Statistics Netherlands

Discussion Paper

Small area estimation with zero-inflated data - a simulation study

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

2015 | 1

Sabine Krieg, Harm Jan Boonstra and Marc Smeets

January 2015

Many target variables in official statistics follow a semicontinuous distribution with a mixture of zeros and continuously distributed positive values, also called zero-inflated variables. When reliable estimates for subpopulations with small samples are required, a model-based small area estimation method can be used, which improves the accuracy of the estimates by borrowing information from other domains. In this paper, two small area estimators are compared in a simulation study with zero-inflated target variables. The first estimator, the EBLUP, can be considered as the standard small area estimator, and is based on a linear model that assumes normal distributions. Therefore it is model-misspecified in our situation. The second estimator is based on a model that takes the zero-inflation into account and is therefore less misspecified. Both estimators are found to improve the accuracy compared to a design-based approach. The gain in accuracy is generally larger for the model that takes the zero-inflation into account. The amount of improvement depends on properties of the population. Furthermore, there are large differences in improvement between the domains.

Key words: Small area estimation, MCMC, linear mixed model, logit

1 Introduction

Traditionally national statistical institutes (NSIs) such as Statistics Netherlands prefer design-based estimation methods since these methods have the advantage of producing approximately design-unbiased estimates. However, the demand for detailed estimates for sub-populations is increasing, while at the same time budgets are under continuous pressure. Therefore, different NSIs started to investigate the possibilities of small area estimation (SAE). This model-based methodology is developed for situations where the sample sizes of the considered sub-populations (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). This estimator 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 situations, different adjustments of the EBLUP and some entirely 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 paper deals with the estimation for target 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 kind of data, the first using a Bayesian approach and the second a frequentist approach. Both approaches are used in this paper, with a small simplification of the method used in Pfeffermann et al. (2008). The two approaches result in different estimates and are therefore considered as two different estimators. They are both based on two models. The first one is a linear mixed model for the non-zero values, the second one is a generalized linear mixed model for the binary zero-indicator. This SAE method for zero-inflated data is compared in this paper with the EBLUP and with a design-based method (the survey regression estimator) in different situations. In the first part of the paper, a model-based simulation is carried out where different populations are created to investigate the properties of the three 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 developed. Then the results of the model-based simulation are described in Section 3. The results of the simulation for the

HBS follows 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 $\mathbf{x}_{ij} = (x_{ij}^1, \dots, x_{ij}^p)^t$. An intercept is always included, i.e., for the model underlying the survey regression estimator, the linear mixed model or the generalized linear mixed model, it can be assumed that $x_{ij}^1 = 1$. Population means $Y_j^{mean} = \frac{1}{N_j}\sum_{i=1}^{N_j} y_{ij}$ for target variable y for all domains $j = 1, \dots, m$ have to be estimated.

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

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

Furthermore, we write U_{nz} and S_{nz} for the part of the population and sample where $y_{ij} \neq 0$. The numbers of elements of domain j in U_{nz} and S_{nz} are denoted by $N_{nz,j}$ and $n_{nz,j}$.

2.2 Survey regression

The survey regression estimator (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 paper the SR is considered as the reference estimation method; 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

$$\widehat{Y}_j^{SR} = \widehat{Y}_j^{HT} + (\mathbf{X}_j^{mean} - \widehat{\mathbf{X}}_j^{HT})^t \widehat{\boldsymbol{\beta}}, \tag{2}$$

where

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y}.$$

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

2.3 Empirical best linear unbiased predictor (EBLUP)

Consider the linear mixed model given by

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

where

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

Here σ_e^2 is the within-area variance parameter, whereas σ_r^2 is the between-domain variance, i.e. the variance of the random domain effects. The model involves random domain intercepts. In general, the model could be extended to include other random effects such as random slopes, but we do not consider this here.

Based on model (3) we consider the Empirical Best Linear Unbiased Predictor (EBLUP), Rao (2003), to estimate the population means Y_j^{mean} for the domains $j = 1, \dots, m$. The small area estimator for Y_j^{mean} is then given by

$$\widehat{Y}_j^{EBLUP} = \mathbf{X}_j^{mean} \widehat{\boldsymbol{\beta}} + \widehat{\vartheta}_j\,. \tag{4}$$

Expressions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\vartheta}_j$ can be found in (Rao, 2003, Section 6.2.1) for the more general situation of random slope models. In (Rao, 2003, Section 7.2) the situation of the random intercept model (3) is discussed.

A refined version of (4) would use predicted values only for the non-sampled part of the population, and the observed values for themselves. However, when sampling fractions are small the difference is negligible, and we simply use (4) in this paper.

The variance parameters σ_r^2 and σ_e^2 are estimated by the method of restricted maximum likelihood (REML). The EBLUP estimator is computed with R (R Development Core Team, 2009), where the function `lmer` of package lme4 is used to fit the linear mixed model.

2.4 A small area estimator for zero-inflated data

In this section, an estimator is developed which takes the zero-inflation into account. There are two approaches to estimate the models, the frequentist approach (Section 2.4.1), which was first described by Chandra and Sud (2012), and the Bayesian approach (Section 2.4.2), first decribed by Pfeffermann et al. (2008). The two approaches result in slightly different estimates and can therefore be considered as different estimators. For both approaches we use the abbreviation ZERO in the rest of the paper, 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 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 linear mixed models. The first model describes the distribution of y^*_{ij}:

$$y^*_{ij} = \mathbf{x}^t_{nz,ij}\boldsymbol{\beta}_{nz} + \vartheta_{nz,j} + e_{ij},\ \text{for } j = 1,\dots,m \text{ and } i = 1,\dots,N_j, \tag{5}$$

where

$$\vartheta_{nz,j} \sim N(0,\sigma^2_{r,nz})\,,\ e_{ij} \sim N(0,\sigma^2_{e,nz})\,.$$

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

$$logit(p_{ij}) = ln(\frac{p_{ij}}{1-p_{ij}}) = \mathbf{x}^t_{z,ij}\boldsymbol{\beta}_z + \vartheta_{z,j},\ \text{for } j = 1,\dots,m \text{ and } i = 1,\dots,N_j, \tag{6}$$

with

$$\vartheta_{z,j} \sim N(0,\sigma^2_{r,z}).$$

Similar as for the EBLUP, we restrict ourself to random domain intercepts. Model (5) is estimated based on the non-zero part of the sample, model (6) is estimated based on the complete sample, resulting in estimates $\hat{\boldsymbol{\beta}}_{nz}$, $\hat{\vartheta}_{nz,j}$, $\hat{\boldsymbol{\beta}}_z$, $\hat{\vartheta}_{z,j}$ for the model parameters and in estimates $\hat{\sigma}_{r,nz}$, $\hat{\sigma}_{e,nz}$, $\hat{\sigma}_{r,z}$ for the variance parameters.

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

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

$$\hat{p}_{ij} = \frac{exp(\mathbf{x}^t_{z,ij}\hat{\boldsymbol{\beta}}_z + \hat{\vartheta}_{z,j})}{1 + exp(\mathbf{x}^t_{z,ij}\hat{\boldsymbol{\beta}}_z + \hat{\vartheta}_{z,j})}. \tag{8}$$

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 non-zero observations, whereas it is applied to predict all population elements, zero or non-zero. In order to reduce the risk of bias, it is therefore important to include variables that predict δ_{ij} also in the model for y^*. In this paper we always use the same predictors $\mathbf{x}$ in both models.

Again, for convenience, prediction in (9) is used for all population elements, including the ones observed. The mixed models can be estimated using 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).

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

2.4.2 The Bayesian approach

The two models (5) and (6) can also be estimated 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 not only to compute point estimates but also measures of accuracy. 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} have 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.

When both MCMC simulations are finished, estimates can be computed for every $\rho = 1, \ldots, r$, and every domain j, where the results of both MCMC simulations are combined:

$$\hat{Y}_{j,\rho} = \sum_{i=1}^{N_j} \hat{y}^*_{ij,\rho}\, \hat{p}_{ij,\rho} \tag{10}$$

with

$$\hat{y}^*_{ij,\rho} = \mathbf{x}^t_{nz,ij}\hat{\boldsymbol{\beta}}_{nz,\rho} + \hat{\vartheta}_{nz,j,\rho},$$

$$\hat{p}_{ij,\rho} = \frac{exp(\mathbf{x}^t_{z,ij}\hat{\boldsymbol{\beta}}_{z,\rho} + \hat{\vartheta}_{z,j,\rho})}{1 + exp(\mathbf{x}^t_{z,ij}\hat{\boldsymbol{\beta}}_{z,\rho} + \hat{\vartheta}_{z,j,\rho})}\,.$$

The estimates $\hat{Y}_{j,\rho}$, $\rho = 1, \ldots, r$ can be used for further analysis. In this paper, the main focus is on the point estimates, which are computed as

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

The mean squared error of the point estimates can be estimated as

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

For this paper, function `MCMCglmm` from the R package of the same name (Hadfield, 2010) is used, with `family="categorical"` for model (6). We use weakly informative default priors as implemented in `MCMCglmm` for the coefficients and variance parameters in both models. In particular, for the random effect variances we use parameter expanded priors implying half-Cauchy priors on the standard-deviation parameter (Gelman, 2006) that improve convergence and mixing of the Markov chain, especially in situations with small random effect variances.

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 sub-models. It is possible that such a model would better fit the data. For this paper 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. Pfeffermann et al. (2008) showed (for one example) that taking the correlation into account only slightly improves the accuracy of the estimates.

3 Model-based simulation

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 model-based simulation is carried out. In this simulation, an artificial population is created from which 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 in total $N = 60000$ elements. The domains are not equally sized. The domain size increases from 30 for the first five domains up to 3250 for the last domain.

The creation of the artificial populations starts with drawing an auxiliary variable x from a normal distribution with $N(2, 2.25)$. 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 $\vartheta_{nz,j}$ and $\vartheta_{z,j}$ for the domains $j = 1, \ldots, m$ are independently distributed following $N(0, \sigma^2_{r,nz})$ and $N(0, \sigma^2_{r,z})$. 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 $\delta_{ij} \sim Be(p_{ij})$ is Bernoulli distributed taking value 1 with probability p_{ij} . Model (6) is extended with residuels $e_{ij,z} \sim N(0, \sigma^2_{z,e})$. 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 $\beta_{0,nz}$, $\beta_{1,nz}$, $\beta_{0,z}$, $\beta_{1,z}$ with subscripts 0 and 1 corresponding to the intercept and x, respectively.

With different choices for $\beta_{0,nz}$, $\beta_{1,nz}$, $\beta_{0,z}$, $\beta_{1,z}$, $\sigma^2_{r,nz}$, $\sigma^2_{r,z}$, $\sigma^2_{e,nz}$, $\sigma^2_{e,z}$ different populations with different fractions of non-zero values, different correlations between target variable and auxiliary information and with small or large random effects can be created. In total, 48 different situations with different parameter sets are investigated.

For each set of parameters, 10 different populations are created, and with each population, a simulation with 500 runs is carried out. In each run, a sample of size $n = 2000$ using simple random sampling without replacement is drawn. By creating different populations, coincidences in the populations have less influence. At the same time, with 500 runs for each population it is possible to analyze the bias and results for different domains, for example domains with large random effects.

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 non-zeros	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

Table 1 Description of the populations, model parameters, fractions non-zeros and population means

Table 1 shows the parameters used to create the different populations, the proportion of non-zero values, and the mean of the population means over all 10 populations and over the 50 domains. In the first 16 populations, the proportion of non-zero values is around 0.5. In the other populations, this fraction is around 0.85 (No. 17 - 32) and around 0.1 (No. 33 - 48), respectively. In the populations with the smaller variance $\sigma^2_{e,z}$ (for example number 1 - 8), the correlation between the auxiliary variable x and p is relatively large (around 0.7) in the other cases it is smaller (around 0.2). Furthermore, a small variance $\sigma^2_{e,nz}$ (for example No. 1 - 4 and 9 - 12) results in a relatively large correlation (around 0.7) between x and y^*, in the other cases, it is around 0.3. In the cases with small variances $\sigma^2_{r,z}$ (or $\sigma^2_{r,nz}$), the random effects $\vartheta_{z,j}$ (or $\vartheta_{nz,j}$) are estimated at zero in a substantial part of the simulations (between 17% and 55% depending on the parameters).

