Nonlinear Analysis: Modelling and Control, Vol. 25, No. 3, 341–357
https://doi.org/10.15388/namc.2020.25.16773

ISSN: 1392-5113
eISSN: 2335-8963

Adaptive composite estimation in small domains

Andrius Čiginas

Institute of Data Science and Digital Technologies,
Vilnius University,
Akademijos str. 4, LT-08663 Vilnius, Lithuania
andrius.ciginas@mif.vu.lt

Received: May 20, 2018 / **Revised:** January 29, 2020 / **Published online:** May 1, 2020

Abstract. Small area estimation techniques are used in sample surveys, where direct estimates for small domains are not reliable due to small sample sizes in the domains. We estimate the domain means by generalized linear compositions of the weighted sample means and the synthetic estimators that are obtained from the regression-synthetic model of fixed effects, based on the domain level auxiliary information. In the proposed method, the number of parameters of optimal compositions is reduced to a single unknown parameter, which is further evaluated by minimizing an empirical risk function. We apply various composite and related estimators to estimate proportions of the unemployed in a simulation study, based on the Lithuanian Labor Force Survey data. Conclusions on advantages and disadvantages of the proposed compositions are obtained from this empirical comparison.

Keywords: small area estimation, synthetic estimator, composite estimator, domain level model, Labor Force Survey.

1 Introduction

The overall sample size of a survey is usually designed to produce reliable estimates of parameters for the whole survey population and for some of its domains, where the reliability is understood as the acceptability of errors of the estimators. However, if there are additional needs to get estimates under more detailed subdivisions of the population, traditional direct estimation methods are not suitable because the direct estimators are based only on the domain sample and the obtained sample sizes can be too small in some of the new domains. For example, in the Lithuanian Labor Force Survey (LFS), the estimated errors of the domain-specific direct estimates of proportions of the unemployed are acceptable at the county level, but the accuracy of the direct estimates is poor at a lower municipality level. It seems that this problem can be solved by allocating the sample according to the subdivision by municipalities, but then the total sample size is too large for the budget of the survey.

© 2020 Author. Published by Vilnius University Press
This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Domains or areas of the population are called small if the unbiased or approximately unbiased direct estimators of parameters have unacceptably large variances in these domains. In the small area estimation (SAE) theory, see [24], alternative indirect estimators, based on implicit or explicit linking models, are considered. Through auxiliary information available from administrative sources or the previous surveys, these estimators borrow sample information from neighbor domains that increase an effective sample size and hence reduce the variances of estimators in the small domain. To efficiently link the additional sample data to the domain of interest, good auxiliary information at the domain or unit level is essential for the underlying models.

Two important classes of the indirect estimators can be distinguished: synthetic and composite estimators. The traditional synthetic estimators are based on the implicit linking models, which rely on the assumptions that small domains resemble the whole population or larger domains. That is, as defined in [10, 24], an estimator is called a synthetic estimator if a reliable direct estimator for a large domain, covering several small domains, is used to derive an estimator for a small domain under the assumption that the small domains have the same characteristics as the large domain. For example, assuming that the true proportions of the unemployed are the same for all municipalities covered by a single county, the direct estimator of that county is a synthetic estimator for each of these municipalities. We consider a different example of the synthetic estimator in Section 2.2, called as a regression-synthetic estimator. The synthetic estimation techniques have been used historically since about the 1970s [10, 11], and they are still popular because of their simplicity and wide applicability [24]. Unfortunately, such an estimation is biased if the implicit model is not approximately true. Therefore, in order to find a trade-off between a bias of the synthetic estimator and a large variance of the direct estimator, the synthetic estimator is linearly combined with the direct estimator as started to be considered in [8, 22] and also in [9], among others. So the specificities of estimation domains are reflected by the direct components of the composite estimators. Many of the estimators, proposed in the later literature, are also linear combinations [24]. There are also new developments as well as generalizations in the synthetic estimation, for instance, [23, 28]. More complex composite estimators are constructed in [15].

