DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 523–542, http://dx.doi.org/10.2478/JOS-2018-0024

Small Area Estimation with a Lognormal Mixed Model under Informative Sampling

Thomas Zimmermann[1] *and Ralf Thomas Münnich*[2]

The demand for reliable business statistics at disaggregated levels, such as industry classes, increased considerably in recent years. Owing to small sample sizes for some of the domains, design-based methods may not provide estimates with adequate precision. Hence, model-based small area estimation techniques that increase the effective sample size by borrowing strength are needed. Business data are frequently characterised by skewed distributions, with a few large enterprises that account for the majority of the total for the variable of interest, for example turnover. Moreover, the relationship between the variable of interest and the auxiliary variables is often non-linear on the original scale. In many cases, a lognormal mixed model provides a reasonable approximation of this relationship. In this article, we extend the empirical best prediction (EBP) approach to compensate for informative sampling, by incorporating design information among the covariates via an augmented modelling approach. This gives rise to the EBP under the augmented model. We propose to select the augmenting variable based on a joint assessment of a measure of predictive accuracy and a check of the normality assumptions. Finally, we compare our approach with alternatives in a model-based simulation study under different informative sampling mechanisms.

Key words: Small area estimation; informative sampling; lognormal mixed model.

1. Motivation

Political and economic decision processes require increasingly reliable information on sub-populations and smaller regions. However, classical sample surveys, in general, can hardly consider all the different subgroups of interest during the design stage of a survey, which may lead to direct estimates of insufficient accuracy due to very small sample sizes in some areas. Under these circumstances, model-based small area estimation methods have become an effective tool to provide accurate estimates for the subpopulations. Detailed overviews are given in Rao and Molina (2015) and Pfeffermann (2013).

In recent years, a few papers on applying small area estimation methods in business statistics have been published (e.g. Hidiroglou and Smith 2005; Krieg et al. 2012; or Ferrante et al. 2016). Furthermore, within the research project BLUE—enterprise and trade statistics (BLUE-ETS), funded under the seventh framework programme of the European Commission, the development of design- and model-based methods for business

[1] Statistisches Bundesamt, Mathematical-Statistical Methods & Research Data Centre, Gustav-Stresemann-Ring 11, D-65189 Wiesbaden, Germany. Email: thomas.zimmermann@destatis.de
[2] University of Trier, Faculty IV – Economics, Economic and Social Statistics Department, Universitätsring 15, D-54286 Trier, Germany. Email: muennich@uni-trier.de
Acknowledgments: The authors thank the editor and two anonymous reviewers whose comments helped improve the readability of the article considerably.

surveys (Bernardini Papalia et al. 2013) was promoted. The successful implementation of small area estimation techniques in business statistics has to account for high concentrations of turnover or sizes of industries in many branches, which may vary considerably among regions. This leads to two problems.

On the one hand, since variables such as turnover or earnings are skewed, classical model assumptions are often violated and model-based estimates are likely to be biased. Transformations of these non-linearities to achieve linear relationships may help to remove, or at least reduce, possible biases of these estimates (Chandra and Chambers 2011). On the other hand, sampling designs are, in general, stratified with respect to industry (NACE classes – Nomenclature statistique des activités économiques dans la Communauté européenne) and likely to business size class. These scattered strata may give rise to highly unequal selection probabilities that are automatically compensated by design-based methods, but which have to be taken into account if model-based methods are to be applied. Moreover, attempts at design-optimization, such as the use of probability proportional to size methods (e.g. Tillé 2006) may have a negative impact on model-based small area estimation methods, as biases can occur if the size variable is not properly accounted for in the modelling process. This issue was illustrated by Burgard et al. 2014 in the context of business surveys. Nonetheless, even if model-based small area estimation techniques typically require ignorable sampling designs, this assumption is rarely valid in the practice of business statistics.

An intuitive explanation of a non-ignorable or informative sampling mechanism is that the model which holds for the sample data, differs from the one which holds in the population (Pfeffermann and Sverchkov 2009, 455). This informativeness arises because the conditional independence assumption between the sample membership and the response variable given the covariates is not satisfied. It should be noted that the informativeness of a sample can also be a result of the response process, which is an issue that we do not cover here, as the focus of our article is informativeness due to the sampling mechanism (see the discussion in Valliant et al. 2000, Sec. 2.6.2).

The consequences of an informative sampling mechanism on model-based small area methods are severe, as biased estimates of a single model parameter such as the intercept may cause biased small area estimates. Therefore, different methods were developed to obtain model-unbiased estimates in the presence of an informative sampling mechanism. Prasad and Rao (1999) as well as You and Rao (2002) considered a design-consistent pseudo-empirical best linear unbiased prediction (EBLUP) estimator that uses survey weights to compensate for the informativeness of the sampling mechanisms within areas. Moreover, Pfeffermann and Sverchkov (2007) proposed modelling the sample weights to correct for its informativeness. Finally, Verret et al. (2015) suggest augmenting the sample model with a variable that is a function of the selection probability.

In this article, we consider two different approaches to extend the empirical best predictor (EBP) under a lognormal model due to Berg and Chandra (2014) in the case of informative sampling. Our first proposal is to consider the EBP under an augmented model. We show how to choose the augmenting variable by a measure of the predictive accuracy, and demonstrate the importance of assessing the normality assumptions of the augmented model. Further, we propose an additional extension of the EBP estimator using survey weighted estimating equations.

The article is structured as follows. In the next section, we present the small area estimation methods of interest. The third section covers the simulation study. After introducing the simulation set-up, the relevant model selection tools and diagnostics needed for finding the augmentation variable are presented. Next, the simulation results using the modelling proposed by Verret et al. 2015 are presented to show the performance of the proposed methods in contrast to the other before-mentioned approaches. The article concludes with a summary and an outlook.

2. Estimation Methods

Our aim is to predict the unknown small area means

$$\mu_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D, \tag{1}$$

where N_d indicates the population size in area d and y_{dj} denotes the value of the variable of interest of unit j within area d. We assume that the following lognormal mixed model holds for the units in the population:

$$\log(y_{dj}) = \mathbf{x}_{dj}^T\boldsymbol{\beta} + v_d + \varepsilon_{dj}, \quad d = 1, \ldots, D, \;\; j = 1, \ldots, N_d. \tag{2}$$

Model (2) is the well-known nested error regression model (Battese et al. 1988) with $\log(y_{dj})$ as the dependent variable. In (2), $\mathbf{x}_{dj}$ denotes the vector of covariates for unit j in area d and $\boldsymbol{\beta}$ refers to the vector of regression parameters. Furthermore, v_d denotes the area-specific random effect and ε_{dj} the idiosyncratic error term. We assume a joint normal distribution on the random components, which are assumed to be independent from each other, that is, $(v_d, \varepsilon_{dj}) \sim N\big(0, \mathrm{diag}\big(\sigma_v^2, \sigma_\varepsilon^2\big)\big)$. To estimate the area means (1), the sample information on y_{dj} is assumed to be available for n_d units in area $d = 1, \ldots, D$, as well as the values of $\mathbf{x}_{dj}$ for all units in the population.

The best predictor (BP), minimizing the mean squared error (MSE), can be derived under the implicit assumption that the sampling mechanism is non-informative, that is, the model for the sampled units is identical to the model which holds for the units in the population. Berg and Chandra (2014) derive a closed-form expression for the BP under Model (2) as

$$\tilde{\mu}_d^{BPLog} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \tilde{y}_{dj}^{BPLog}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{3}$$

$$\tilde{y}_{dj}^{BPLog} = \exp\left(\mathbf{x}_{dj}^T\boldsymbol{\beta} + \tilde{v}_d + 0.5\sigma_\varepsilon^2(\gamma_d/n_d + 1)\right), \tag{4}$$

where $\gamma_d = \sigma_v^2/\big(\sigma_v^2 + \sigma_\varepsilon^2/n_d\big)$, $\tilde{v}_d = \gamma_d\big(\bar{l}_d - \bar{\mathbf{x}}_d^T\boldsymbol{\beta}\big)$ with $\bar{l}_d = n_d^{-1}\sum_{j=1}^{n_d}\log(y_{dj})$ and $\bar{\mathbf{x}}_d = n_d^{-1}\sum_{j=1}^{n_d}\mathbf{x}_{dj}$.