In addition to the simulation with 48 different parameter sets, a few special cases are investigated. First, the simulations with the the first four parameter sets are repeated with population size $N = 180000$ and samples of size $n = 6000$. The sizes of the domains are three times the sizes of the domains in the population with 60000 elements. 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 $N = 60000$ and $n = 2000$. Third, the simulations with the first four parameter sets are repeated using a Bayesian approach (with independent random effects, $N = 60000$ and $n = 2000$). Here, only a single population is created, for which a simulation with 1000 runs is carried out. The frequentist approach is applied to the same 1000 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 x is added. Therefore, for the EBLUP $\mathbf{x}_{ij} = (1, x_{ij}, \bar{x}_j)^t$. For the other estimators we use $\mathbf{x}_{ij} = (1, x_{ij})^t$, 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

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

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 model based simulation, $\nu = 500$.

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

$$sd_j = \sqrt{\sum_{q=1}^{\nu} (\hat{Y}_{j,q} - \bar{Y}_j)^2 \,/\nu}, \tag{13}$$

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

$$bias_j = \sum_{q=1}^{\nu} (Y_j^{mean} - \hat{Y}_{j,q}) \,/\nu\,. \tag{14}$$

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 SR with 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 above-mentioned 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 10 populations with the same parameters, the mean of these measures over the 10 populations is computed.

3.3 Results

Table 2 shows the mean of the bias and rmse over the domains and over the 10 created populations for SR, EBLUP and ZERO-F. In the first part of the tables, where SR is compared with EBLUP and ZERO-F, the first five domains are excluded. The table shows that in all cases considered, both SAE methods are more accurate than the SR, and ZERO-F is more accurate than EBLUP. The gain in accuracy strongly depends on the properties of the population. The following points are noticed:

- SR is generally approximately design-unbiased. This is also the case here. Small non-zero 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 SR is very large in the cases with small $\sigma_{r,z}$ (odd numbers). In those cases, the rmse is often more than halved by the SAE methods. In the case of large $\sigma_{r,z}$, the rmse of the SAE methods is usually around 10% smaller than the rmse of SR.
- In some cases, the gain in accuracy of ZERO-F with respect to EBLUP in the five smallest domains is substantially larger than in the other domains. Therefore, it is important to compare 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 EBLUP is only 5% - 10%.
- Substantially larger gains with ZERO-F instead of EBLUP are possible in the case of large $\sigma_{r,nz}$, small $\sigma_{r,z}$ and a small residual variance $\sigma^2_{e,nz}$, especially if the non-zero-fraction is around 0.5 or 0.85 (number 3, 11, 19, 27).
- Substantially larger gains with ZERO-F instead of EBLUP are also possible in the case of a small residual variance $\sigma^2_{e,z}$, if the non-zero-fraction is around 0.1 or 0.85 (number 17-24, 33-40). There, the gain is larger if the non-zero-fraction is around 0.1, than if it is 0.85.
- All together, the possible gain with ZERO-F instead of EBLUP depends only slightly on the non-zero-fraction.

Another way to summarize the results about the rmse is to compute the ratios $\frac{rmse_{EBLUP}}{rmse_{ZERO-F}}$ 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 paper is the comparison of EBLUP and ZERO-F, such a comparison is not done for SR and SEA 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 EBLUP in at least 65% of the domains.
- In the cases with large $\sigma_{r,z}$ (even numbers) and a non-zero fraction of around 0.5, the differences between the domains are relatively small with a 10% quantile of between 0.96 and 1.03 a 90% quantile between 1.05 and 1.2.
- Also in the cases of large residual variances $\sigma^2_{e,z}$ and $\sigma^2_{e,nz}$ and a non-zero fraction of around 0.5 (number 13 and 15), the differences between the domains are relatively small.
- For a non-zero 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 $\sigma^2_{e,z}$, non-zero fraction of around 0.85, (number 25 and 29).
- In many cases with small $\sigma_{r,z}$ (odd numbers) 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 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.

		bias	bias	bias	rmse	rmse	rmse	rmse	rmse
	Domains	6:50	6:50	6:50	6:50	6:50	6:50	1:50	1:50
No.		SR	EBLUP	ZERO-F	SR	EBLUP	ZERO	EBLUP	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
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

Table 2 Mean absolute bias and mean rmse

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

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 numbers), the gain in accuracy of ZERO-F with respect to EBLUP generally depends strongly on the size of the random effects $\vartheta_{z,j}$ with a larger gain in the domains with the smallest (most negative) and the largest random effects. This gain is caused by both a smaller bias and a smaller standard deviation of ZERO-F in these domains. In situations with around 50% non-zero target variables, this gain in accuracy is similar in the domains with the smallest and the largest random effects. In situations with around 85% non-zero target variables, this gain is larger in the domains with the largest random effects. In situations with around 10% non-zero target variables, it is the opposite. 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 $\vartheta_{nz,j}$ is visible. The gain in accuracy of ZERO-F with respect to the EBLUP does not depend 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.

To demonstrate the described dependence of the gain in accuracy on the size of the random effects $\vartheta_{z,j}$, figures for two situations are shown. Figures 1, 2, and 3 show the rmse, bias and sd for situation 4 of SR, EBLUP and ZERO-F. The figures show the results for one of the ten populations created using the same parameters. The domains are ordered by the size of the random effects $\vartheta_{z,j}$. For the first and the last few domains (i.e. domains the the smallest and largest random effects $\vartheta_{z,j}$), ZERO-F is more accurate than EBLUP, whereas the accuracy of both estimators is almost equal for the domains in between, as Figure 1 shows. Figure 2 shows that EBLUP and ZERO-F are biased in the first and the last few domains, where the bias for ZERO-F is mostly smaller than for EBLUP, especially for the last few domains. Figure 3 shows that the sd of ZERO-F is smaller than the sd of the EBLUP with the largest gain in precision in the first few domains.

The Figures 4, 5, and 6 show similar results for situation 22. Here, the rmse of ZERO-F is smaller than the rmse of EBLUP in almost all domains with the largest gain in accuracy in the last few domains (i.e. the domains the the largest random effects $\vartheta_{z,j}$). In this situation, domains with large random effects $\vartheta_{z,j}$ are domains with almost no zeros in the target variable. In these domains, the gain in precision (measured by sd, Figure 6) of ZERO-F with respect to EBLUP is large. In the domains with the smallest random effects $\vartheta_{z,j}$ the bias of the EBLUP is much larger than the bias of ZERO-F.

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 **x** is added are special in many cases. There, the rmse of EBLUP and 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. In Figure 7 this is demonstrated for situation 22. Here, domain 16 is the domain where the 0.9-quantile of the vector **x** is added.

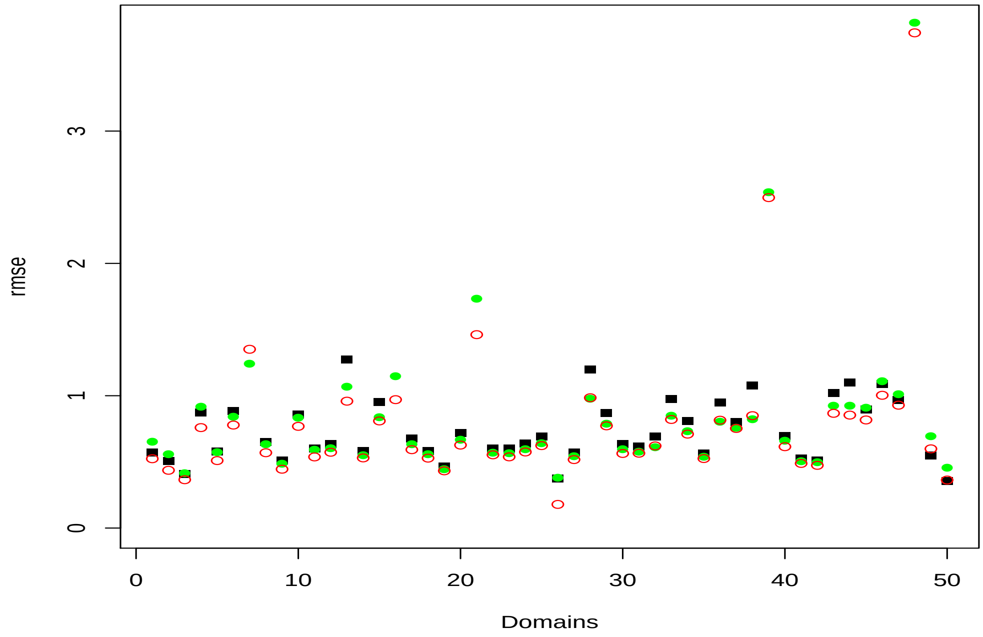

Figure 1 Rmse, situation 4, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)

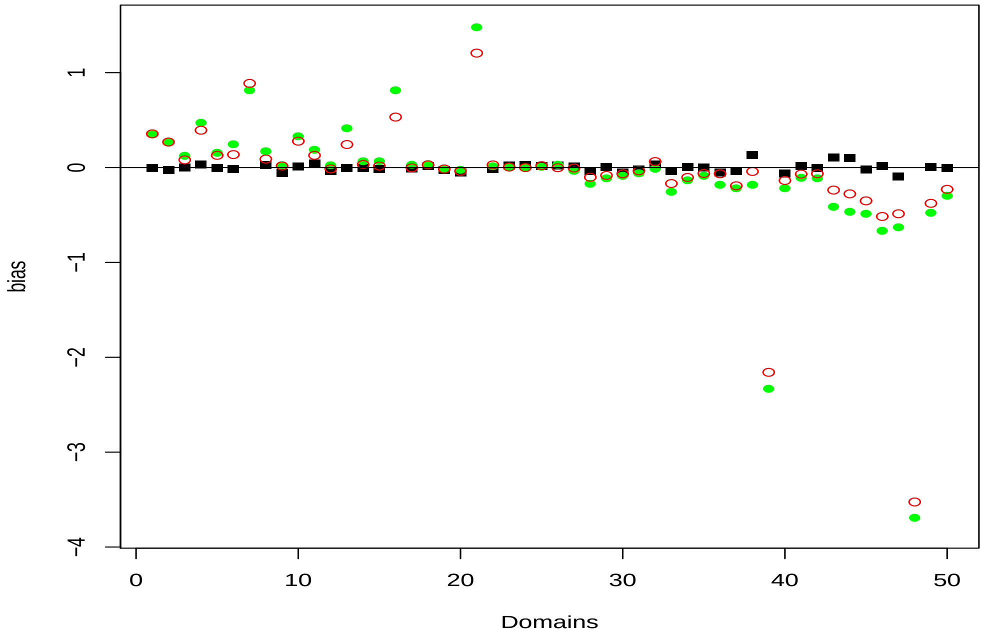

Figure 2 Bias, situation 4, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)

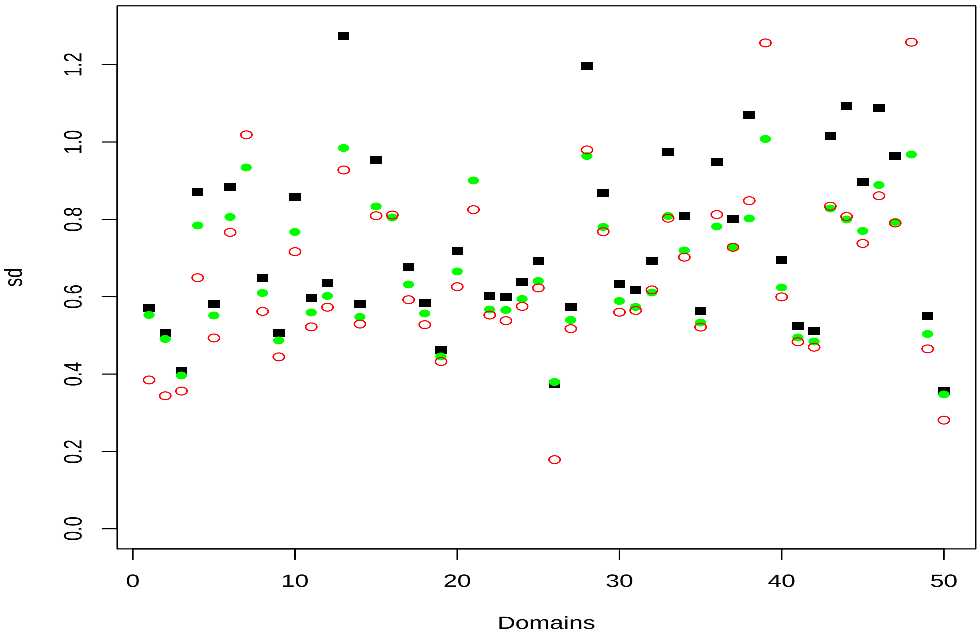

Figure 3 **Sd, situation 4, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)**