Assuming that the auxiliary information is available only at the domain level, we focus ourselves on the estimation of coefficients in composite estimators of domain means. In Section 2.3, we discuss the known empirical approximations to the optimal linear combination, where the direct component is represented by the weighted sample mean from Section 2.1 and the synthetic part is the regression-synthetic estimator, and we offer an adaptive approach to the approximation method of [8]. In [12], estimators are defined using the values of parameters, minimizing the estimated mean square errors (MSEs) that are called the empirical risk functions. Following that idea, our adaptation is based on the minimization of an empirical risk function. We also consider generalized linear compositions and propose analogous adaptive estimators of their coefficients. That is, in Section 2.4, we extend the estimation method of [4]. The regression-synthetic model of fixed effects, which we incorporate in all the compositions, is a special case of the famous Fay–Herriot (FH) domain level model [9]. The FH model is an example of the explicit linking models, and its developments are reviewed in [5]. Estimating additional domain-

specific parameters, the authors of [9] derived the empirical best linear unbiased predictors (EBLUPs) of the domain means, see Section 2.5. Since these indirect estimators are expressed as composite estimators, we include them in further comparisons.

To compare the presented synthetic and composite estimators, the Lithuanian LFS data are used in Section 3. It is clear from a survey of literature that quite various SAE techniques have been analyzed and applied in LFSs of different countries. We shall mention only several instances, related to our context of the study. Earlier examples of [11] and [8] on the estimation of unemployment rates are based on the United States and Canada data, respectively. In the more recent study in [1] of the Italian LFS data, the traditional synthetic estimation is applied. Using the Spanish data, the authors of [19] applied much more complex SAE models, and the paper [26] compared various methods, also including the composite estimation of [8]. The synthetic and composite estimators have been analyzed using the Iranian data in [13]. Among other estimators, EBLUP, based on the FH model, is applied to the Dutch [2], Lithuanian [14], and Swiss [21] unemployment data. The conclusions of our comparative simulation study are discussed in Section 4.

2 Synthetic and composite estimation

2.1 Basic notation

The set $\mathcal{U} = \{1, \dots, N\}$ consists of the labels of elements of the finite survey population. Let y be a continuous or discrete study variable with the fixed values $y_1, \dots, y_N$, assigned to the corresponding elements. In order to estimate the means or proportions in the population and its subsets, the sample $s = \{k_1, \dots, k_n\} \subset \mathcal{U}$ of size $n < N$ is drawn according to the sampling design, denoted by the function $p(\cdot)$, with the values $p(s)$ specifying the probabilities of drawing all possible samples. The inclusion into the sample probabilities

$$\pi_k = \mathrm{P_p}\{k \in s\} = \sum_{s:\ k \in s} p(s), \quad k \in \mathcal{U},$$

are assumed strictly positive. Here the symbol $\mathrm{P_p}$, and hereafter $\mathrm{E_p}$, $\mathrm{var_p}$, and $\mathrm{MSE_p}$ denote probability, expectation, variance, and MSE according to the sampling design, respectively. That is, the probability distribution of any estimator, calculated from the sample s, is induced by the given sampling design $p(\cdot)$. Therefore, the characteristic $\mathrm{var_p}(\cdot)$ is called the sampling variance or design variance, and the like.

Let $\mathcal{U} = \mathcal{U}_1 \cup \dots \cup \mathcal{U}_M$ be a partition of the population into M domains, where $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ as $i \neq j$. There are N_i elements in the domain $\mathcal{U}_i$, and $\sum_{i=1}^M N_i = N$. In general, if $p(\cdot)$ generates samples without replacement, the sample $s_i = s \cap \mathcal{U}_i$ is of the random size $0 \leqslant n_i \leqslant N_i$. Here $s = s_1 \cup \dots \cup s_M$ and $\sum_{i=1}^M n_i = n$. We consider the means

$$\mu_i = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} y_k, \quad i = 1, \dots, M, \tag{1}$$

as target parameters, where N_i are assumed to be known. In the cases where the sampling design $p(\cdot)$ was not constructed in order to ensure samples of fixed sizes in the domains, small (or even zero) n_i can be obtained. Then the accuracy of direct estimators of (1), such as the weighted sample means

$$\hat{\mu}_i^{\mathrm{H}} = \frac{1}{\widehat{N}_i} \sum_{k \in s_i} \frac{y_k}{\pi_k}, \quad \text{where } \widehat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k},\ i = 1, \dots, M, \tag{2}$$

is questionable because their variances are large.