The BP comprises two parts: Inside the brackets is the sum of the sampled units within an area plus the sum of the best predictions, given the model and the available data for the non-sampled units. However, the BP defined by (3) cannot be computed in practice, as the model parameters $\boldsymbol{\xi} = \big(\boldsymbol{\beta}^T, \sigma_v^2, \sigma_\varepsilon^2\big)^T$ are generally not known and have to be replaced by

estimates $\hat{\boldsymbol{\xi}}^T$. These estimates are obtained from fitting Model (2) to the sample data using a suitable estimation method, for example, (restricted) maximum likelihood ((RE)ML) or the method of moments. Berg and Chandra (2014) suggest using REML to fit Model (2) and obtain the empirical best predictor (EBP) as:

$$\hat{\mu}_d^{EBPLog} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d} \hat{y}_{dj}^{EBPLog}\right], \quad d = 1, \ldots, D \quad \text{where} \tag{5}$$

$$\hat{y}_{dj}^{EBPLog} = \exp\left(\mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}} + \hat{v}_d + 0.5\hat{\sigma}_\varepsilon^2(\hat{\gamma}_d/n_d + 1)\right). \tag{6}$$

The MSE of the EBP can be given as an orthogonal decomposition into two parts according to (Berg and Chandra 2014):

$$\mathrm{MSE}\left(\hat{\mu}_d^{EBPLog}\right) = \mathrm{E}\left(\tilde{\mu}_d^{BPLog} - \mu_d\right)^2 + \mathrm{E}\left(\hat{\mu}_d^{EBPLog} - \tilde{\mu}_d^{BPLog}\right)^2 = M_{1d} + M_{2d}. \tag{7}$$

Berg and Chandra (2014) derived a closed-form expression for the leading term M_{1d}, as well as a linear approximation to M_{2d}. It should be noted that the expressions for both terms depend on the unknown parameter vector $\boldsymbol{\xi}$. Replacing the unknown $\boldsymbol{\xi}$ by estimates $\hat{\boldsymbol{\xi}}$ leads to a naive estimator of the MSE, owing to a bias of the leading term. Thus, a bias-correction for M_{1d}, based on the estimated $\hat{\boldsymbol{\xi}}$, is needed. One option in this regard is due to Berg and Chandra (2014), who proposed to evaluate the leading term at a modified value of the regression intercept. Alternatively, the jackknife approach from Jiang et al. (2002) can be employed. Besides correcting the bias of the leading term, this method also provides an estimate of the M_{2d}-term. This is convenient since the linear approximation to M_{2d} is cumbersome, owing to the presence of double sums. Moreover, a parametric bootstrap approach could be used to estimate the MSE of the EBP as well.

When the assumption of a non-informative sampling mechanism does not hold, the EBP (5) may suffer from severe biases, as the model validated for the sample no longer applies to the population. To overcome this problem, we consider two approaches based on extending the lognormal mixed model and a design-consistent estimation of some of the model parameters. Our first proposal is to apply the approach from Verret et al. (2015) to the context of lognormal mixed models. The basic idea of this approach is to include the selection probabilities p_{dj}, or a suitably defined function of them, $g(p_{dj})$, among the covariates in the Model (2). This gives rise to the augmented model for the sample data defined as:

$$\log(y_{dj}) = \mathbf{x}_{dj}^T\boldsymbol{\beta}_0 + g(p_{dj})\kappa_0 + v_{0d} + \varepsilon_{0dj}, \quad d = 1, \ldots, D, \; j = 1, \ldots, n_d, \tag{8}$$

where $\boldsymbol{\beta}_0$ denotes the vector of regression parameters associated with $\mathbf{x}_{dj}$, κ_0 is the regression parameter associated with $g(p_{dj})$ and $(v_{0d}, \varepsilon_{0dj}) \sim N\left(0, \mathrm{diag}\left(\sigma_{0v}^2, \sigma_{0\varepsilon}^2\right)\right)$. If the Model (8) can be validated for the sample data, it also holds for the units in the population, as the response and the sample membership are independent, conditional on the selection probabilities (Verret et al. 2015; Skinner 1994). Thus, expressions for the BP and EBP under the augmented model are obtained by replacing $\mathbf{x}_{dj}$ with $\mathbf{x}_{dj}^* = \left(\mathbf{x}_{dj}^T, g(p_{dj})\right)^T$, $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_0^T, \kappa_0\right)^T$ and $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}^* = \left(\hat{\boldsymbol{\beta}}_0^T, \hat{\kappa}_0\right)^T$ in Expressions (3) and (5). In a similar vein, the

methods to estimate the MSE of the EBP are applicable to estimate the MSE of the EBP under the augmented model as well.

It should be noted that using the EBP under the augmented model requires knowledge of the selection probabilities for all units in the population. In some situations, the model analyst might only have access to the selection or inclusion probabilities for the sampled units. A simple alternative for such cases is to consider survey-weighted estimates of the model parameters using the design weights equal to the inverse inclusion probabilities, $w_{dj} = \pi_{dj}^{-1}$. Our proposal follows You and Rao (2002), as we first obtain unweighted estimates of σ_v^2 and σ_ε^2 and then in a second step produce an estimate of $\boldsymbol{\beta}$ as the solution from survey weighted estimating Equations (SWEE). Our estimate of $\boldsymbol{\beta}$ is then given by:

$$\hat{\boldsymbol{\beta}}^{SWEE} = \sum_{d=1}^{D}\sum_{j=1}^{n_d}(w_{dj}\mathbf{x}_{dj}(\mathbf{x}_{dj} - \hat{\gamma}_{dw}\bar{\mathbf{x}}_{dw})^T)^{-1}\left(\sum_{d=1}^{D}\sum_{j=1}^{n_d} w_{dj}(\mathbf{x}_{dj} - \hat{\gamma}_{dw}\bar{\mathbf{x}}_{dw})\bar{l}_{dw}\right), \quad (9)$$

where $\bar{l}_{dw}$ and $\bar{\mathbf{x}}_{dw}$ are Hájek-type estimators of the domain means of the logarithm of the variable of interest and the vector of covariates, respectively. A survey-weighted prediction of the random effect is then obtained via:

$$\hat{v}_{dw} = \hat{\gamma}_{dw}\left(\bar{l}_{dw} - \bar{\mathbf{x}}_{dw}^T\hat{\boldsymbol{\beta}}^{SWEE}\right), \quad (10)$$

where $\hat{\gamma}_{dw} = \hat{\sigma}_v^2/\left(\hat{\sigma}_v^2 + \hat{\sigma}_\varepsilon^2\sum_{j=1}^{n_d}\tilde{w}_{dj}^2\right)$ and $\tilde{w}_{dj} = w_{dj}/\sum_{j=1}^{n_d} w_{dj}$. The predictions for the non-sampled units are then obtained via

$$\hat{y}_{dj}^{SWEE} = \exp\left(\mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}}^{SWEE} + \hat{v}_{dw} + 0.5\hat{\sigma}_\varepsilon^2\left(\hat{\gamma}_{dw}\sum_{j=1}^{n_d}\tilde{w}_{dj}^2 + 1\right)\right), \quad (11)$$

and used in a predictor of the small area means as follows:

$$\hat{\mu}_d^{SWEE} = \frac{1}{N_d}\left[\sum_{j\in S_d} y_{dj} + \sum_{j\notin S_d}\hat{y}_{dj}^{SWEE}\right], \quad d = 1, \ldots, D. \quad (12)$$

3. Simulation Study

3.1. Simulation Set-up

We consider a model-based simulation study to evaluate the different approaches to estimate the small area means in the presence of an informative sampling mechanism. The populations are drawn according to Model (2) with one covariate, that is, $\mathbf{x}_{dj} = (\mathbf{1}, x_{dj})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. In order to study the robustness of our findings, we consider three parameter settings in accordance with Berg and Chandra, which are summarized in Table 1. The parameters were chosen by Berg and Chandra (2014), such that the first two moments of the y_{dj} resemble the number of chickens per segment in a survey of the United States Department of Agriculture in the 1960s. For each setting we generate $R = 10{,}000$ finite populations and from each population one sample is drawn. In this article, we do not consider setting 1 of Berg and Chandra (2014), since this setting did not provide further information, in contrast to settings 2 to 4 (Zimmermann 2018). Our auxiliary variable was

Table 1. Parameter specifications for our simulation study.

