relative rmse with all domains included (Columns 5–7) and mean relative rmse with Domains 3 and 4 excluded (Columns 5–7, between brackets) for three variables.

	rmse			rel. rmse		
Fixed	SR	EBLUP	ZERO-B	SR	EBLUP	ZERO-B
Clothes	6.294	5.570	4.900	0.125 (−)	0.130 (−)	0.097 (−)
Men's clothes	3.530	3.295	2.945	0.372 (0.239)	0.675 (0.238)	0.369 (0.183)
Motor fuel	3.834	3.432	3.284	0.074 (−)	0.072 (−)	0.068 (−)

Domain 4 (single women, 65 years and older). There, however, the EBLUP and ZERO-B have a similar rmse, which is much larger than the rmse of the SR.

To summarize the results, the mean of the rmse and the mean of the relative rmse over the domains are computed. Since the relative rmse for Domain 3 and 4 is extremely large for men's clothes, for this variable the mean relative rmse is also computed with these domains excluded. The results are shown in Table 10. Based on this table, it can be concluded that ZERO-B is the most accurate estimator. The fact that the mean relative rmse of ZERO-B is almost equal to the one of the SR for men's clothes (Column 5 and 7) is caused by the extremely large values for Domain 3 and 4 (compare the numbers between brackets). The EBLUP is also more accurate than the SR, but the possible gain is smaller than the one achieved by using ZERO-B. The possible gain in accuracy also varies between the target variables.

4.3. mse Estimates and Coverage

Figure 4 compares the model-based rmse obtained from the MCMC simulation computed using (11) with the design-based rmse based on the simulation for the variable clothes. Even though the two rmse concepts are quite different, it can still be useful to have good frequentist properties of the model-based rmse and of model-based intervals. In most domains the model-based rmse is on average somewhat larger than the simulation rmse.

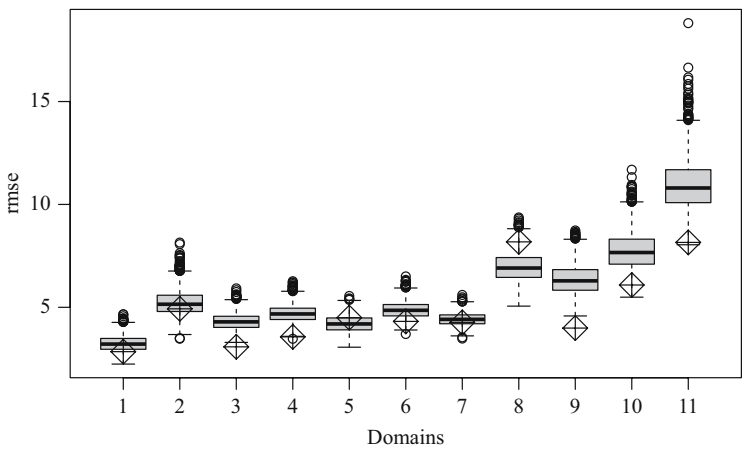

Fig. 4. Boxplot of MCMC estimates for root mse and rmse based on simulation (diamonds), clothes.

Table 11. Coverage for 95% highest posterior density intervals.

	Domain										
	1	2	3	4	5	6	7	8	9	10	11
Clothes	96.4	97.7	99.3	98.4	91.6	96.1	95.3	89.6	99.5	99.3	99.2
Men's clothes	90.0	94.4	97.4	96.9	92.7	98.5	93.4	96.6	98.2	98.6	96.3
Motor fuel	95.4	99.5	98.6	97.0	95.4	98.8	94.9	93.5	96.3	98.5	97.0

For Domain 8 it is smaller on average. Nevertheless the model-based rmse seems a useful measure of accuracy for the domain estimates, even in a repeated sampling sense. For the other two variables, no figures are shown because the results are similar.