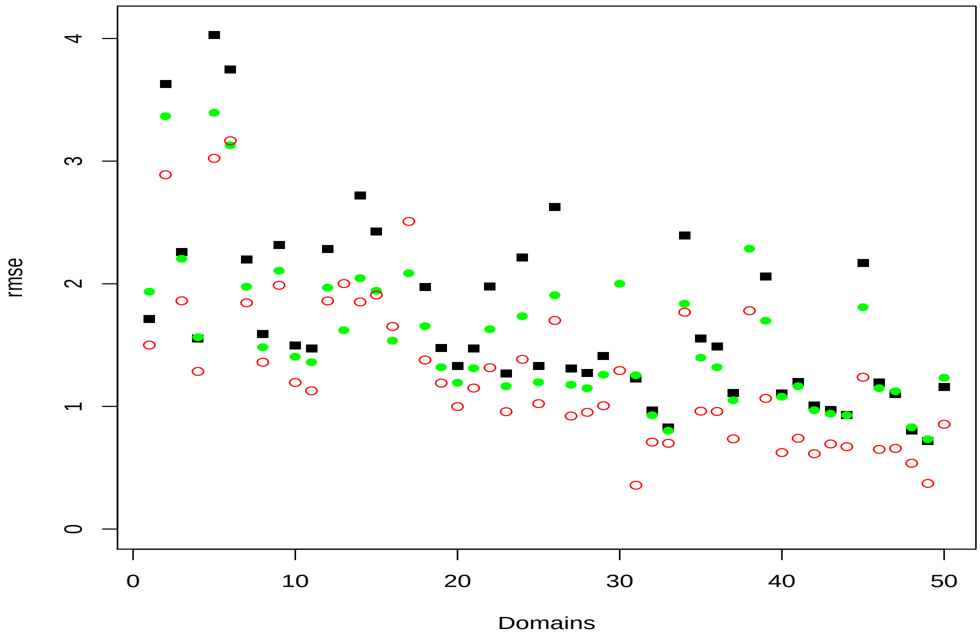

Figure 4 **Rmse, situation 22, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)**

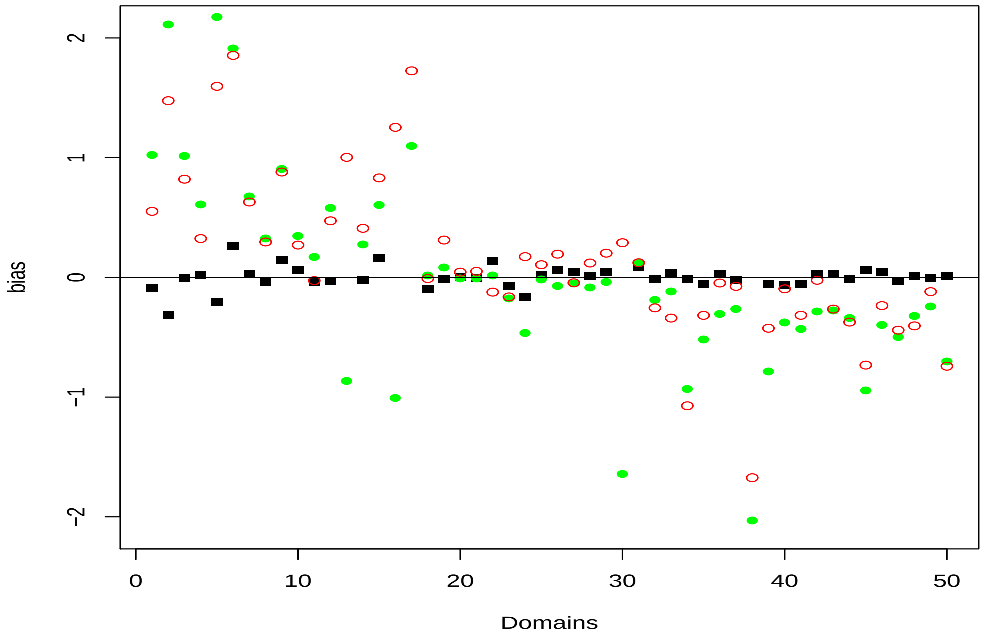

Figure 5 Bias, situation 22, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)

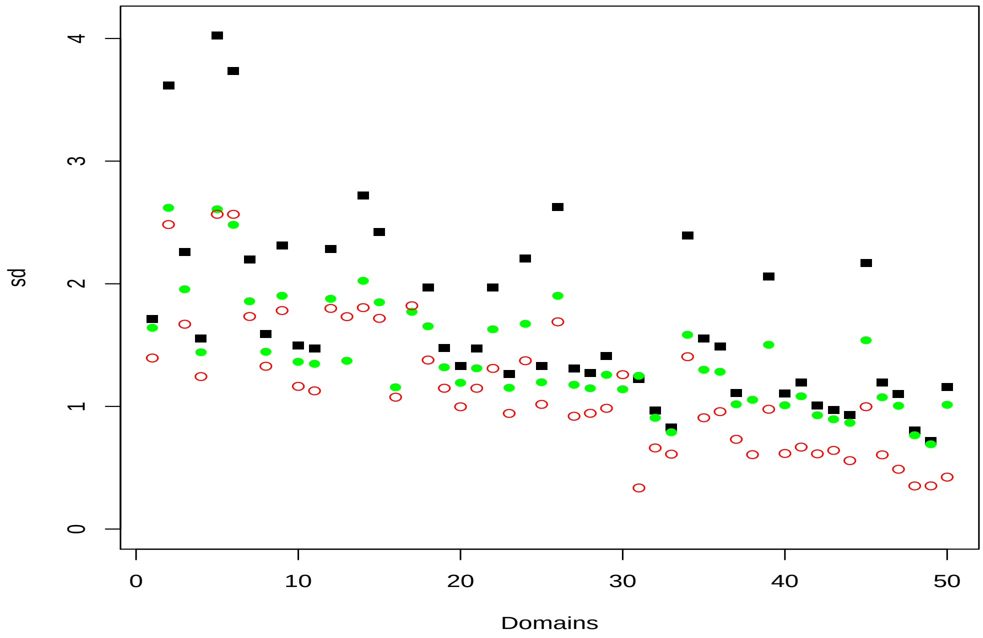

Figure 6 Sd, situation 22, domains ordered by random effect $\vartheta_{z,j}$, (black: SR, green: EBLUP, red: ZERO-F)

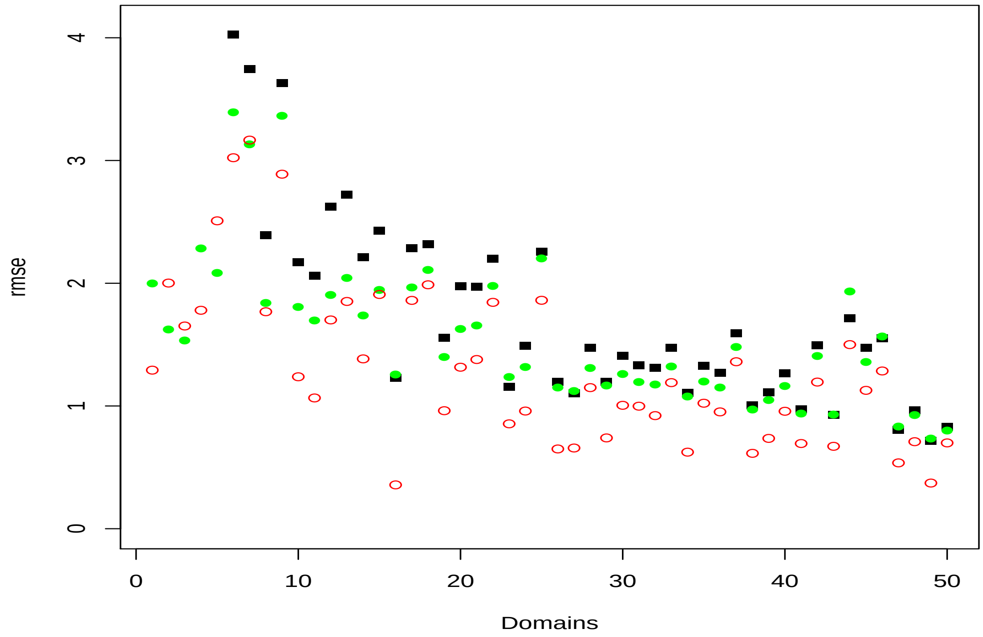

Figure 7 Rmse, situation 22, domains ordered by domain size, (black: SR, green: EBLUP, red: ZERO-F)

3.5 Results for larger populations and for correlated random effects

The simulations for the first four situations are repeated for larger populations (N=180000) and larger samples (n=6000). The results for these simulations are very similar to those for the smaller populations and samples discussed in the previous subsection and therefore not included in detail. As expected, the possible gain in accuracy by using SAE methods instead of SR is smaller when the sample size increases. The gain in accuracy of ZERO-F with respect to EBLUP is more or less equal to that with smaller sample sizes.

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 4 (for correlation 0.5) and Table 5 (for correlation 0.9). The accuracy of SR is, as expected, not affected by this correlation. The effect on the accuracy of EBLUP and ZERO-F is also small. Only for the EBLUP in situation number 3, there is 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 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.

		bias	bias	bias	rmse	rmse	rmse	rmse	rmse
	Domains	6:50	6:50	6:50	6:50	6:50	6:50	1:50	1:50
No.		SR	EBLUP	ZERO-F	SR	EBLUP	ZERO	EBLUP	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

Table 4 Mean absolute bias and mean rmse, correlation 0.5

		bias	bias	bias	rmse	rmse	rmse	rmse	rmse
	Domains	6:50	6:50	6:50	6:50	6:50	6:50	1:50	1:50
No.		SR	EBLUP	ZERO-F	SR	EBLUP	ZERO	EBLUP	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 5 Mean absolute bias and mean rmse, correlation 0.9

3.6 Results for Bayesian approach

Finally, the simulations for the first four situations are repeated with the ZERO-B. For these simulations, a single population is created for each situation, the number of runs in the simulations being 1000. The mean absolute bias, mean sd and mean rmse over the domains of the ZERO-F and ZERO-B are shown in Table 6. 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 8, 9, 10, 11 show boxplots of the mcmc-estimates for the rmse for all 1000 runs of the simulation. The simulation rmses, 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 and 4, Figure 9 and 11), the mcmc rmse-estimates track the simulation rmse very well. In those cases, the variation of the mcmc rmse-estimates 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 and 3, Figure 8 and 10), the bulk of the distribution of mcmc rmse-estimates often deviates from the simulation rmse. The mcmc rmse-estimates do not vary much over the domains in these cases. Nevertheless, the rmse-estimates are of the right magnitude and can be useful as an indication for the accuracy of the estimates.