Setting	β_0	β_1	μ_x	σ_x	σ_v	$\sigma_v^2\sigma_\varepsilon^{-2}$
2	−1.62	0.9	3.253	1.58	0.35	0.16
3	−1.62	0.9	3.253	1.24	0.71	0.45
4	−1.62	0.9	3.253	1.24	0.46	0.15

drawn as $x_{dj} \sim N(\mu_x, \sigma_x^2)$, where the values of μ_x and σ_x^2 for a given setting can be obtained from Table 1. Moreover, the last column of Table 1 shows the ratio of the variance components. Altering this ratio, as well as the variance of the auxiliary variable, enables us to study the impact of relative magnitudes of the variance parameters on the predictors. A difference in our simulation set-up from the one by Berg and Chandra (2014) is that we draw x_{dj} once per setting and then consider it fixed for all populations in this setting. Fixing the values of the auxiliary variable in the population seems reasonable, as they are most often obtained from registers and thus can be considered fixed. We follow Verret et al. (2015) in terms of the population structure and sampling designs. Thus, we consider $D = 99$ areas, where each area comprises $N_d = 100$ elements and allow for different sample sizes within areas in the following way: five elements are sampled in areas 1 to 33, seven elements are drawn from areas 34 to 66 and nine elements are sampled from areas 67 to 99. This allows us to study the impact of varying area-specific sample sizes on the choice of the augmenting variable. Note that this sampling mechanism avoids non-sampled areas, but the area-specific sample sizes are sufficiently small.

Our sampling design is based on selection probabilities proportional to an invariant Asparouhov-type size variable (Asparouhov 2006). This sampling mechanism allows us to fine-tune the informativeness of the sampling design easily by specifying a parameter. Hence, one can compare different degrees of informativeness in a straightforward manner, which would be difficult otherwise. The size variable is given by:

$$b_{dj} = \left[1 + \exp\left(-0.5\left\{\frac{1}{\alpha}\varepsilon_{dj} + \sqrt{1 - \frac{1}{\alpha^2}}\varepsilon_{dj}^*\right\}\right)\right]^{-1}, \tag{13}$$

where $\varepsilon_{dj}^* \sim N(0, \sigma_\varepsilon^2)$, $\mathrm{Cov}\left(\varepsilon_{dj}, \varepsilon_{dj}^*\right) = 0 \ \forall \ d, j$ and $\alpha \in \{1; 1.25; 2; 1,000\}$. It should be noted that α controls the degree of informativeness associated with the size variable defined by (13). For $\alpha = 1$ the b_{dj} is solely depend on the individual error term ε_{dj}, which gives rise to highly informative samples. On the other extreme, for $\alpha = 1{,}000$, the size measure hardly depends on ε_{dj}, such that the sampling mechanism is virtually ignorable. In comparison to previous studies using the size measure (13), we decided to include a value of $\alpha = 1.25$ instead of $\alpha = 3$. The reason for doing so is that Zimmermann (2018) found highly different results when α took a value of 2 instead of 1, whereas the differences for values of 2 and 3 were much less pronounced. In addition to the invariant size variable (13), a non-invariant Asparouhov-type size variable and a Pfeffermann-Sverchkov-type size variable have also been considered in the literature (Verret et al. 2015; Pfeffermann and Sverchkov 2007). Under both approaches, the size variable, and thus the selection probabilities, depend not only on the idiosyncratic error term ε_{dj}, but also on the random

effect v_d. Since earlier simulations showed only small differences between using a non-invariant and an invariant size variable, we decided to focus on the invariant case.

The selection probabilities and the inclusion probabilities are given by $p_{dj} = b_{dj}/\sum_{j=1}^{N_d} b_{dj}$ and $\pi_{dj} = n_d p_{dj}$, respectively. Note that this results in design weights given by $w_{dj} = \pi_{dj}^{-1} = (n_d p_{dj})^{-1}$. The samples were then obtained by applying Midzuno's method, which is described in detail in Tillé (2006, Sec. 6.3.5). An advantage of this particular method for our purposes is that a fast C++ implementation is available in the R package `simFrame` (Alfons et al. 2010).

3.2. *Model Selection and Diagnostics*

The simulation results in the study of Verret et al. (2015) highlight the importance of finding a suitable variable for augmentation. This can be achieved by applying a combination of model selection techniques and model diagnostics. It should be noted that a number of tests in informativity of samples have been proposed in the literature, for both single-level and two-level models, such as (2) (e.g., Skinner 1994; Pfeffermann and Sverchkov 2007). However, in our simulation setting, we control the degree of informativeness and, therefore, our focus is to find the best choice of $g(p_{dj})$. A similar approach can be found in Verret et al. (2015), who plot the residuals from an ordinary least squares (OLS) regression of the variable of interest on the covariates $\mathbf{x}_{dj}$ against the potential choices of $g(p_{dj})$. If the scatter plot reveals a functional relationship between the OLS residuals and a choice of $g(p_{dj})$, this is taken as evidence for informative sampling. Moreover, the authors discuss that a particular choice of the augmenting variable would work well, provided that the relationship is linear and the scatter is not too wide. This is based on the argument that if a perfectly linear relationship could be found for a particular $g(p_{dj})$, then including it among the covariates would eliminate any residual variation. We show the residual plot for one sample under setting 2 with $\alpha = 1$ in Figure 1 and note a systematic relationship between any choice of the augmenting variable and the residuals. Hence, the analysis of the residuals clearly indicates informative sampling. Furthermore, non-linearities are evident for w_{dj} and p_{dj}^{-1}, while it is difficult to establish a clear ordering between $\log(p_{dj})$ and p_{dj}. Thus, we additionally compare the values of the conditional Akaike information criterion (cAIC) for the different specifications of $g(p_{dj})$ as discussed by Zimmermann (2018). The cAIC was proposed by Vaida and Blanchard (2005) for model selection among linear mixed models and is a measure of the predictive accuracy conditional on the random effects, which is of utmost importance in small area estimation. It is defined as

$$\text{cAIC} = -2 \log g(\mathbf{y}|\hat{\boldsymbol{\xi}}(\mathbf{y}), \hat{\mathbf{v}}(\mathbf{y})) + 2K, \tag{14}$$

with $\log g(\mathbf{y}|\hat{\boldsymbol{\xi}}(\mathbf{y}), \hat{\mathbf{v}}(\mathbf{y}))$ as the log-likelihood conditional on the model parameters $\hat{\boldsymbol{\xi}}$ and the predictions of the random effects $\hat{\mathbf{v}}$, which are both functions of the response vector $\mathbf{y}$ (Vaida and Blanchard 2005, 355). Moreover, K denotes a penalty term regarding the model complexity, where the precise expression of K depends on the method used to fit the model. The cAIC allows comparison of different non-nested models, where a smaller value indicates a higher predictive accuracy. In Section 1 of this article, we stated that a sampling mechanism is informative if the response vector and the sample membership are