The coverages of 95% highest posterior density intervals are shown in Table 11 for all three variables. Most coverages are not very far from 95%, although some undercoverage and overcoverage occurs depending on the domain and the variable of interest. Intervals based on the normal approximation using the model-based mse have also been computed. They are quite similar to the highest posterior density intervals, although slightly wider, and in almost all cases their coverages are close to those for the highest posterior density intervals.

5. Conclusion

Model-based small-area estimation (SAE) can be considered as an alternative to approximately design-unbiased estimation if the sample size is too small for producing reliable design-based estimates. Zero-inflated target variables occur in many surveys by national statistical institutets. Therefore, in this article three SAE methods are compared with each other and with a design-based estimator in a simulation study using zero-inflated variables. The first SAE method is the EBLUP (Rao 2003), which is the most common SAE method but ignores zero inflation. The second and third SAE method, developed by Pfeffermann et al. (2008) and Chandra and Sud (2012), take the zero inflation explicitly into account. They are based on the same models but use the Bayesian and the frequentist approach respectively. They result in similar point estimates and are referred to in abbreviation as ZERO. The general conclusion is that in the case of zero-inflated variables, an improvement of accuracy can be achieved with all SAE estimators compared with design-based methods. So the performance of the EBLUP is often satisfactory even though the model of the EBLUP is misspecified since it ignores the zero inflation. Generally ZERO is more accurate than the EBLUP. In a simulation with artificial populations, the properties of the populations can be controlled. There, ZERO is less model misspecified. The amount of improvement in accuracy of ZERO compared with the EBLUP depends on the properties of the entire population and the domains. In some populations, the improvement is negligible; in others, it is substantial. In all considered simulations, there are also some domains where the EBLUP is more accurate than ZERO.

The accuracy of the point estimates of ZERO under the frequentist approach or under the Bayesian approach is almost equal, which means that the statistician's taste can be the

deciding factor. A disadvantage of the Bayesian approach is that the computation time is higher, while an advantage is that information about the accuracy of the estimates follows directly. For the frequentist approach, no formula for the mean squared error has been developed so far. Parametric bootstrapping can be applied as proposed by Chandra and Sud (2012), which is also computationally intensive. The mean squared error estimates under the Bayesian approach do not always track the simulation error accurately. However, the mean squared error estimates seem to be useful as an indication of accuracy.

In a second simulation, real data of the Household Budget Survey (HBS) of Statistics Netherlands are used. The considered target variables, expenditures for three products, are zero inflated. In this simulation, the properties of the population cannot be controlled. Model misspecification is now more pronounced for ZERO since this estimator takes only one particular deviation from normality (zero inflation) into account, but no other possible deviations. Nevertheless, ZERO is the most accurate estimator for the majority of the domains. Contrary to the first simulation, in the simulation with HBS data there are some domains where the design-based estimator is substantially more accurate than ZERO. Such domains are very rare in the simulation with artificial data. Further results of both simulations are similar.

ZERO as used in this paper assumes a normal distribution of the nonzero part of the population. This assumption is not quite met in our application to the HBS. A suitable transformation applied to the target variable, as described in Dreassi et al. (2012) and Chandra and Chambers (2011a), could improve the model and the estimates. Furthermore, a model that replaces the normal distribution for random effects by one with wider tails might be able to better accommodate outlying random effects to prevent overshrinkage. In the continuation of this research, these potential improvements can be implemented and the results can be compared with those in this paper. Other research questions are how the estimators work if a complex design of the survey and different response probabilities must be taken into account.

6. References

Bafumi, J. and A. Gelman. 2006. "Fitting Multilevel Models When Predictors and Group Effects Correlate." Manuscript prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago. DOI: http://dx.doi.org/10.2139/ssrn.1010095.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1). Doi: http://dx.doi.org/10.18637/jss.v067.i01.

Battese, G., R. Harter, and W. Fuller. 1988. "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data." *Journal of the American Statistical Association* 83: 28–36. Doi: http://dx.doi.org/10.2307/2288915.