No.	freq bias	mcmc bias	freq sd	mcmc sd	freq rmse	mcmc rmse
1	0.239	0.232	0.162	0.168	0.307	0.306
2	0.222	0.218	0.644	0.649	0.714	0.716
3	0.255	0.247	0.208	0.213	0.350	0.348
4	0.234	0.230	0.660	0.665	0.735	0.737

Table 6 Bias, sd and rmse of ZERO-F and ZERO-B, mean over the domains, for bias mean of absolute values

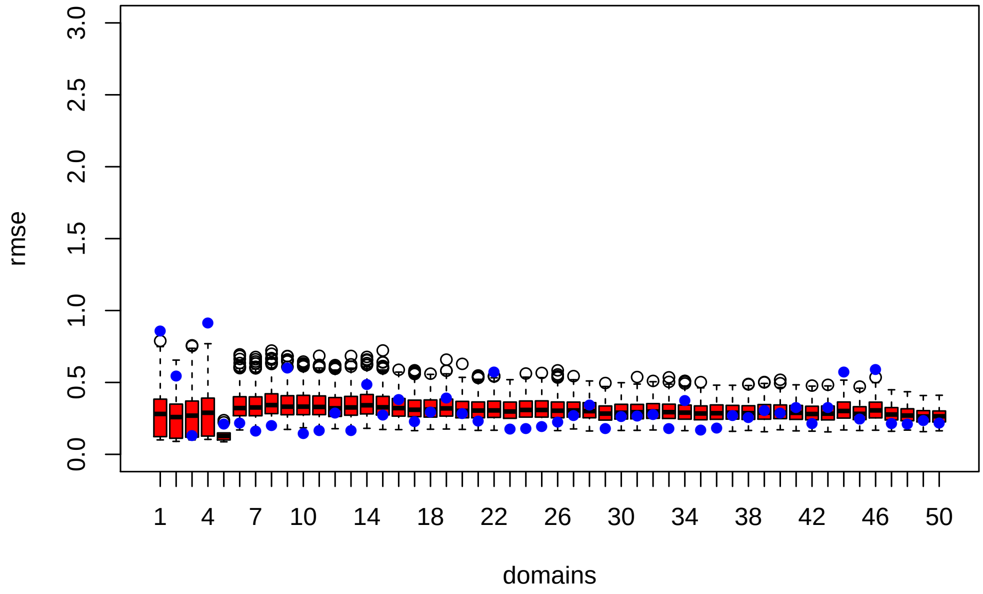

Figure 8 Boxplots of mcmc estimates for rmse of ZERO-B from 1000 simulation runs and rmse based on simulation (blue bullets), situation 1

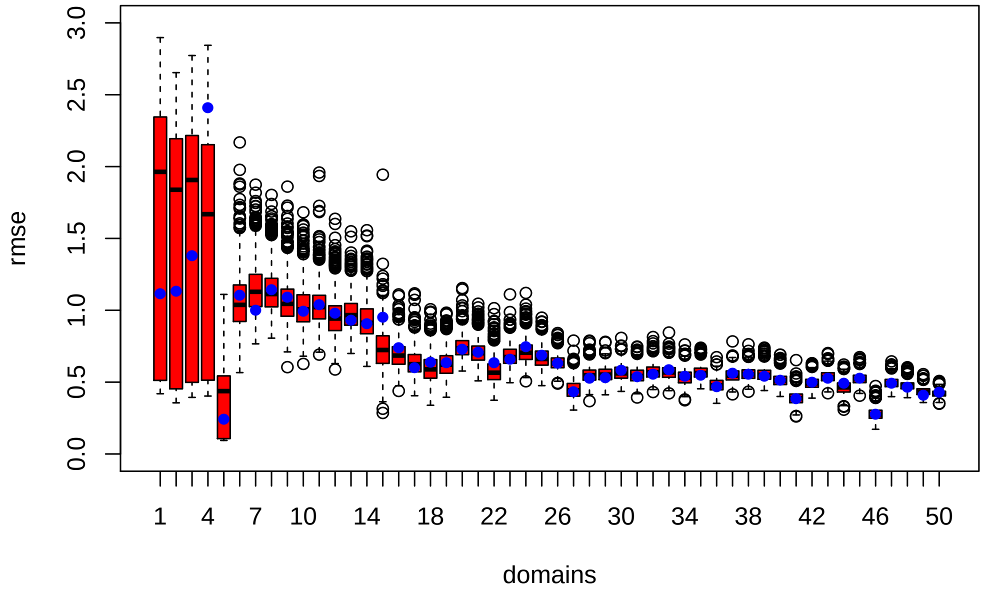

Figure 9 Boxplots of mcmc estimates for rmse of ZERO-B from 1000 simulation runs and rmse based on simulation (blue bullets), situation 2

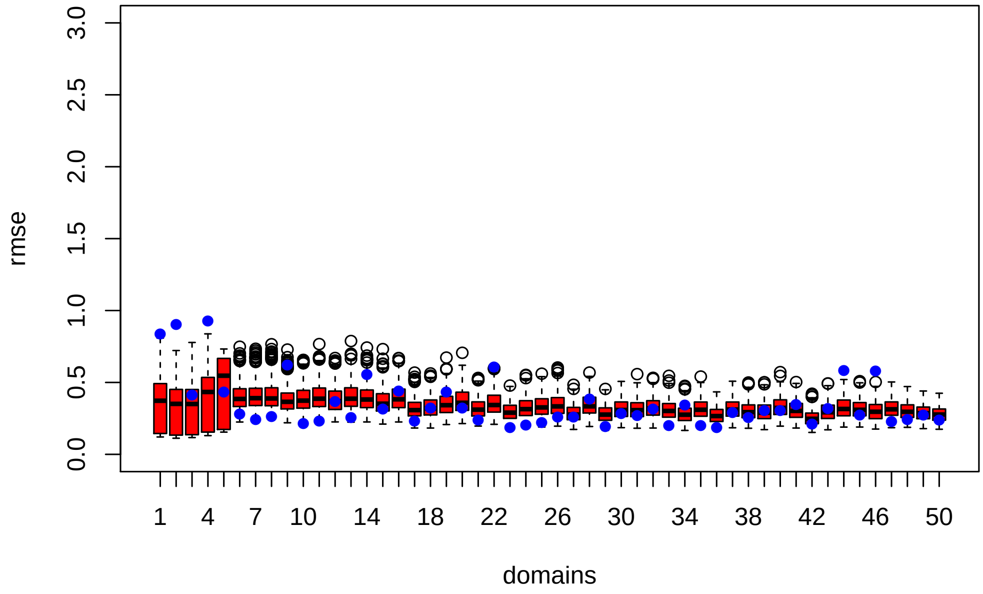

Figure 10 Boxplot of mcmc estimates for rmse of ZERO-B from 1000 simulation runs and rmse based on simulation (blue bullets), situation 3

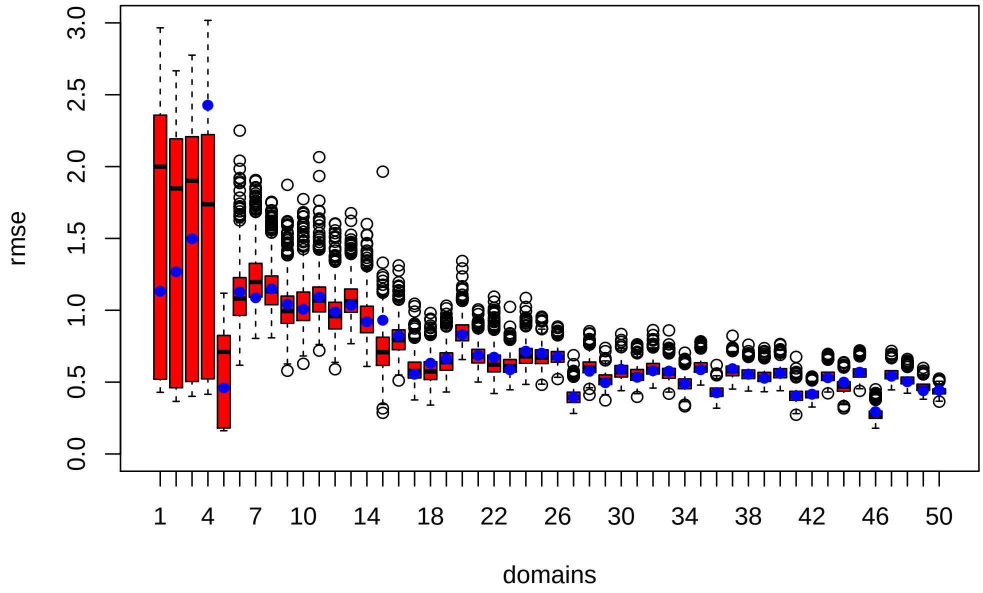

Figure 11 Boxplot of mcmc estimates for rmse of ZERO-B from 1000 simulation runs and rmse based on simulation (blue bullets), situation 4

4 Simulation with HBS data

4.1 Design of HBS

The aim of the HBS is to measure the expenditures of households, for example rent, insurance premium, expenditures for furniture, clothes, books or food. Some of these expenses are on a regular basis (often the same amount of money is paid every month for rent and insurance premium). Other expenses are quite regular, although with varying amounts of money spent. This often concerns cheaper products (food is bought almost every week). Finally, there are also expenses which are more rare, for example furniture or clothes. These products are often, but not always, relatively expensive. Therefore, the Dutch 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 euro.

The HBS has been redesigned repeatedly with the aim to increase response rates and decrease costs. Since 2012, the diary for large expenditures is 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, with 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 7 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 paper, we consider a classification in $m = 11$ types of households. Table 7 shows these domains and their sizes in the artificial population. In the simulation, samples of size $n = 5000$ are drawn by simple random sampling without replacement. Complications caused by different response probabilities which occur in practice are avoided. In the simulation 3000 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 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 bike do not buy motor fuel.

Table 8 shows the percentages non-zero expenditures, the means of the non-zero expenditures, and the overall expenditure means for the three target variables and the

11 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.

No.	Description	Population size
1	single man, younger than 65 years	12976
2	single man, 65 years or older	2985
3	single woman, younger than 65 years	11176
4	single woman, 65 years or older	8141
5	couple, main wage earner younger than 65 years	18781
6	couple, main wage earner 65 years or older	10514
7	couple with child(ren), all children younger than 18 years	19803
8	couple with child(ren), at least one child 18 years or older	8020
9	one-parent family, all children younger than 18 years	4006
10	one-parent family, at least one child 18 years or older	2291
11	other households	1307

Table 7 Population size per type of household in artificial population of 100,000 households

	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

Table 8 Percentage non-zero expenditures, mean of non-zero expenditures and overall mean for three target variables and 11 household types

4.2 Model specification

In the simulation, the following four model specifications for the fixed effects are considered (For the definition of the auxiliary variables see Appendix I):

$$\begin{aligned}
fixed1 &= 1 + Income + Quarter4 + EconCat3 && (15)\\
fixed2 &= 1 + Income + EconCat7 + LivSit3 + Quarter4 + IncomeCat6 && (16)\\
fixed3 &= 1 + Income + EconCat7 + LivSit3 + Quarter4 + IncomeCat6 && (17)\\
&\quad +IncomeMean &&\\
fixed4 &= 1 + Income + EconCat7 + LivSit3 + Quarter4 + IncomeCat6 && (18)\\
&\quad +Income * EconCat7 + Age4 + HhSize5 + AgeChild4 + Sex2 &&
\end{aligned}$$

With these four specifications, the methods can be compared in different situations, and the influence of different specifications can be studied. The first specification is a relatively simple one, which could be used if no more auxiliary information is known. The last specification is quite complex and possibly too large given the sample size. The second specification was found as a reasonable one for some samples. By comparing the second and the third specification, the influence of the domain mean of Income can be investigated.

These predictor models are used for SR, EBLUP and ZERO-F. As mentioned before, we use the same auxiliary variable models for both models underlying the ZERO-F. ZERO-B is only applied with specification $fixed2$ because of the computation time of this approach.

4.3 Results: accuracy of the estimators using a frequentist approach

4.3.1 Clothes

We start with the results for the variable Clothes. To explore the distribution of the point estimates, boxplots are shown in Figures 12 and 13 for all domains under the three methods and four predictor models. In the figures, the point estimates are divided by the true domain means to have all results around the value 1. As expected, SR is (approximately) unbiased. Furthermore, the differences between the predictor models are very small for the SR. On the other hand, the predictor model has a larger influence for EBLUP and ZERO-F. These estimators are biased. This bias is often combined with smaller variance (in the figures visible as a smaller box). For the domains 5-7 (some of the larger domains, see Table 7 for the domain sizes) the bias is generally small as is the variance reduction with respect to SR. For most of the domains, the bias is larger for EBLUP than for ZERO-F. In addition, $fixed4$ has the largest variance, both for EBLUP and for ZERO-F.