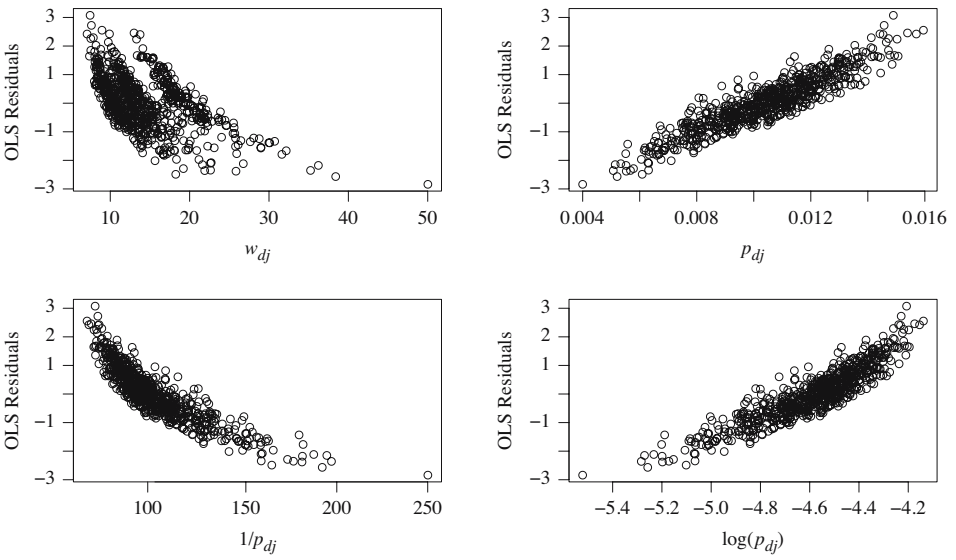

Fig. 1. Residual plots for setting 2 when $\alpha = 1$.

correlated after conditioning on the covariates. Hence, a model which includes a function of the selection probabilities should not yield a higher predictive accuracy compared to a similar model without an additional covariate under a non-informative design. Moreover, we can compare various specifications of $g(p_{dj})$ in terms of their predictive accuracy as measured by the cAIC to find the most suitable variable to augment the model.

We illustrate this procedure for one particular population generated from setting 2, where one sample is drawn using each of the values of α. Note that in the case of $\alpha = 2$, it is the same sample that has been used to generate Figure 1. The results of this comparison are displayed in Table 2. It is easily seen that the EBP under the non-augmented model, shown in the column with "-" on top, is dominated by all other specifications of $g(p_{dj})$ for all values of α except $\alpha = 1{,}000$. Thus, under informative sampling, applying an augmented model yields a higher predictive accuracy. Regarding the choice of $g(p_{dj})$, the specification $g(p_{dj}) = p_{dj}$ clearly dominates the other alternatives for $\alpha \in \{1;1.25\}$. Note that the distinct advantage in predictive accuracy of $g(p_{dj}) = p_{dj}$ is in line with the residual plot in Figure 1.

In addition to the predictive accuracy of an augmented model, the validity of the model assumptions are also critical. This issue is especially important under a lognormal mixed model, where the normality assumptions are exploited in the derivation of the BP. For this purpose, the transformed residuals may be studied to jointly assess the normality

Table 2. The conditional AIC for different choices of $g(p_{dj})$ for setting 2.

α	-	w_{dj}	p_{dj}	p_{dj}^{-1}	$\log(p_{dj})$
1	1839.20	621.32	−2629.31	193.68	−752.99
1.25	1859.77	1402.98	1213.87	1295.93	1232.22
2	1870.34	1777.23	1730.07	1742.21	1732.88
1000	1880.36	1881.91	1882.25	1882.33	1882.27

Table 3. Shapiro-Wilk test of normality of the transformed residuals for setting 2.

α		-	w_{dj}	p_{dj}	p_{dj}^{-1}	$\log(p_{dj})$
1	W	0.9985	0.8794	0.8942	0.8671	0.8829
	p-value	0.8401	0.0000	0.0000	0.0000	0.0000
1.25	W	0.9980	0.9971	0.9977	0.9983	0.9978
	p-value	0.5761	0.2634	0.4710	0.7503	0.4905
2	W	0.9983	0.9986	0.9969	0.9967	0.9965
	p-value	0.7227	0.8698	0.2035	0.1681	0.1394
1,000	W	0.9984	0.9984	0.9984	0.9983	0.9983
	p-value	0.7829	0.7867	0.7702	0.7610	0.7611

assumption on both random components of the mixed model (Battese et al. 1988). The transformed residuals are defined as

$$\hat{u}_{dj} = (\log(y_{dj}) - \hat{\tau}_d \bar{l}_d) - (\mathbf{x}_{dj} - \hat{\tau}_d \bar{\mathbf{x}}_d)^T \hat{\boldsymbol{\beta}}, \tag{15}$$

where $\hat{\tau}_d = 1 - (1 - \hat{\gamma}_d)^{1/2}$. Under the nested-error regression model (2), the transformed residuals are approximately distributed as $\hat{u}_{dj} \sim N(0, \hat{\sigma}_\varepsilon^2)$. Following Battese et al. (1988), the normality assumption can be inspected graphically via quantile-quantile (QQ) plots, or statistically by applying the Shapiro-Wilk test for normality (Shapiro and Wilk 1965).

In this article, we decided to use the Shapiro-Wilk test for the transformed residuals, whose results for the same samples that were used to compare the cAIC are presented in Table 3. Here, the test-statistic W, as well as the associated *p*-value for each value of α and for each choice of $g(p_{dj})$, are shown. A striking aspect is that the null hypothesis of normality is rejected for all candidates to augment the model in the case of $\alpha = 1$. Interestingly, the null hypothesis is not rejected for the non-augmented model with $\alpha = 1$, which highlights that augmentation, while increasing the predictive accuracy of the model and accounting for informative sampling can, in fact, violate the normality assumptions. Moreover, for any other choice of α, none of the test results lead to a rejection of the null hypothesis.

The comparison of the cAIC showed that specifying $g(p_{dj}) = p_{dj}$ yields the highest predictive accuracy among the choices considered. Furthermore, the Shapiro-Wilk tests did not indicate particular advantages of one of the choices of $g(p_{dj})$. Hence, we decided to focus on p_{dj} to augment the model in the simulation study.

3.3. Simulation Results

An overview of the different estimators analyzed in this study is given in Table 4. We performed the MSE estimation by means of the jackknife approach presented in Jiang et al. (2002).

Table 4. Estimators in our study.

Abbreviation	Description
EBP	EBP under the non-augmented model
Augmented	EBP under the augmented model
SWEE	Predictor based on survey-weighted parameter estimates

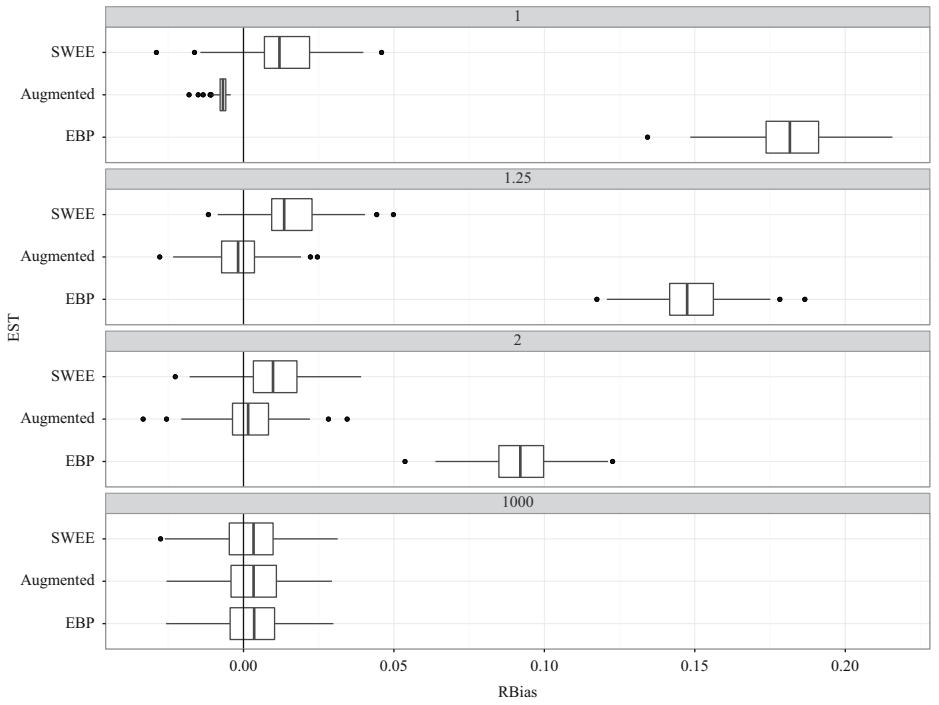

Fig. 2. Relative bias under setting 2.