2.2 Synthetic estimator

We assume that, for the domain $\mathcal{U}_i$, the auxiliary information is available as the vector of characteristics $\mathbf{x}_i = (x_{i1}, \dots, x_{iP})'$. This assumption narrows a choice of direct estimators to that similar to (2), however one can expect to improve the weighted sample means by borrowing sample information for the domains through the given auxiliary information. It can be done by applying the regression-synthetic estimators (similar to that in [24, Sect. 3.2.2]), which are constructed following the next two steps. First, let us model the direct estimate $\hat{\mu}_i^{\mathrm{H}}$ as a random variable Y_i of the form

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, M, \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$ are unknown parameters. Here, assuming the deterministic relation $\mu_i \approx \mathbf{x}_i' \boldsymbol{\beta}$, the random variable ε_i is treated as a sampling error in the domain. Thus, we write $\mathrm{E_p}(\varepsilon_i) = 0$ and $\mathrm{var_p}(\varepsilon_i) = \mathrm{var_p}(\hat{\mu}_i^{\mathrm{H}}) = v_i^2$. Despite that the sampling errors may not be independent if the domains cut across the sampling design, the model is usually simplified further assuming that the errors are independent and normally distributed, that is, $\varepsilon_i \sim \mathcal{N}(0, v_i^2)$, and we fix these assumptions, too. Similar assumptions are used in the more general domain level model [24, Sect. 4.2], and we consider its special case including only the fixed effects. Having consistent estimators $\hat{v}_i^2$ of the sampling variances v_i^2, denote by

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_P)' = \left(\sum_{i=1}^{M} \frac{\mathbf{x}_i \mathbf{x}_i'}{\hat{v}_i^2} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{x}_i \hat{\mu}_i^{\mathrm{H}}}{\hat{v}_i^2}$$

the feasible generalized least squares estimator of the parameter $\boldsymbol{\beta}$, obtained from the data $(\hat{\mu}_i^{\mathrm{H}}, \mathbf{x}_i)$, $i = 1, \dots, M$. Second, the synthetic estimators of (1) are defined as

$$\hat{\mu}_i^{\mathrm{S}} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}, \quad i = 1, \dots, M. \tag{4}$$

These estimators rely on the synthetic assumption that the parameter $\boldsymbol{\beta}$ is the same across all domains. Therefore, though their variances are generally small, compared to that of direct estimators (2), contributions of biases to MSEs of (4) can be substantial if the synthetic assumption is not realistic. Moreover, the biases not necessarily decrease as the sample sizes n_i increase.

2.3 Composite estimators

To find a compromise between larger variances of approximately unbiased estimators (2) and biases of synthetic estimators (4), their linear compositions

$$\tilde{\mu}_i^{\mathrm{C}} = b_i \hat{\mu}_i^{\mathrm{H}} + (1 - b_i)\hat{\mu}_i^{\mathrm{S}}, \quad i = 1, \dots, M, \tag{5}$$

with M unknown coefficients $0 \leqslant b_i \leqslant 1$, are a very common approach considered in the literature. Further, following [12], we use the notion "risk function" when dealing with MSE functions or their combinations. Minimizing the risk function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\mu}_i^{\mathrm{C}})$ with respect to b_i, the optimal weight is expressed as the population characteristic [24]

$$\begin{aligned} b_i^* &= \frac{\mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}}) - \mathrm{E}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{H}} - \mu_i)(\hat{\mu}_i^{\mathrm{S}} - \mu_i)}{\mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{H}}) + \mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}}) - 2\,\mathrm{E}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{H}} - \mu_i)(\hat{\mu}_i^{\mathrm{S}} - \mu_i)} \\ &\approx 1 - \frac{v_i^2}{v_i^2 + \mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}})}, \end{aligned} \tag{6}$$

where the approximation is based on the assumptions that $\mathrm{E}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{H}} - \mu_i)(\hat{\mu}_i^{\mathrm{S}} - \mu_i) \ll \mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}})$ and that $\hat{\mu}_i^{\mathrm{H}}$ is nearly unbiased. However, even using the last standard approximation, evaluation of (6) from the sample data is seldom efficient because it is difficult to construct stable estimators of $\mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}})$ without additional assumptions on similarities among domains as originally proposed in [20], see [24], or an additional modeling like in [16]. Straightforward estimators of MSEs of (4) are [24, Sect. 3.2.5]

$$\mathrm{mse}\big(\hat{\mu}_i^{\mathrm{S}}\big) = \big(\hat{\mu}_i^{\mathrm{S}} - \hat{\mu}_i^{\mathrm{H}}\big)^2 - \hat{\sigma}^2\big(\hat{\mu}_i^{\mathrm{S}} - \hat{\mu}_i^{\mathrm{H}}\big) + \hat{\sigma}^2\big(\hat{\mu}_i^{\mathrm{S}}\big), \tag{7}$$