More formally, the rmse can be computed for all domains, all methods and all predictor models. The results are shown in Figure 14. The rmse of the EBLUP is particularly large in the first two domains. In these domains, the mean expenditure is small (Table 8). Looking at the other domains, the results are mixed. The accuracy of SR is almost equal for all predictor models, as already concluded from Figure 12 and 13. This is not true for

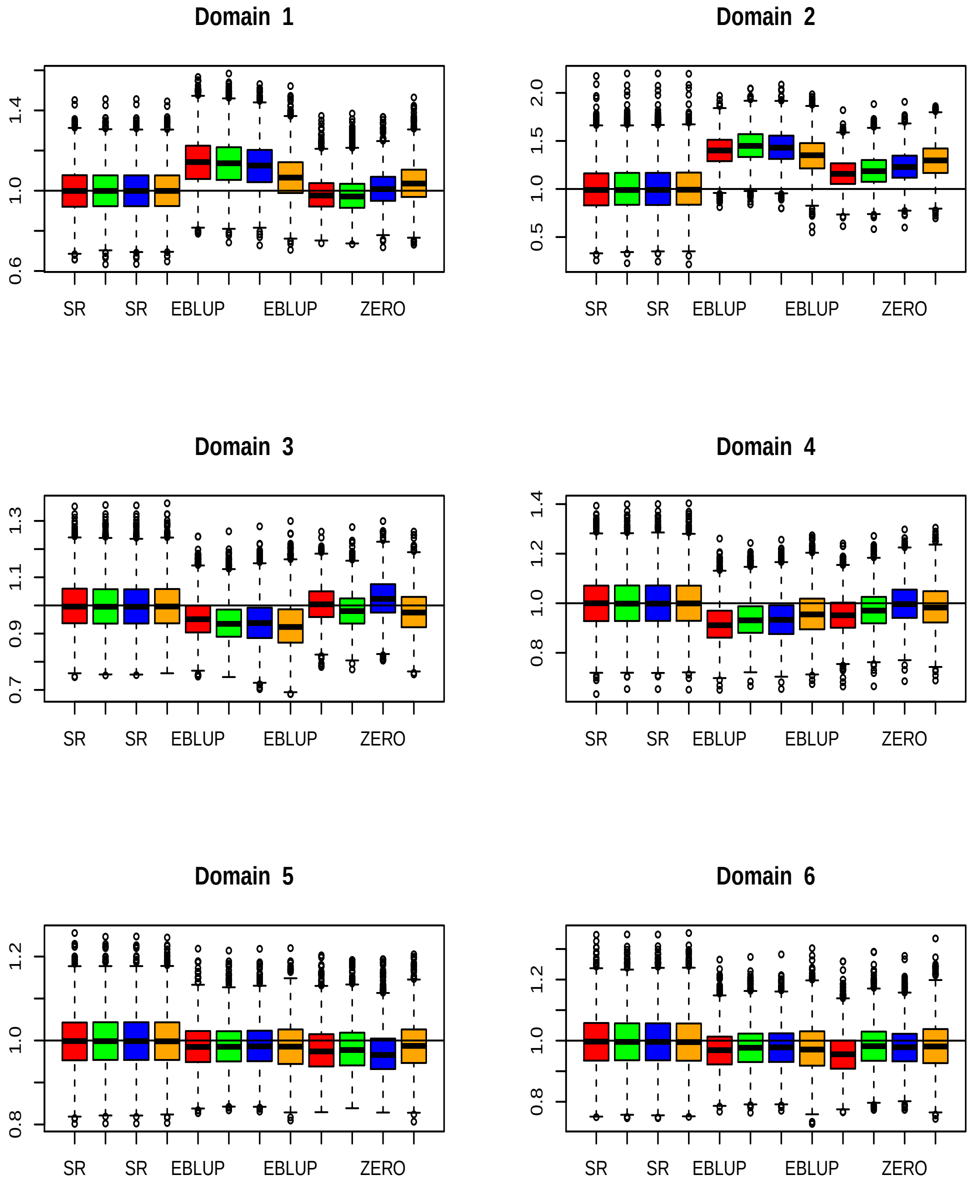

Figure 12 Boxplot of simulated point estimates (relative to population mean) for three estimators with four different fixed effects for Clothes (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 1 - 6

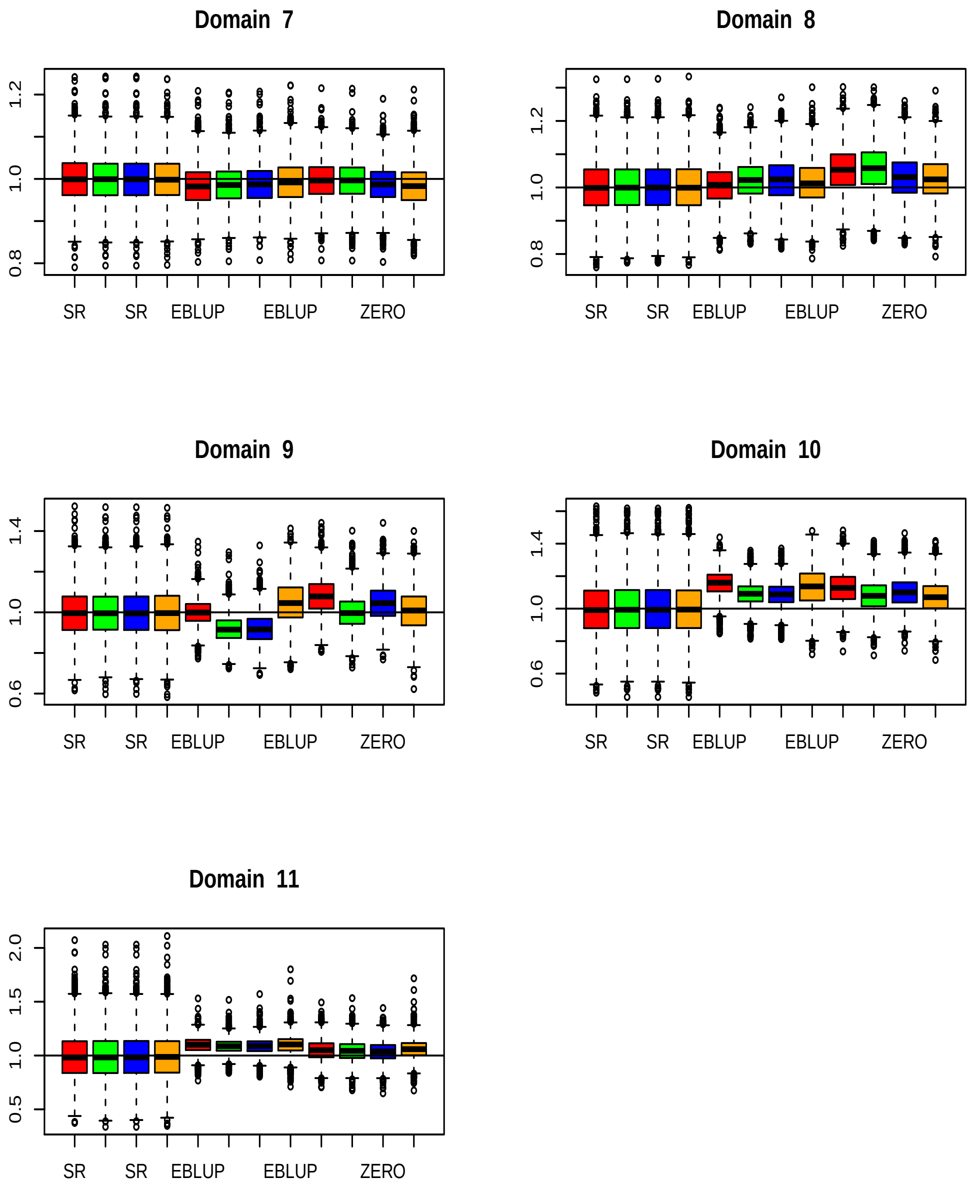

Figure 13 Boxplot of simulated point estimates for three estimators with four different fixed effects for Clothes (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 7 -11

the other methods. For example, EBLUP is much more accurate under $fixed4$ than under the other predictor models for domain 1, whereas for domain 10, this estimator is less accurate under $fixed4$ than under $fixed2$ and $fixed3$. ZERO-F is, for example, most accurate under $fixed4$ for domain 8. Furthermore, for all predictor models there are domains where EBLUP is more accurate than ZERO-F and other domains where ZERO-F is more accurate. SR is less accurate than EBLUP and ZERO-F for most but not all domains. Especially for the domains 5-7 the gain in accuracy is quite small.

There are different ways to summarize the results over all domains. One possibility is to compute the mean of the rmse over the domains; the results are shown in Table 9, column 2 - 4. Based on this table, one can conclude that ZERO-F is substantially more accurate than EBLUP, which is more accurate than SR. The second predictor model is the most accurate one for EBLUP and ZERO-F. This means that adding the domain mean of income does not improve the accuracy in this case. Furthermore, a very large model can increase the rmse, as is well known.

Another possibility to summarize the results over all domains is to compute the mean of the relative rmse over the domains (Table 9, column 5 - 7). Also according to this measure, the second predictor model is the most accurate one for ZERO-F. For EBLUP, the results are different. According to this measure, it is less accurate than SR. This is caused by the large variation between the domains and the large rmse for domains 1 and 2, the domains with small mean expenditure (Figure 14).

From Figure 14 it follows that there is no estimator or prediction model that is best for all domains, and depending on which domains are considered more important, a different estimator is optimal. ZERO-F is the best one if the mean rmse and the mean relative rmse are considered. Furthermore, this estimator is more accurate than the other estimators in more than half of the domains.

		rmse			relative rmse	
Fixed	SR	EBLUP	ZERO-F	SR	EBLUP	ZERO-F
1	6.321	5.671	5.177	0.125	0.129	0.104
2	6.294	5.570	4.888	0.125	0.130	0.099
3	6.294	5.661	4.918	0.125	0.130	0.103
4	6.286	5.971	5.219	0.124	0.131	0.112

Table 9 Mean rmse and mean relative for four choices of predictor models and three estimation methods, variable Clothes

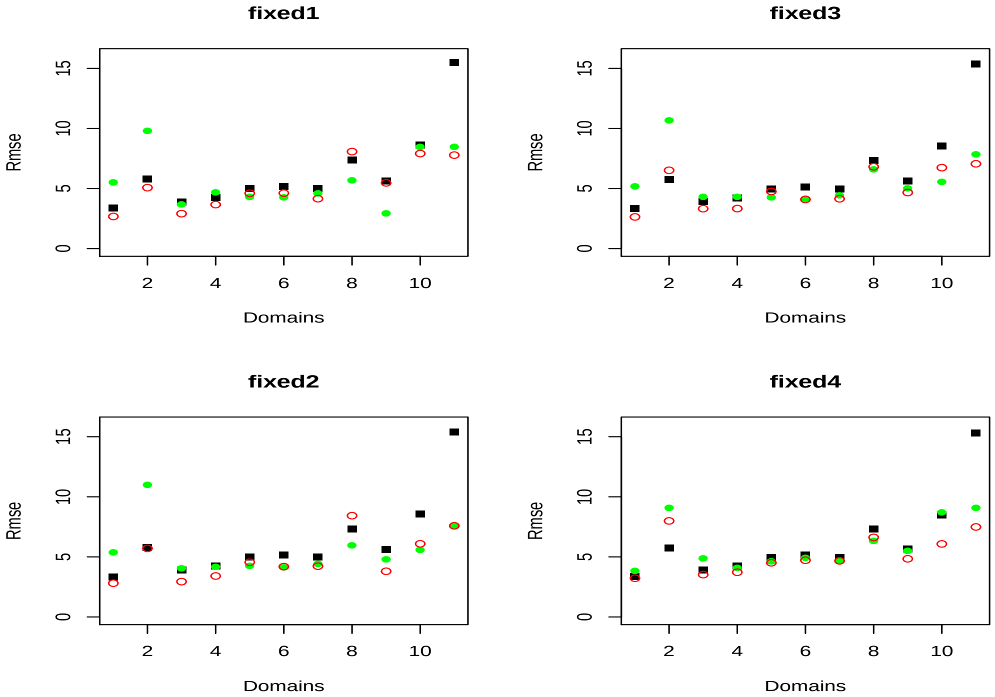

Figure 14 Root mse of three estimators with four different fixed effects for Clothes (black: SR, green: EBLUP, red: ZERO-F)

4.3.2 Men's clothes

In Figure 15 and 16 the boxplots of the estimates for Men's Clothes are shown. The general pattern is similar as for Clothes with almost no bias and very small differences between the model specifications for SR, relatively small differences between the methods and relatively small bias of the SAE methods in domains 5-7 and a large bias of the EBLUP in some domains. Here, bias and variance of EBLUP are especially large for domain 3 and 4, domains with almost no purchase of Men's clothes (compare Table 8). The variances under predictor model $fixed4$ seem less inflated for Men's Clothes than for Clothes.