To report the results of our simulation study, we consider several different quality measures related to the accuracy of point estimates and the reliability of confidence intervals for each area. A common measure to estimate the bias of a point estimator is the relative bias. It is given by

$$\mathrm{RB}(\hat{\mu}_d) = \frac{\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{r,d} - \mu_{r,d})}{\frac{1}{R}\sum_{r=1}^{R}\mu_{r,d}}, \quad d = 1, \ldots, D, \tag{16}$$

where $\hat{\mu}_{r,d}$ and $\mu_{r,d}$ denote the estimated and the true mean for area d in replication r, respectively. The relative bias takes values from $-\infty$ to ∞, while a relative bias close to zero is desirable, indicating that the point estimates are on average identical to the true values. Another quality measure is the relative root mean squared error (RRMSE), which measures the variability of the point estimates and is computed as

$$\mathrm{RRMSE}(\hat{\mu}_d) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_{r,d} - \mu_{r,d})^2}}{\frac{1}{R}\sum_{r=1}^{R}\mu_{r,d}}, \quad d = 1, \ldots, D. \tag{17}$$

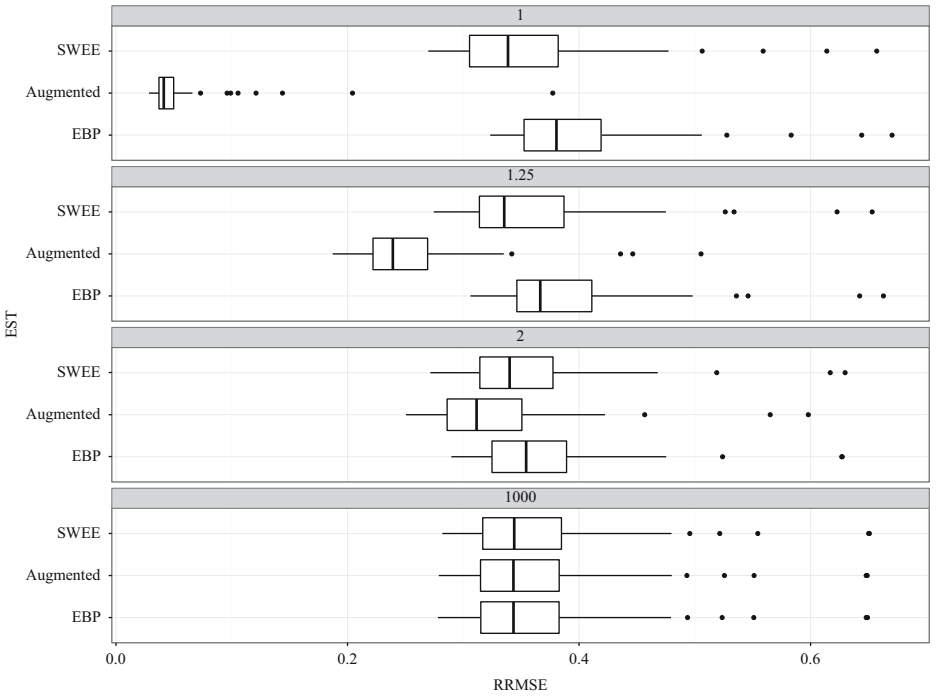

Fig. 3. RRMSE under setting 2.

Confidence intervals with a nominal coverage rate of 95% were constructed via

$$\mathrm{CI}(\hat{\mu}_{r,d})_{0.95} = \left[\hat{\mu}_{r,d} - \sqrt{\widehat{\mathrm{MSE}}(\hat{\mu}_{r,d})} \cdot t_{0.975,D}; \hat{\mu}_{r,d} + \sqrt{\widehat{\mathrm{MSE}}(\hat{\mu}_{r,d})} \cdot t_{0.975,D}\right], \quad d = 1, \ldots, D, \tag{18}$$

with $t_{0.975,D}$ as the 97.5%-quantile of the t-distribution with D degrees of freedom. A useful summary to evaluate the quality of the confidence intervals is the fraction of intervals that cover the true area mean. This quantity is called the average coverage rate, evaluated at the confidence level of 95% and given by

$$\mathrm{ACR}(\hat{\mu}_d)_{0.95} = \frac{1}{R}\sum_{r=1}^{R} I\left(\mu_{r,d} \in \mathrm{CI}(\hat{\mu}_{r,d})_{0.95}\right), \quad d = 1, \ldots, D, \tag{19}$$

where $I(A)$ denotes the indicator function, that is, $I(A) = 1$ if condition A is met and $I(A) = 0$ otherwise.

As we obtained similar results under the different settings, we focus on setting 2 in the following and report the results under settings 3 and 4 in the appendix.

The relative biases of the different domain estimates for the selected values of α are shown in Figure 2. Here, the numbers on top of each panel correspond to the values of α such that the degree of informativeness decreases from the uppermost to the lowermost panel. We note that the EBP without augmentation is severely and systematically upwardly biased for $\alpha = 1$. As α increases and thus, the level of informativeness

decreases, the biases of this non-augmented EBP are less severe. With regard to the SWEE predictor, a tendency to overestimate the area means can be seen for highly informative sampling mechanisms. Furthermore, the relative biases of the SWEE predictor are much smaller as compared to the EBP without augmentation. The best results in terms of the relative biases are achieved for the EBP under the augmented model for informative sampling mechanisms. Nonetheless, small but systematic negative biases can be seen for this method with $\alpha = 1$. For higher values of α no systematic biases are visible for the EBP under the augmented model. Finally, for $\alpha = 1{,}000$ very similar results are obtained using all estimation methods. In this case, all approaches yield unbiased estimates on average.

The results in terms of the RRMSE of the area estimates are presented in Figure 3. It is immediately obvious that for values $\alpha \in \{1;1.25,2\}$, the EBP without augmentation is dominated by both other estimation methods. Moreover, the EBP under the augmented model outperforms the SWEE predictor for these values of α. The differences between these two estimation methods is most pronounced for $\alpha = 1$, where the EBP under the augmented model yields very precise estimates. For a value of $\alpha = 1{,}000$, no visible differences can be seen between any of the methods. Altogether, applying the EBP under an augmented model in combination with an informative sampling mechanism leads to an improvement compared to the EBP without augmentation under a non-informative sampling mechanism.

To assess the quality of the precision estimates as well, we show the quality of 95% confidence interval coverage rates in Figure 4. In this graph, each panel depicts one combination of α and the estimation method, where the average coverage rates are shown

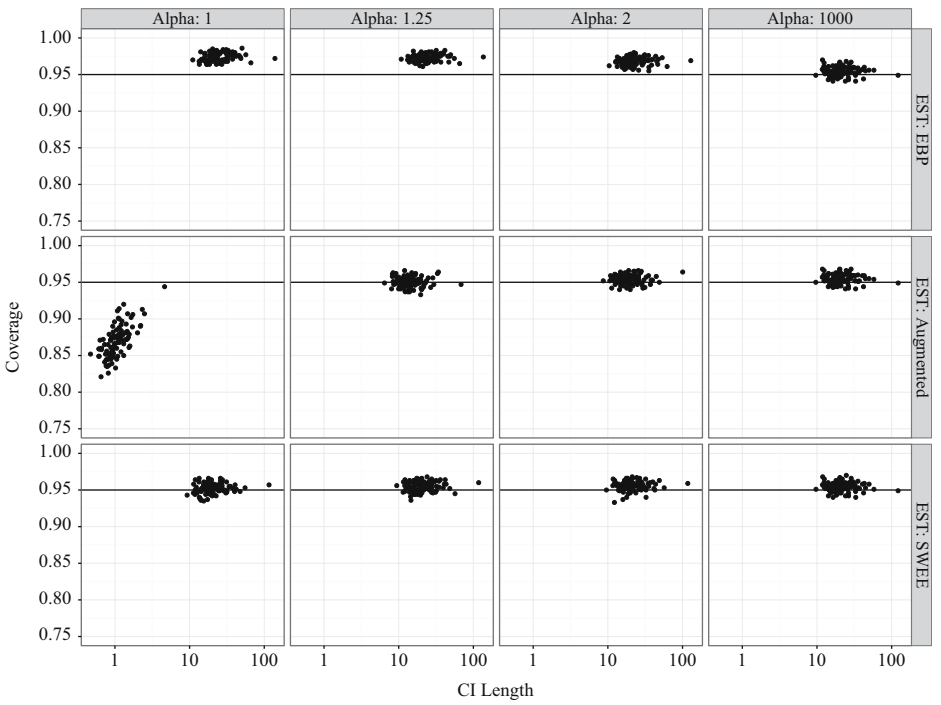

Fig. 4. Quality of confidence intervals under setting 2.