where $\hat{\sigma}^2(\cdot)$ stands for an estimator of the design variance $\mathrm{var}_{\mathrm{p}}(\cdot)$, and their approximations

$$\mathrm{mse}\big(\hat{\mu}_i^{\mathrm{S}}\big) \approx \big(\hat{\mu}_i^{\mathrm{S}} - \hat{\mu}_i^{\mathrm{H}}\big)^2 - \hat{v}_i^2 \tag{8}$$

derived using the assumption $\mathrm{var}_{\mathrm{p}}(\hat{\mu}_i^{\mathrm{S}}) \ll v_i^2$. However, these estimators are unstable for small sample sizes because they include direct estimators (2) of the domain means. Therefore, applying the approximation from (6) and simplified formula (8), the resulting estimators

$$\hat{b}_i^* = 1 - \frac{\hat{v}_i^2}{(\hat{\mu}_i^{\mathrm{S}} - \hat{\mu}_i^{\mathrm{H}})^2} \tag{9}$$

of optimal coefficients (6) are unstable as well, and the efficiency of compositions may be lost.

To get robust estimators of compositions (5), convenient for automatic applications, one can reduce the number M of unknown parameters. One of the known methods, proposed in [22], is to assume a common weight $b_i = \phi$ in (5) and minimize the risk function $R_1(\phi) = \sum_{i=1}^M \mathrm{MSE}_{\mathrm{p}}(\tilde{\mu}_i^{\mathrm{C}})$ with respect to ϕ. Replacing unknown population characteristics by (8) in the expression of minimizer of the approximated risk [24, Sect. 3.4.1], the result

$$\hat{\phi}^* = 1 - \frac{\sum_{i=1}^M \hat{v}_i^2}{\sum_{i=1}^M (\hat{\mu}_i^{\mathrm{S}} - \hat{\mu}_i^{\mathrm{H}})^2} \tag{10}$$

should be more stable as compared to (9). The second method is sample size dependent (SSD) estimation. According to the proposal of [8], the weights $b_i = \phi_i(\delta)$ in (5) also depend on the single parameter δ, that is,

$$\phi_i(\delta) = \begin{cases} 1 & \text{if } \widehat{N}_i/N_i \geqslant \delta, \\ \widehat{N}_i/(\delta N_i) & \text{otherwise.} \end{cases} \tag{11}$$

In this case, the problem of explicit minimization of the risk function $R_2(\delta) = \sum_{i=1}^{M} \mathrm{MSE}_{\mathrm{p}}(\tilde{\mu}_i^{\mathrm{C}})$ with respect to δ is complicated. Therefore, subjective choices of the parameter, such as general $\delta = 1$, are proposed in the literature. On the other hand, the empirical risk function

$$r_2(\delta) = \sum_{i=1}^{M} \bigl(\bigl(\tilde{\mu}_i^{\mathrm{C}} - \hat{\mu}_i^{\mathrm{H}}\bigr)^2 - \hat{\sigma}^2\bigl(\tilde{\mu}_i^{\mathrm{C}} - \hat{\mu}_i^{\mathrm{H}}\bigr) + \hat{\sigma}^2\bigl(\tilde{\mu}_i^{\mathrm{C}}\bigr) \bigr), \tag{12}$$

based on estimators (7) of MSEs, where the compositions $\tilde{\mu}_i^{\mathrm{C}} = \tilde{\mu}_i^{\mathrm{C}}(\delta)$ are treated as synthetic estimators, can be minimized numerically. The use of simplification (8) is not suitable here because then the empirical risk would be minimized for values of δ close to zero. The solution $\hat{\delta}^* = \arg\min_{\delta>0} r_2(\delta)$ should give more efficient weights $\hat{\phi}_i^* = \phi_i(\hat{\delta}^*)$ than that, based on a predefined value of δ. We examine the latter estimation method in the simulation study.