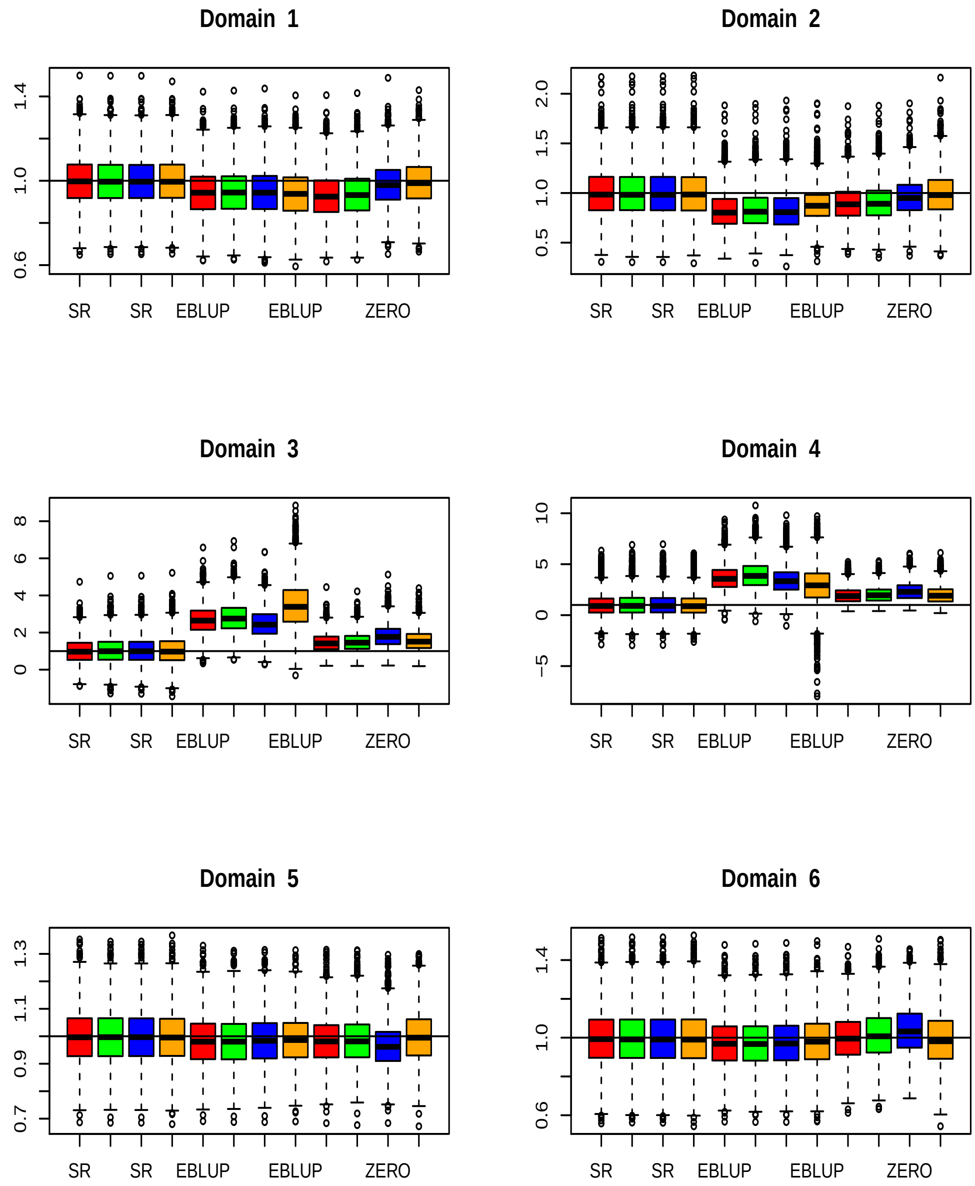

Figure 15 Boxplot of simulated point estimates for three estimators with four different fixed effects for Men's Clothes (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 1 - 6

In Table 10 the mean and relative rmse are shown for Men's Clothes. The rmse for all

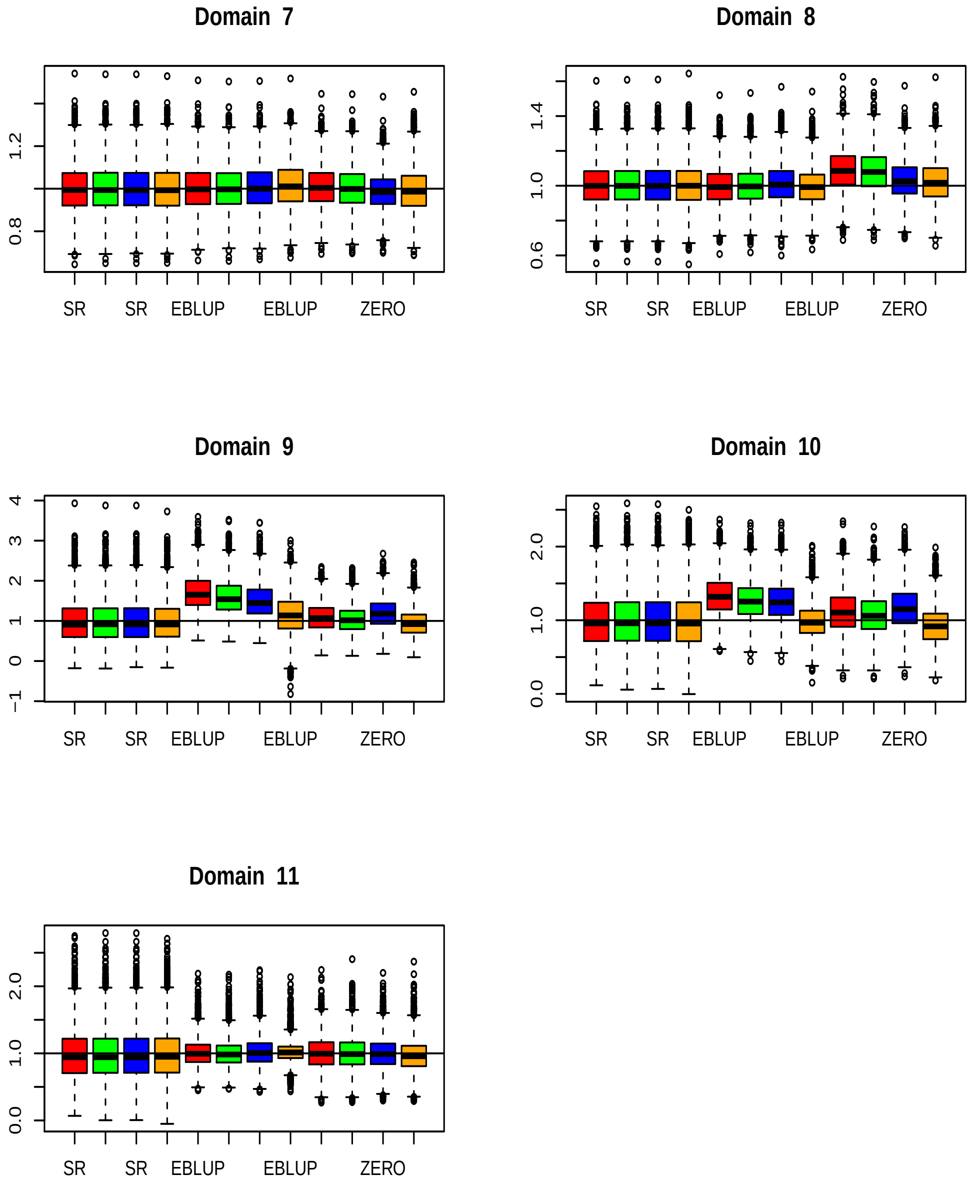

Figure 16 Boxplot of simulated point estimates for three estimators with four different fixed effects for Men's Clothes (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 7 -11

domains is shown in Figure 17. The two domains (single women, younger than 65 and at least 65, no 3 and 4) with very small expenditures (Table 8) have very large relative rmse, which strongly influences the mean relative rmse. Therefore, the mean relative rmse is also computed with these domains excluded. Nevertheless, it is an important result that ZERO-F can estimate these two domains much more accurately than EBLUP, and that EBLUP is strongly biased for these domains. The mean relative rmse with all domains included indeed shows some unexpected results, with ZERO-F being less accurate than SR for $fixed3$, which is caused by these two domains (compare Figure 17). The large difference between SR and EBLUP is also caused by these domains. For this target variable, the results are again mixed over the domains. Nevertheless, it can be concluded that ZERO-F is more accurate than the other two estimators in a large part of the domains. Particularly, EBLUP has a larger bias than ZERO-F, causing the lesser accuracy of EBLUP. For this target variable, $fixed4$ is the best predictor model for EBLUP. For ZERO-F, the differences between the predictor models are smaller.

		rmse			rel. rmse			rel. rmse	
Fixed	SR	EBLUP	ZERO-F	SR	EBLUP	ZERO-F	SR	EBLUP	ZERO-F
1	3.533	3.379	3.012	0.363	0.649	0.329	0.240	0.256	0.191
2	3.530	3.295	2.991	0.372	0.675	0.337	0.239	0.238	0.186
3	3.529	3.312	2.858	0.370	0.598	0.397	0.239	0.234	0.191
4	3.529	2.939	2.887	0.370	0.659	0.345	0.238	0.189	0.182

Table 10 Mean rmse (first three columns) mean relative rmse with all domains included (column 5-7) and mean relative rmse with domains 3 and 4 excluded (last three columns) for four choices of fixed effects and three methods, variable Men's Clothes

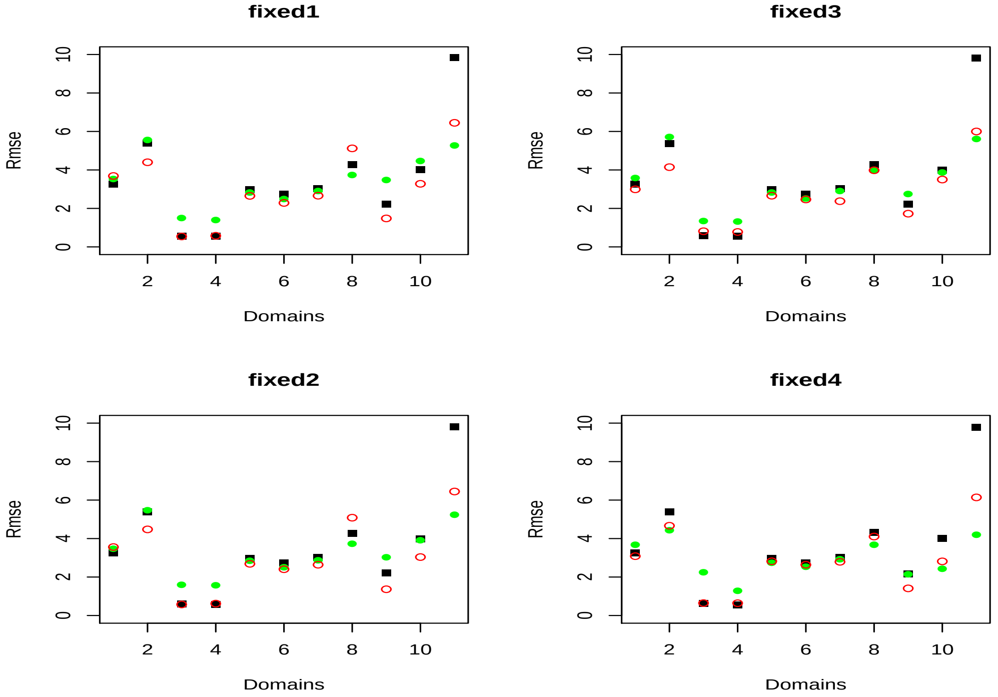

Figure 17 Root mse of three estimators with four different fixed effects for Men's Clothes (black: SR, green: EBLUP, red: ZERO-F)

4.3.3 Motor Fuel

The results for Motor Fuel are shown in Figures 18, 19, Table 11 and Figure 20. The general pattern of the boxplots are similar as for the other target variables. However, adding the domain mean of income as auxiliary variable ($fixed3$) has a large effect on the estimates in this case. It increases the bias substantially, for example in domain 1, 2, 9, 10, but also reduces the variance. This makes this predictor model worth considering for this target variable. According to the measures mean rmse and mean relative rmse the ZERO-F is again the most accurate estimator, except for $fixed4$, where the mean relative rmse of SR is smaller. For this model specification, the bias for domains 2 and 10 is large for EBLUP and ZERO-F. For this target variable, the three estimators are more or less equally accurate for domains 1, 3, 5-8, and the gain in accuracy of ZERO-F with respect to the other estimators is realized in a small part of the domains.