Table 5. Median of the relative biases of MSE estimates for setting 2.

EST	$\alpha = 1$	$\alpha = 1.25$	$\alpha = 2$	$\alpha = 1{,}000$
EBP	0.05	0.10	0.12	0.02
Augmented	−0.80	−0.02	0.01	0.01
SWEE	−0.05	−0.02	−0.00	0.01

on the y-axis and the average length of the intervals on the x-axis using a logarithmic scale. The horizontal line refers to the nominal coverage rate of 95%. Hence, the points should ideally be on the horizontal line and in the left part of the panel, as this would indicate short confidence intervals that meet the nominal rate.

It can be seen from Figure 4 that for a given value of $\alpha < 1{,}000$, the shortest confidence intervals are obtained from the EBP under the augmented model. In the case of $\alpha = 1$, this method yields very short intervals, but they do not meet the desired nominal rate. This undercoverage can, in principle, arise due the biased point estimates or the fact that the normality assumptions cannot be maintained, which may further affect the quality of the MSE estimates. As pointed out by a referee, the low values for the RRMSEs are an indication that the MSE estimates are too low. Indeed, the median relative biases of the MSE estimates shown in Table 5 reveal that for $\alpha = 1$ the MSE estimates of the EBPunder the augmented model are clearly too small. For all other choices of α, however, the nominal rate of 95% coverage is achieved by the EBP under the augmented model. Moreover, the EBP without augmentation yields overcoverage for values of $\alpha < 1{,}000$. Hence, the MSE estimates are not efficient. With respect to the SWEE predictor, we note that the confidence intervals meet the nominal rate for all values of α and their length is in between the one of the EBP with and without augmentation. For a value of $\alpha = 1{,}000$, the confidence intervals produced by all methods are very similar and meet the nominal rate.

4. Conclusion and Outlook

In business statistics, skewed variables of interest and informative sampling designs play a considerable role, especially when applying small area methods. To better cope with the skewness of the variable of interest, the EBP under the lognormal mixed due to Berg and Chandra (2014) can be used. However, this method is based on the implicit assumption of a non-informative sampling design.

We proposed two extensions of the EBP to alleviate biases owing to an informative sampling mechanism. Our first proposal was based on the EBP under an augmented model, whereas our second strategy used design weights to derive the model parameters. Furthermore, we have demonstrated how the selection of the augmented variable can be guided by a measure of predictive accuracy and a check of the normality assumptions.

The results from the simulation study have shown that the EBP under the augmented model leads to significant improvements in the estimation process once the sampling design is informative. The improvements can be seen in all three simulation settings by a lower RRMSE and shorter confidence intervals. Only in the case of very high informativeness, does the EBP under the augmented model suffer from small biases and undercoverage of the confidence intervals. It should be noted that the case of $\alpha = 1$

implies that the selection probabilities are determined solely from the unexplained error term ε_{dj}. In applications, the relationship between the error term of a regression model and the selection probabilities will hardly be very close. Hence, the case $\alpha = 1$ reflects an extreme scenario. Nevertheless, it is interesting to note that the Shapiro-Wilk test indicated a departure from normality in this case. Hence, our diagnostic tool provides important information for properly applying the EBP derived from the augmented model in this case.

Moreover, our second alternative, the SWEE predictor, also achieved an improvement upon the EBP without augmentation in the presence of informative sampling mechanisms. This finding is particularly convenient, as the SWEE predictor does not require access to the selection probabilities for the non-sampled units. Therefore, this method is also applicable in situations where the model analyst only has access to the survey weights for the sampled units.

Furthermore, it should be noted that in our simulation study the informativeness of the sampling mechanisms was induced by sampling with selection probabilities proportional to an Asparouhov-type size measure, which enabled us to settle the degree of informativeness precisely. In many business surveys, stratified random sampling with sample sizes allocated to strata using the Neyman allocation, a take-all stratum for highly influential businesses, or similar approaches are used (cf. Hidiroglou and Lavallee 2009). The issue of informative sampling may still arise and a function of the selection probabilities can be used to augment the model. An alternative option to account for informative sampling in the case of stratified random sampling is to include stratum membership indicators among the covariates. However, this approach may be impractical if the number of strata is very large and the sample sizes within strata very small, which is frequently the case in business surveys with fine stratifications.

Although this article focuses on lognormal models, the approach of augmenting could be easily applied to other model transformations. A condition for the augmenting modelling approach to work is that the augmented model can be validated for the sample data. Once the validity is established, we know that the model holds for the non-sampled part as well.

Future research may focus in two areas. On the one hand, it would be interesting to see how the augmentation and testing methods presented in this article can be considered in the design stage. This may become especially relevant in the future, where more and more design-based and model-based methods are applied simultaneously in business surveys, knowing that design optimization may lead to informativeness. On the other hand, though many national statistical offices are still too conservative to apply Bayesian methods, the popularity of hierarchical Bayes methods may prompt research on adopting the above methods to Bayesian small area methods.

Appendix

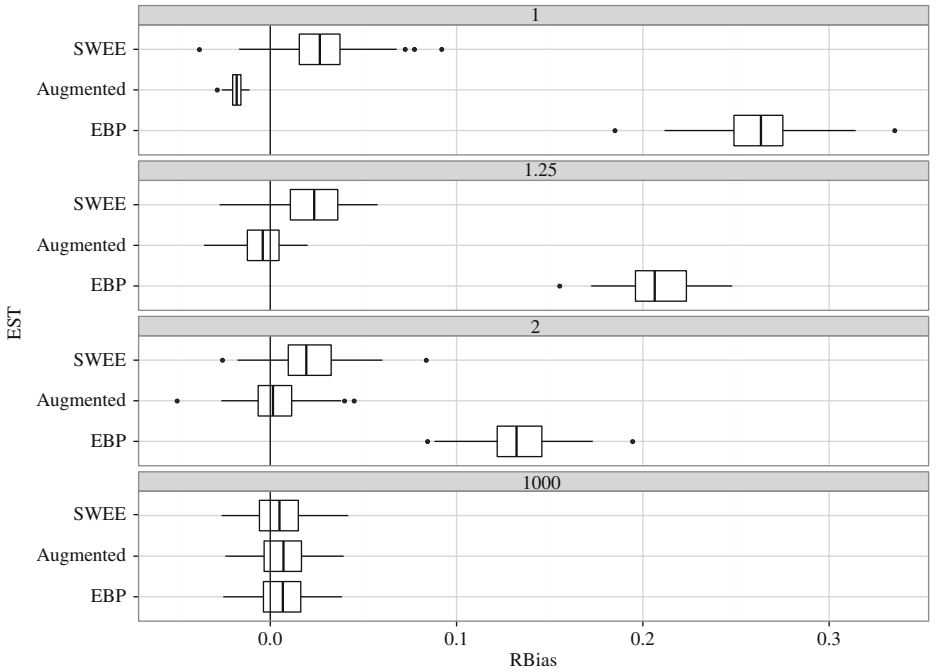

Fig. 5. Relative bias under setting 3.