2.4 Adaptive generalized composite estimator

Since parameters (11) of compositions (5) are estimated, as well as (9) and (10) are based on the given sample, the resulting estimated compositions are viewed as adaptive estimators. Consider the adaptive estimation of generalized compositions of (2) and (4) of the form

$$\sum_{i=1}^{M} a_{ij}\hat{\mu}_i^{\mathrm{S}} + \sum_{i=1}^{M} b_{ij}\hat{\mu}_i^{\mathrm{H}}, \quad j = 1, \dots, M, \tag{13}$$

with the coefficients $0 \leqslant a_{ij}, b_{ij} \leqslant 1$, satisfying $\sum_{i=1}^{M}(a_{ij} + b_{ij}) = 1$. Similar combinations are discussed in [18], pointing that simple compositions are improved in this way, due to the larger flexibility when combining estimators. However, straightforward evaluations of optimal coefficients of general compositions lead to instabilities, similar to that in the case of simple combinations (5), and a large number of parameters adds more complexity. Indeed, there are two matrices of unknown weights in (13). Our idea is to construct estimators of the weights a_{ij} and b_{ij}, dependent on a single unknown parameter for all $i, j = 1, \dots, M$, similarly as in SSD estimation of [8], and to evaluate the optimal value of this parameter by minimizing an empirical risk function.

Given the sample s, let us create two sets of size $2M$ that contain the known numbers. The first set is $\mathcal{P} = \{\hat{\tau}_1, \dots, \hat{\tau}_M, \hat{\tau}_{M+1}, \dots, \hat{\tau}_{2M}\}$, where $\hat{\tau}_i = \hat{\mu}_i^{\mathrm{S}}$, $i = 1, \dots, M$, are the synthetic estimates, and the elements $\hat{\tau}_{i+M} = \hat{\tau}_i$, $i = 1, \dots, M$, are their replications. The second set is $\mathcal{R} = \{\hat{\mu}_1^{\mathrm{H}}, \dots, \hat{\mu}_M^{\mathrm{H}}, \hat{\tau}_{M+1}, \dots, \hat{\tau}_{2M}\}$. Arrange both sets according to

the values in the set $\mathcal{P}$, so that we could obtain $\mathcal{P} = \{\hat{\tau}_{k_1}, \hat{\tau}_{k_1+M}, \dots, \hat{\tau}_{k_M}, \hat{\tau}_{k_M+M}\}$ with the ordering $\hat{\tau}_{k_1} = \hat{\tau}_{k_1+M} \leqslant \dots \leqslant \hat{\tau}_{k_M} = \hat{\tau}_{k_M+M}$, and then $\mathcal{R} = \{\hat{\mu}^{\mathrm{H}}_{k_1}, \hat{\tau}_{k_1+M}, \dots, \hat{\mu}^{\mathrm{H}}_{k_M}, \hat{\tau}_{k_M+M}\}$. To simplify the notation of permuted indices, temporarily change the numeration of domains from $\{k_1, \dots, k_M\}$ to the set of odd indices $\mathcal{I}_{\mathrm{H}} = \{1, \dots, 2M-1\}$, and let the remaining indices $\{k_1 + M, \dots, k_M + M\}$, changed to the set of even numbers $\mathcal{I}_{\mathrm{S}} = \{1, \dots, 2M\} \backslash \mathcal{I}_{\mathrm{H}}$, be an alternative (auxiliary) numeration of all the domains. Next, we apply the transformation

$$\tilde{\mu}^{\mathrm{HS}}_i = \begin{cases} \hat{\tau}_i + \theta^{1/2}(\hat{\mu}^{\mathrm{H}}_i - \hat{\tau}_i) & \text{if } i \in \mathcal{I}_{\mathrm{H}}, \\ \hat{\tau}_i & \text{if } i \in \mathcal{I}_{\mathrm{S}} \end{cases}$$

to the values of set $\mathcal{R}$, where $\theta \geqslant 0$ is a new parameter. Using the fixed estimates $\hat{v}^2_i$, taken from (4), write the stochastic model

$$\widetilde{Y}_i = \hat{\tau}_i + \tilde{\varepsilon}_i, \quad \text{where } \tilde{\varepsilon}_i \sim \begin{cases} \mathcal{N}(0, \theta\hat{v}^2_i) & \text{if } i \in \mathcal{I}_{\mathrm{H}}, \\ \mathcal{N}(0, 0) & \text{if } i \in \mathcal{I}_{\mathrm{S}} \end{cases} \tag{14}$$

are independent random components, which simulates the dependence between the transformed value $\tilde{\mu}^{\mathrm{HS}}_i$, modeled as the random variable $\widetilde{Y}_i$, and the corresponding value $\hat{\tau}_i$, taken from the set $\mathcal{P}$. Here $\mathcal{N}(0, 0)$ denotes the degenerate distribution with zero mean. Model (14) is an estimated and extended regression (3), which generates synthetic estimates (4) and their randomized analogues. For the value $\theta = 1$ and the set of indices $\mathcal{I}_{\mathrm{H}}$, this model imitates direct estimates (2). Introduce the characteristics of random variables (14):