	rmse			rel. rmse		
Fixed	SR	EBLUP	ZERO-F	SR	EBLUP	ZERO-F
1	3.896	3.626	3.412	0.0748	0.0726	0.0700
2	3.834	3.432	3.273	0.0740	0.0716	0.0697
3	3.823	3.283	3.186	0.0736	0.0731	0.0692
4	3.802	3.685	3.524	0.0733	0.0768	0.0741

Table 11 Mean rmse and mean relative rmse for four choices of fixed effects and three methods, variable Motor Fuel

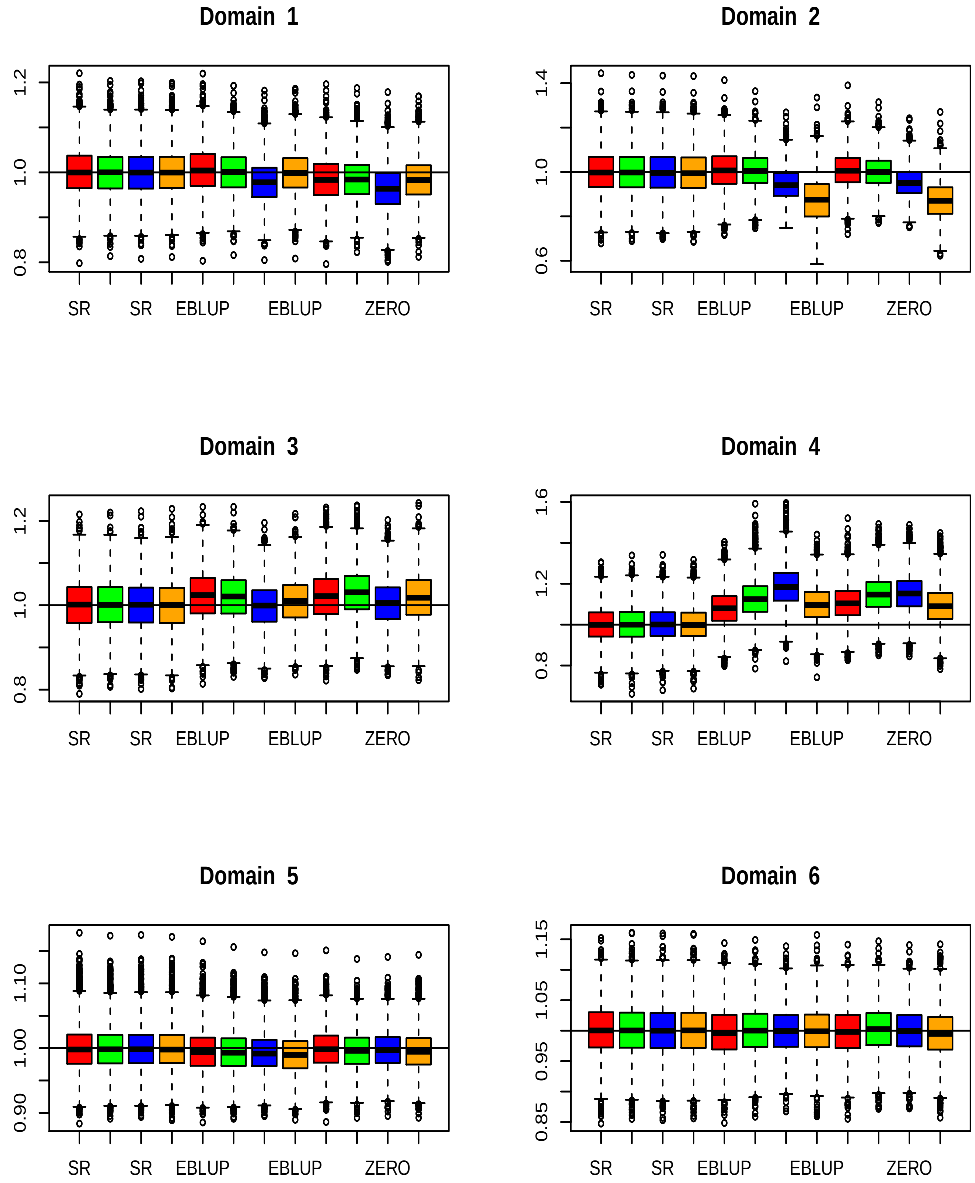

Figure 18 Boxplot of simulated point estimates for three estimators with four different fixed effects for Motor Fuel (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 1 - 6

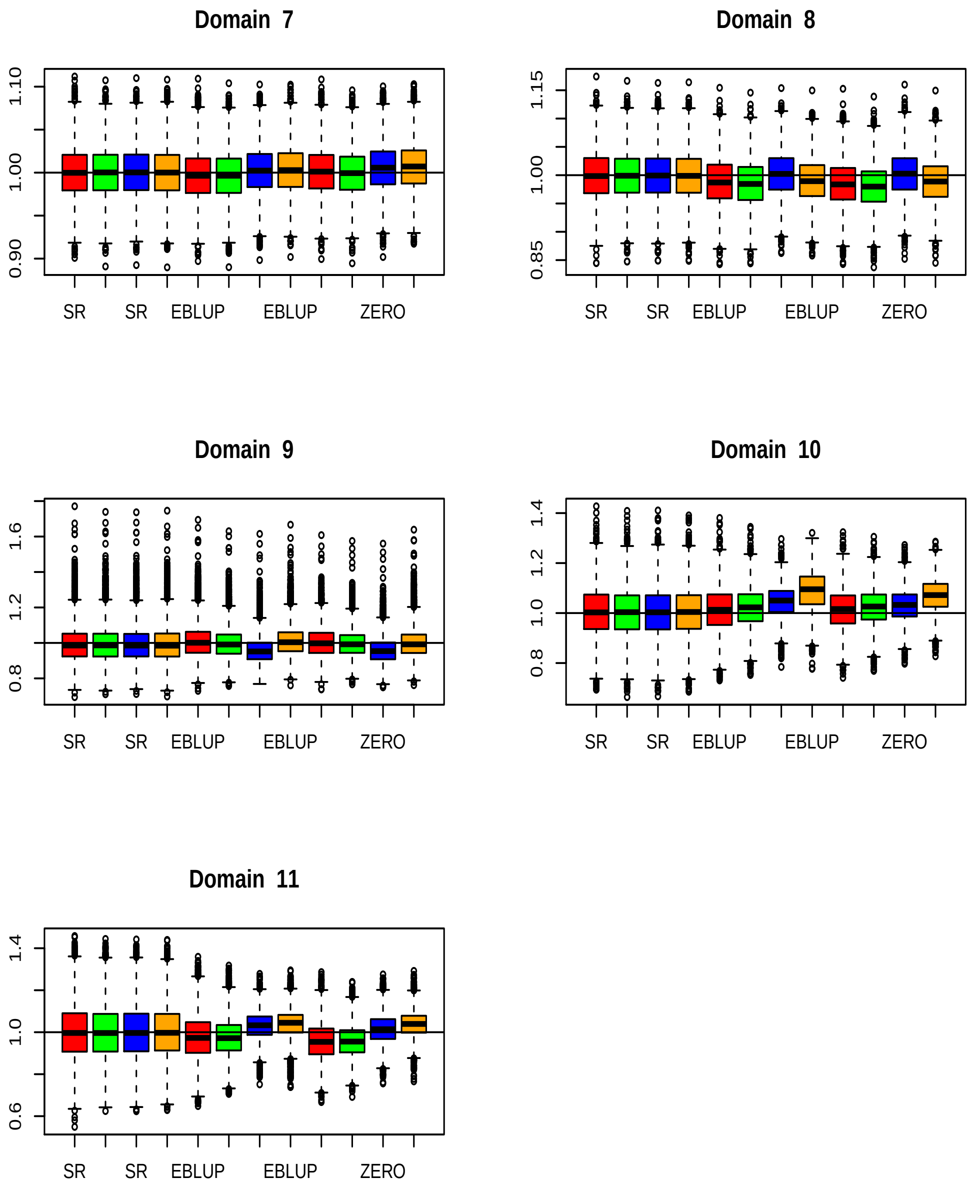

Figure 19 Boxplot of simulated point estimates for three estimators with four different fixed effects for Motor Fuel (red: $fixed1$, green: $fixed2$, blue: $fixed3$, orange: $fixed4$), domains 7 -11

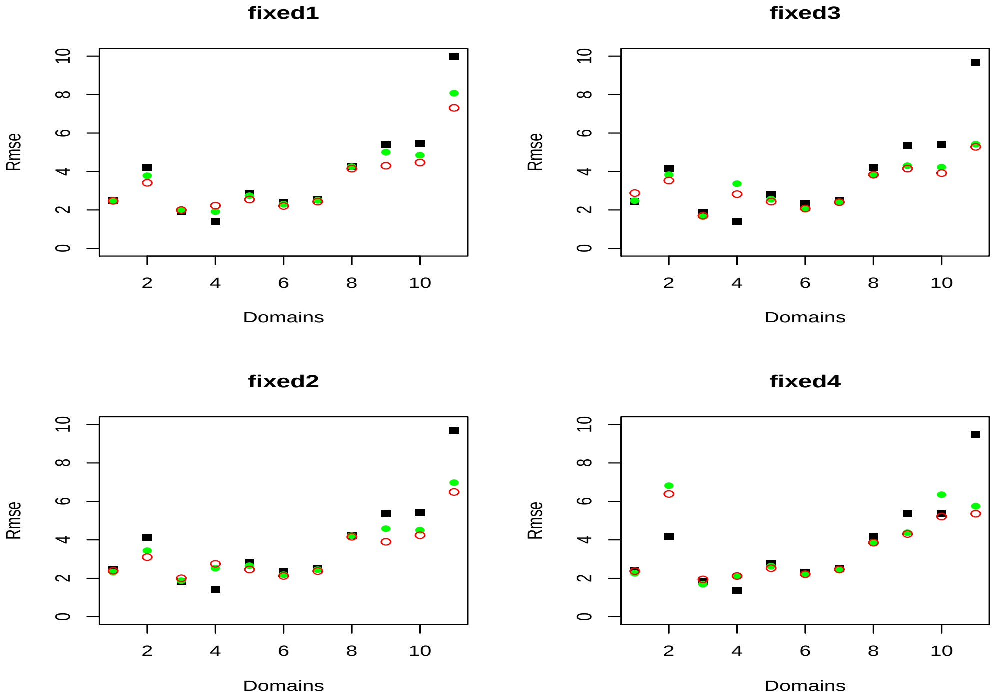

Figure 20 Root mse of three estimators with four different fixed effects for Motor Fuel (black: SR, green: EBLUP, red: ZERO-F)

4.4 Results with Bayesian approach

ZERO-B is only applied in a simulation with the second predictor model (16). In the simulation 3000 samples of size $n = 5000$ are drawn from the population. The accuracy (measured as mean of the rmse over the domains) of ZERO-B is very similar to the accuracy of ZERO-F, as Table 12 shows. The mean absolute bias of ZERO-B is larger than the bias of ZERO-F, whereas the standard deviation is smaller for Clothes and Men's clothes. This is different to the results found in the model based simulation (Table 6). All together, there is no preference for one of the approaches if the goal is restricted to computation of point estimates. The main advantage of the Bayesian approach in this application is that it provides information about the accuracy of the estimates. Figures 21, 22 and 23 compare the mcmc-estimates of the rmse with the rmse based on the simulation. The mcmc-estimates overestimate the (true) simulation rmse for some domains and underestimate it for other domains. Nevertheless the mcmc-estimates of the rmse are useful information about the accuracy of the estimates for the domains as they are in the right magnitude.

	bias		sd		rmse	
	freq	mcmc	freq	mcmc	freq	mcmc
Clothes	1.97	2.11	4.29	4.25	4.89	4.89
Men's clothes	0.84	1.10	2.78	2.64	2.99	2.98
Motor fuel	0.97	0.97	3.00	3.03	3.27	3.26