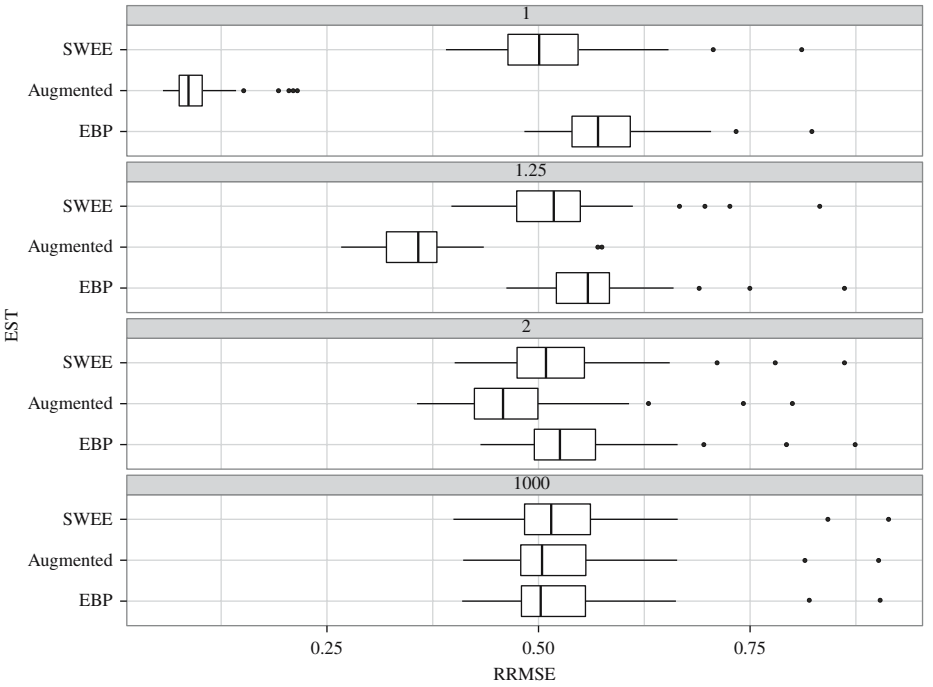

Fig. 6. RRMSE under setting 3.

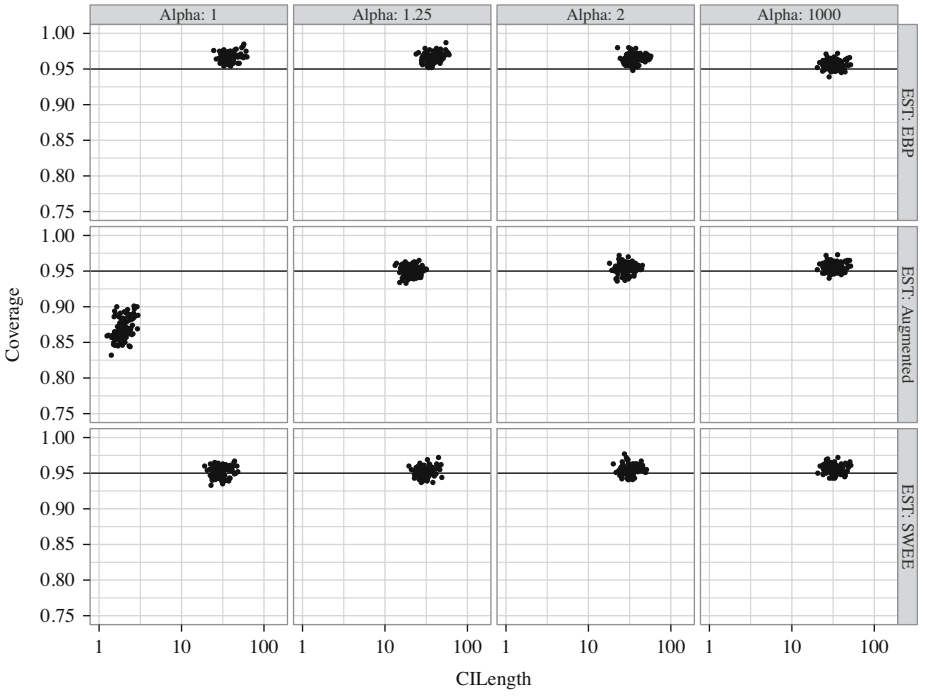

Fig. 7. Quality of confidence intervals under setting 3.

Table 6. Median of the relative biases of MSE estimates for setting 3

EST	$\alpha = 1$	$\alpha = 1.25$	$\alpha = 2$	$\alpha = 1000$
EBP	0.22	0.22	0.25	0.02
Augmented	−0.84	−0.03	0.05	0.01
SWEE	−0.03	−0.03	0.05	0.01

Table 7. Median of the relative biases of MSE estimates for setting 4

EST	$\alpha = 1$	$\alpha = 1.25$	$\alpha = 2$	$\alpha = 1000$
EBP	0.09	0.16	0.18	0.04
Augmented	−0.89	−0.09	0.00	0.05
SWEE	−0.13	−0.06	−0.01	0.04

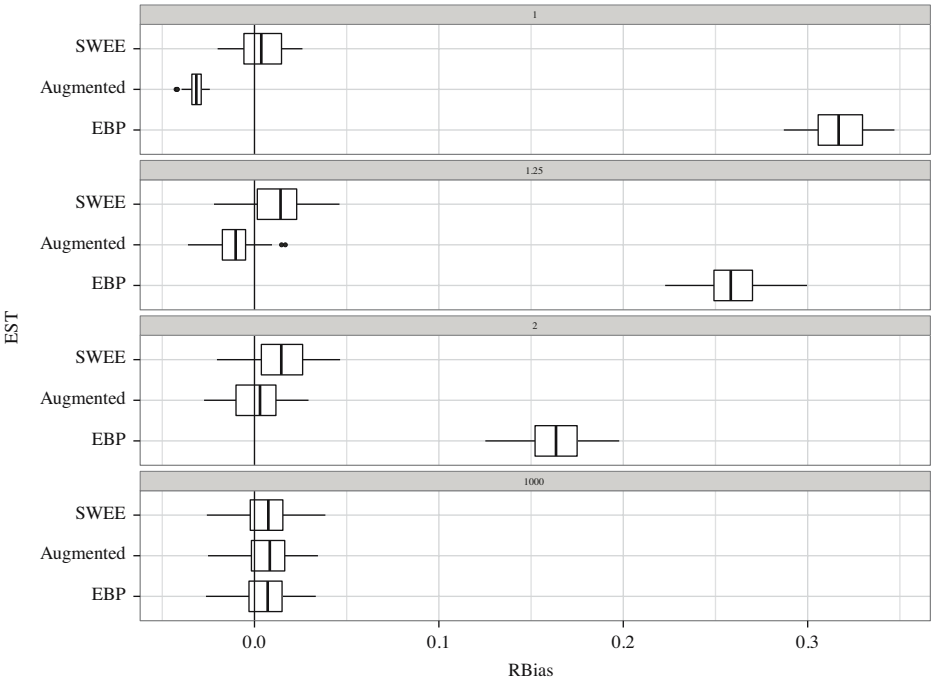

Fig. 8. Relative bias under setting 4.

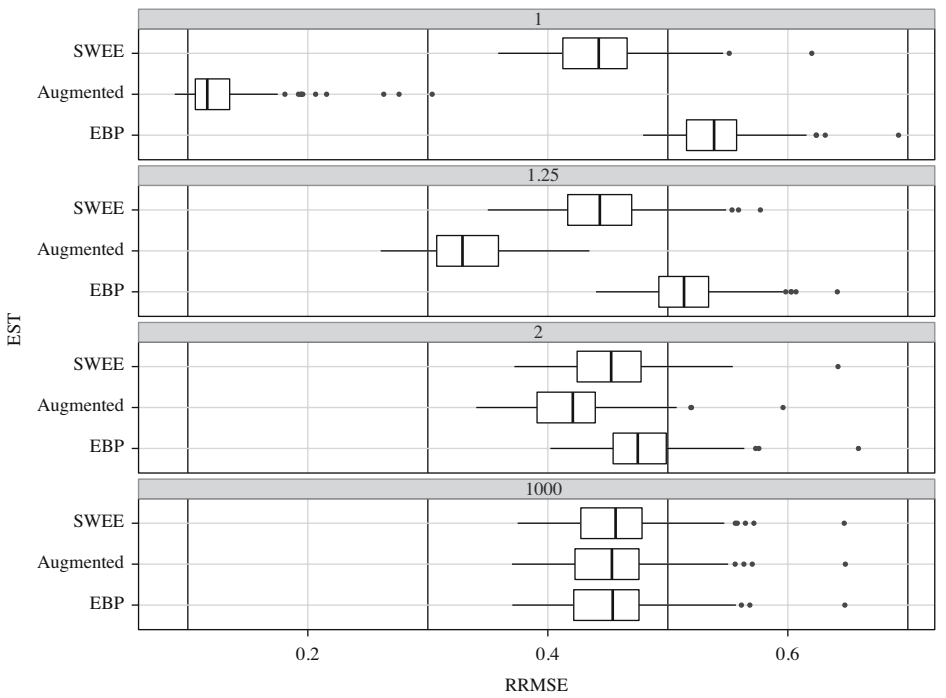

Fig. 9. RRMSE under setting 4.