Boonstra, H., J. van den Brakel, B. Buelens, S. Krieg, and M. Smeets. 2008. "Towards Small Area Estimation at Statistics Netherlands." *Metron* LXVI (1): 21–49. Available at: https://www.researchgate.net/profile/Jan_Brakel/publication/227458249_Towards_small_area_estimation_at_Statistics_Netherlands/links/0c96052f8fdda8aedd000000.pdf.

Chambers, R. and N. Tzavidis. 2006. "M-Quantile Models for Small Area Estimation." *Biometrika* 93: 255–268. Doi: http://dx.doi.org/10.1093/biomet/93.2.255.

Chandra, H. and R. Chambers. 2011a. "Small Area Estimation for Skewed Data in Presence of Zeros." *The Bulletin of Calcutta Statistical Association* 63: 249–252.

Chandra, H. and R. Chambers. 2011b. "Small Area Estimation under Transformation to Linearity." *Survey Methodology* 37: 39–51.

Chandra, H. and U. Sud. 2012. "Small Area Estimation for Zero-Inflated Data." *Communications in Statistics - Simulation and Computation* 41: 632–642. Doi: http://dx.doi.org/10.1080/03610918.2011.598991.

Dreassi, E., A. Petrucci, and E. Rocco. 2012. "Small Area Estimation for Semicontinuous Skewed Georeferenced Data." Technical Report Working paper 2012/05, Dipartimento di Statistica "G. Parenti", Florence. Available at: http://local.disia.unifi.it/ricerca/pubblicazioni/working_papers/2012/wp2012_05.pdf.

Eurarea. 2004. *Project Reference Volume, deliverable d7.1.4*. Technical report, EURAREA consortium.

Gelfand, A. and A. Smith. 1990. "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85: 398–409. Doi: http://dx.doi.org/10.1080/01621459.1990.10476213.

Gelman, A. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1: 515–534. Doi: http://dx.doi.org/10.1214/06-BA117A.

Gelman, A., D.V. Dyk, Z. Huang, and W. Boscardin. 2008. "Using Redundant Parameterizations to Fit Hierarchical Models." *Journal of Computational and Graphical Statistics* 17: 95–122. Doi: http://dx.doi.org/10.1198/106186008X287337.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741. Doi: http://dx.doi.org/10.1109/TPAMI.1984.4767596.

Hadfield, J.D. 2010. "Mcmc Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33: 1–22. Doi: http://dx.doi.org/10.18637/jss.v033.i02.

Neuhaus, J. and C. McCulloch. 2006. "Separating Between- and Within-Cluster Covariate Effects by Using Conditional and Partitioning Methods." *Journal of the Royal Statistical Society B* 68: 859–872. Doi: http://dx.doi.org/10.1111/j.1467-9868.2006.00570.x.

Pfeffermann, D., B. Terryn, and F. Moura. 2008. "Small Area Estimation under a Two Part Random Effects Model with Application to Estimation of Literacy in Developing Countries." *Survey Methodology* 34: 67–72.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "Coda: Convergence Diagnosis and Output Analysis for mcmc." *R News* 6: 7–11. Available at: http://oro.open.ac.uk/id/eprint/22547 (accessed 13 October, 2016).

R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. Technical Report, R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org. (accessed 13 October, 2016).

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

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

Sinha, S.K. and J.N.K. Rao. 2009. "Robust Methods for Small Area Estimation." *Canadian Journal of Statistics* 37: 381–399. Doi: http://dx.doi.org/10.1002/cjs.10029.

Woodruff, R. 1966. "Use of a Regression Technique to Produce Area Breakdowns of the Monthly National Estimates of Retail Trade." *Journal of the American Statistical Association* 61: 496–504. Doi: http://dx.doi.org/10.2307/2282839.

Received March 2015
Revised April 2016
Accepted August 2016