Table 12 Mean rmse, sd and bias of ZERO-F and ZERO-B, three variables

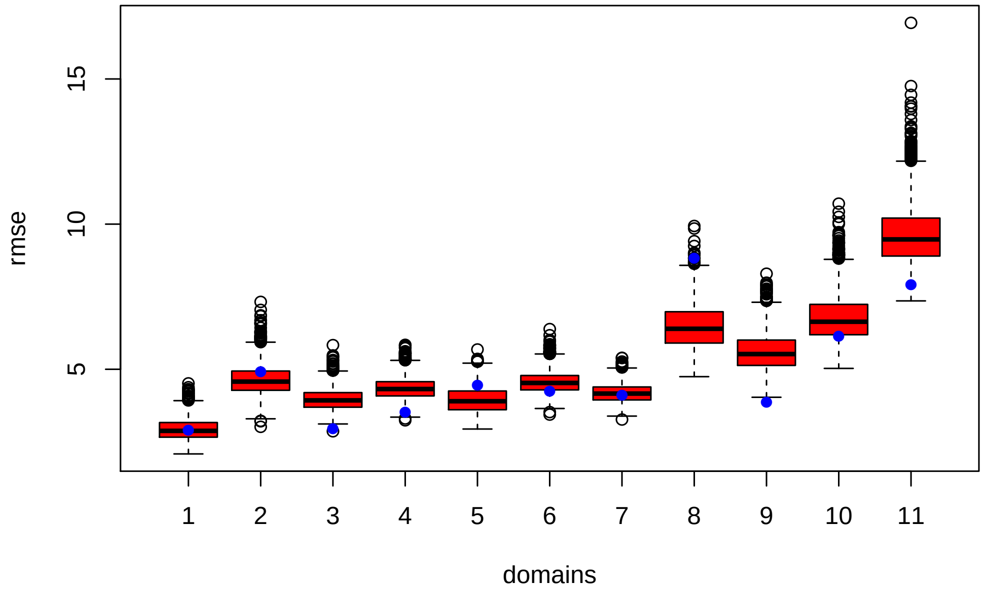

Figure 21 Boxplot of mcmc estimates for root mse and rmse based on simulation (blue bullets), Clothes

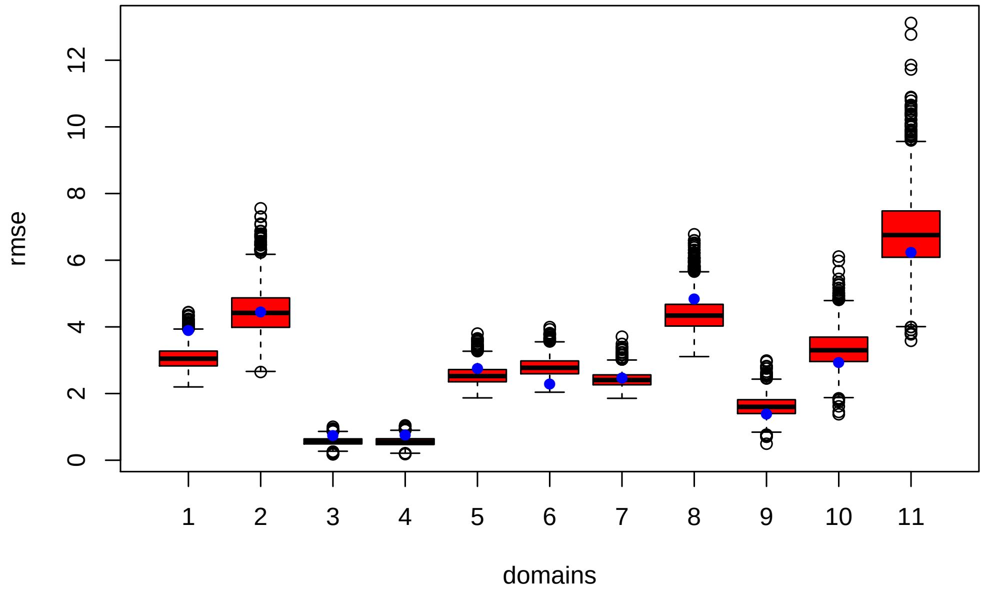

Figure 22 **Boxplot of mcmc estimates for root mse and rmse based on simulation (blue bullets), Men's Clothes**

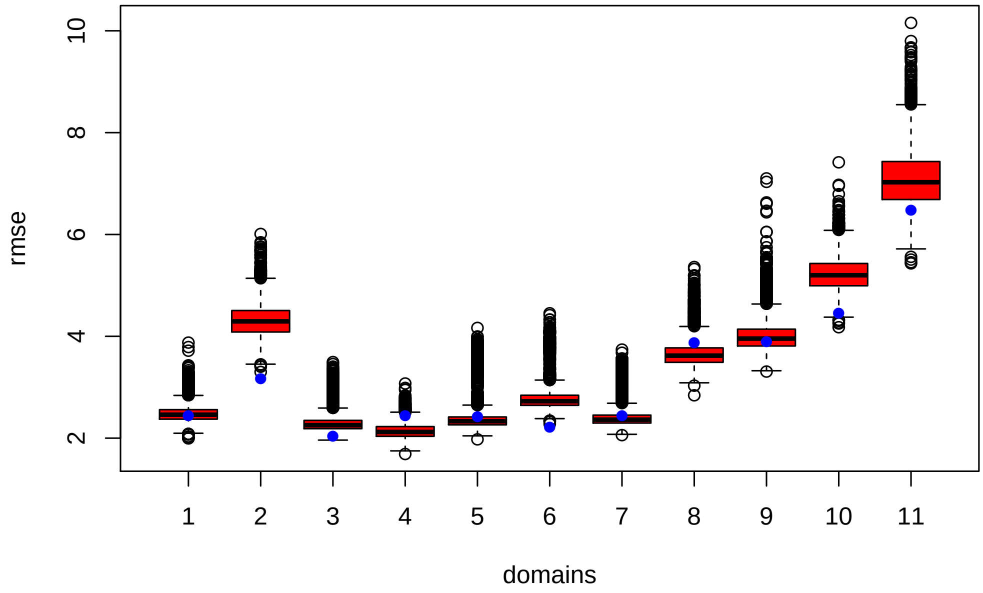

Figure 23 **Boxplot of mcmc estimates for root mse and rmse based on simulation (blue bullets), Motor fuel**

4.5 Comparison with results of model-based simulation

The general result is that for zero-inflated data ZERO is more accurate than EBLUP or SR, both in the model-based simulation and in the simulation with HBS-data. The amount of gain strongly depends on properties of the population, with sometimes only a very small improvement. The amount of improvement in the HBS-simulation is in the same magnitude as in many model-based simulations. Furthermore, the amount of gain varies over the domains, and there are domains where ZERO is less accurate than the other estimators. Domains where SR is more accurate than ZERO are very rare in the model-based simulation, and there, the differences are small. In the HBS-simulation, there are some domains where SR is substantially more accurate than ZERO. Whereas in the model-based simulation we can be sure that the model for ZERO is well specified, model-misspecification is possible also for ZERO in the HBS-simulation. One model assumption is that the random effects are normally distributed. Presumably, this assumption is not met. It is remarkable that the domains where SR is more accurate than ZERO have typically the largest fractions of zeros in the data. For such domains a random effects distribution with wider tails might work better. The model-based simulation shows that ZERO is more accurate than EBLUP especially in domains with large random effects $\vartheta_{z,j}$ (depending on other properties of the population). Also in the HBS-simulation, there are some domains with a large gain in accuracy of ZERO with respect to EBLUP. These are domains with large random effects.

The Bayesian and the frequentist approaches are equally accurate in both applications. The Bayesian approach is especially useful for the estimation of the mse and other measures of accuracy. Both in the model-based simulation and in the HBS-simulation, the mse-estimates perform well in most of the domains, and provide a reasonable indication of the accuracy for all domains.

5 Conclusion

The model-based small area estimation methods can be considered as an alternative to the approximately design-unbiased estimation methods if the sample size is too small for reliable design-based estimates. In many surveys of national statistical institutes, there are zero-inflated target variable. Therefore, three SAE methods are compared with each other and with a design-based estimation method in a simulation study using such a target variable. The first SAE method is the EBLUP (Rao, 2003), which is the most common SAE method but it ignores zero-inflation. The second and third SAE method, developed by Pfeffermann et al. (2008) and Chandra and Sud (2012), take the zero-inflation 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 shortly called ZERO. The general conclusion is that with all SAE estimators an improvement of the accuracy can be achieved 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 EBLUP. In a model-based simulation, the properties of the population can be controlled. There, we can be sure that ZERO is not model-misspecified. The amount of improvement in accuracy of ZERO compared with EBLUP depends on the properties of the entire population and of the properties of 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.

In a design-based simulation, real data of the Household Budget Survey 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 also possible 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.

The accuracy of the point estimates of ZERO under the frequentist approach or under the Bayesian approach is almost equal, which means that the taste of the statistician can be deciding. A disadvantage of the Bayesian approach is that the computation time is higher, an advantage is that information about the accuracy of the estimates follows directly. For the frequentist approach no formula for the mean squared error is 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 mse-estimates seem to be useful as an indication of accuracy.

ZERO as used in this paper assumes a normal distribution of the non-zero part of the population. If this assumption is not met, a transformation of the data can be applied, as described in Dreassi et al. (2012) and Chandra and Chambers (2011a). In the continuation of this project, this estimator will be applied and compared with the estimators considered 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 and how they work for other surveys. An interesting example is the Structural Business Survey. This survey measures the total production of the Dutch enterprises and describes the cost-benefit-structure of different subpopulations of enterprises. In this survey, the differences in domain size are large. Furthermore, a stratified sample design is used with large differences in inclusion probability.

References

Bafumi, J. en Gelman, A. (2006). Fitting Multilevel Models When Predictors and Group Effects Correlate. Manuscript prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago.

Battese, G., Harter, R., en Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association* 83 (401), 28--36.

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

Chandra, H. en Chambers, R. (2011a). Small area estimation for skewed data in presence of zeros. *The Bulletin of Calcutta Statistical Association* ??, ???

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

Chandra, H. en Sud, U. (2012). Small area estimation for zero-inflated data. *Communications in Statistics - Simulation and Computation* 41, 632--642.

Dreassi, E., Petrucci, A., en Rocco, E. (2012). *Small area estimation for semicontinuos skewed georeferenced data*. Working paper 2012/05.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1 (3), 515--533.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* 33 (2), 1--22.

Neuhaus, J. en McCulloch, C. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society B* 68 (5), 859--872.

Pfeffermann, D., Terryn, B., en Moura, F. (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.

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

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

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

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

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 (314), 496--504.

Appendices

I Auxiliary information HBS

The disposable income for each household is known from the tax agency, which is the yearly income in Euro. This income can be very low or even negative in some cases, where the household probably lives from other sources of income unknown at the tax-agency, for example savings. On the other hand, there are households with a very high income. In the models, a linear relation between income and expenditures is assumes. This assumption is only reasonable for normal incomes. A pragmatic solution to fit the extreme incomes together with the normal incomes, is replacing the given disposable income by

$$Income = \begin{cases} 20000 & \text{if } disposable\ income < 6000 \\ disposable\ income & \text{if } 6000 \leq disposable\ income < 200000 \\ 200000 & \text{if } 200000 \leq disposable\ income \end{cases}$$

A disposable income of 6000 euro or less is not enough to survive. An analysis of the data shows that households with such extremely low disposable income spend more than households with a realistic low income. Therefore a higher income of 20000 is assumed for these households in the assumption that these households have some other source of income unknown at the tax agency.

IncomeCat6 is a categorial auxiliary variable derived from disposable income. The following classification is used:

- 12000 euro or less
- 12000 euro - 18000 euro
- 18000 euro - 25000 euro
- 25000 euro - 35000 euro
- 35000 euro - 50000 euro
- more than 50000 euro

Furthermore, the mean income (*IncomeMean*) over the domains is used as auxiliary information.

The classification for living situation (*LivSit3*) is

- home-owner
- tenant, with rent subsidy
- tenant, without rent subsidy

The economic category in 7 categories (*EconCat7*) is based on the economic category of the main wage earner, the classification is as follows:

- employee, including public servant
- managing director
- self-employed or or other working person
- benefit, for example social benefit and unemployment benefit
- retirement
- student
- no personal income

The economic category in 3 categories (*EconCat7*) is the following contraction of the economic category in 7 categories

- managing director or no personal income
- working, including students
- non-working

This contraction is based both on substantive considerations and an analysis of the spending patterns.

Quarter4 is of course the classification in the four quarters of the year. *Sex2* and *Age4* are based on the properties of the main wage earner, the classification of age is

- 30 years or younger
- 31 - 50 years
- 51 - 64 years
- 65 years or older

AgeChild4 is the age class of the youngest child with the following classification

- 5 years or younger
- 6 - 10 years
- 11 - 17 years
- 18 years or older

Household size (*HhSize5*) is a classification in five categories with the last one 5 persons or more.