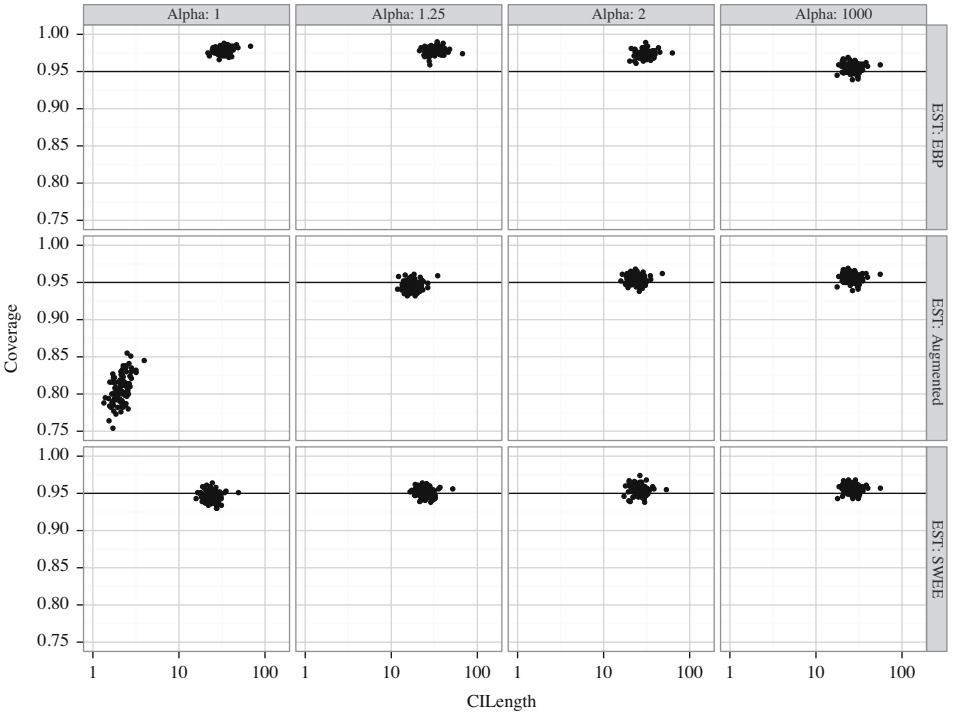

Fig. 10. Quality of confidence intervals under setting 4.

5. References

Alfons, A., M. Templ, and P. Filzmoser. 2010. "An Object-Oriented Framework for Statistical Simulation: The R Package simFrame." *Journal of Statistical Software* 37: 1–36. Doi: http://dx.doi.org/10.18637/jss.v037.i03.

Asparouhov, T. 2006. "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics Theory and Methods* 35: 439–460. Doi: http://dx.doi.org/10.1080/03610920500476598.

Battese, G.E., R.M. Harter, and W.A. Fuller. 1988. "An Error Component 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/10.1080/01621459.1988.10478561.

Berg, E. and H. Chandra. 2014. "Small Area Prediction for a Unit-Level Lognormal Model." *Computational Statistics & Data Analysis* 78: 159–175. Doi: http://dx.doi.org/10.1016/j.csda.2014.03.007.

Bernardini Papalia, R., C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Hernandez-Vazquez, M. Ferrante, J. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk, T. Zimmermann, and S. Zins. 2013. *Best Practice Recommendations on Variance Estimation and Small Area Estimation in Business Surveys*. Technical report, SSH-CT-2010-244767-BLUE-ETS. Available at: http://

www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf (accessed on 12 September 2017).

Burgard, J.P., R. Münnich, and T. Zimmermann. 2014. "The Impact of Sampling Designs on Small Area Estimates for Business Data." *Journal of Official Statistics* 30: 749–771. Doi: http://dx.doi.org/10.2478/jos-2014-0046.

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

Ferrante, M.R., C. Trivisano, and E. Fabrizi. 2016. "Bayesian Small Area Estimation Methods for Business Survey Statistics." In Proceedings of the 60th World Statistics Congress of the International Statistical Institute, 26–31 July 2015, 86–91, Rio de Janeiro.

Hidiroglou, M.A. and P. Lavallee. 2009. "Sampling and Estimation in Business Surveys." In *Handbook of Statistics*, Volume 29 A, edited by D. Pfeffermann and C.R. Rao, Chapter 17, 441–470. Elsevier.

Hidiroglou, M.A. and P. Smith. 2005. "Developing Small Area Estimates for Business Surveys at the ONS." *Statistics in Transition* 7: 527–539.

Jiang, J., P. Lahiri, and S.-M. Wan. 2002. "A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation." *The Annals of Statistics* 30: 1782–1810. Doi: http://dx.doi.org/10.1214/aos/1043351257.

Krieg, S., V. Blaess, and M. Smeets. 2012. "Small Area Estimation of Turnover of the Structural Business Survey." Discussion paper 201203, Statistics Netherlands. Available at: https://www.cbs.nl//media/imported/documents/2012/07/2012-03-x10-pub.pdf (accessed on 12 September 2017).

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation." *Statistical Science* 28: 40–68. Doi: http://dx.doi.org/10.1214/12-STS395.

Pfeffermann, D. and C.R. Rao. 2009a. *Handbook of Statistics: Sample Surveys: Design, Methods and Applications*, Volume 29A. Elsevier.

Pfeffermann, D. and C.R. Rao. 2009b. *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Volume 29A. Elsevier.

Pfeffermann, D. and M. Sverchkov. 2007. "Small-Area Estimation under Informative Probability Sampling of Areas and within the Selected Areas." *Journal of the American Statistical Association* 102: 1427–1439. Doi: http://dx.doi.org/10.1198/016214507000001094.

Pfeffermann, D. and M. Sverchkov. 2009. "Inference under Informative Sampling." In *Handbook of Statistics*, Volume 29B, edited by D. Pfeffermann and C.R. Rao, Chapter 39, 455–487. Elsevier.

Prasad, N.G.N. and J.N.K. Rao. 1999. "On Robust Small Area Estimation Using a Simple Random Effects Model." *Survey Methodology* 25: 67–72.

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons. Doi:10.1002/9781118735855.

Shapiro, S.S. and M.B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52: 591–611. Doi: http://dx.doi.org/10.2307/2333709.

Skinner, C. 1994. "Sample Models and Weights." In Proceedings of the Section on Survey Research Methods: American Statistical Association, 13–18 August 1994. 133–142.

Toronto. Available at: http://ww2.amstat.org/sections/srms/Proceedings/papers/1994_018.pdf (accessed on 12 September 2017).

Tillé, Y. 2006. *Sampling Algorithms*, Springer Series in Statistics. New York: Springer.

Vaida, F. and S. Blanchard. 2005. "Conditional Akaike Information for Mixed-Effects Models." *Biometrika* 92: 351–370. Doi: http://dx.doi.org/10.1093/biomet/92.2.351.

Valliant, R., A.H. Dorfman, and R.M. Royall. 2000. *Finite Population Sampling and Inference: a Prediction Approach*. John Wiley.

Verret, F., M.A. Hidiroglou, and J.N.K. Rao. 2015. "Model-Based Small Area Estimation under Informative Sampling." *Survey Methodology* 41: 333–347.

You, Y. and J.N.K. Rao. 2002. "A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights." *The Canadian Journal of Statistics* 30: 431–439. Doi: http://dx.doi.org/10.2307/3316146.

Zimmermann, T. 2018. *The Interplay between Sampling Design and Statistical Modelling in Small Area Estimation*. PhD thesis, University of Trier.

Received October 2016
Revised September 2017
Accepted October 2017