$$\tilde{p}_{i;j} = \tilde{p}_{i;j}(\theta) = \mathrm{P_p}\{\widetilde{R}_i = j\}, \quad i, j = 1, \dots, 2M, \tag{15}$$

where $\widetilde{R}_1, \dots, \widetilde{R}_{2M}$ are ranks of $\widetilde{Y}_1, \dots, \widetilde{Y}_{2M}$. Treating the synthetic estimate $\hat{\tau}_i$ as the initial size (of the parameter μ_i) of the domain $\mathcal{U}_i$, where $i \in \mathcal{I}_{\mathrm{H}}$ or $i \in \mathcal{I}_{\mathrm{S}}$, this domain takes the ith position in the sequence $\mathcal{U}_1, \dots, \mathcal{U}_{2M}$ of all domains and their duplicates. Then the collection $\{\tilde{p}_{i;j}, j = 1, \dots, 2M\}$ of probabilities models a change of the ith initial position to the new jth position. A similar interpretation is that the probability $\tilde{p}_{i;j}$ describes the distance from the domain $\mathcal{U}_i$ to the domain $\mathcal{U}_j$ of interest by means of the difference in their sizes $\hat{\tau}_i$ and $\hat{\tau}_j$, taking into account the distances among all domains. A relatively large value of this probability suggests that both these domains are similar, that is, relatively more information can be borrowed from the domain $\mathcal{U}_i$ to increase the effective sample size in the domain $\mathcal{U}_j$. Therefore, we use characteristics (15) as the weights in the new family of estimators of means (1), defined as

$$\tilde{\mu}^{\mathrm{HR}}_{k_j} = \tilde{\mu}^{\mathrm{HR}}_{k_j}(\theta) = \sum_{i=1}^{2M} \tilde{p}_{i;2j}\tilde{\mu}^{\mathrm{HS}}_i, \quad j = 1, \dots, M, \tag{16}$$

which depends on the parameter θ. Here $k_1, \dots, k_M$ are the original labels of domains. Since these estimators are alternatively expressed as

$$\sum_{i \in \mathcal{I}_{\mathrm{H}}} \big((1 - \theta^{1/2})\tilde{p}_{i;2j}(\theta) + \tilde{p}_{i+1;2j}(\theta)\big)\hat{\mu}^{\mathrm{S}}_i + \theta^{1/2} \sum_{i \in \mathcal{I}_{\mathrm{H}}} \tilde{p}_{i;2j}(\theta)\hat{\mu}^{\mathrm{H}}_i, \tag{17}$$

$j = 1, \dots, M$, they are partially estimated generalized compositions (13). That is, there is a single unknown parameter θ in estimators (16) instead of $2M^2$ unknown coefficients in the general combinations. Probabilities (15) also depend on the estimated sampling variances (variabilities) $\hat{v}_i^2$, $i \in \mathcal{I}_{\mathrm{H}}$, of the direct estimators. A domain with a relatively large variability tends to borrow sample information from a larger number of neighbor domains. The tuning constant θ controls this number of neighbors through model (14). The probabilities $\tilde{p}_{i;j}(\theta)$ tend to $\mathbb{I}(i = j)$ as the smoothing parameter $\theta \to 0$, and then (16) reduces to regression-synthetic estimators (4). Here $\mathbb{I}(\cdot)$ is the indicator function. In the latter case, all the domains have no neighbors according to our interpretation.

Due to the complexity of weights (15), dependent on the sample, it is difficult to get expressions of the risk function $R_3(\theta) = \sum_{j=1}^{M} \mathrm{MSE}_{\mathrm{p}}(\tilde{\mu}_{k_j}^{\mathrm{HR}})$. The minimizer θ^* of this function can be estimated by minimizing the sample-based risk function

$$r_3(\theta) = \sum_{j=1}^{M} \bigl(\bigl(\tilde{\mu}_{k_j}^{\mathrm{HR}} - \hat{\mu}_{k_j}^{\mathrm{H}}\bigr)^2 - \hat{\sigma}^2\bigl(\tilde{\mu}_{k_j}^{\mathrm{HR}} - \hat{\mu}_{k_j}^{\mathrm{H}}\bigr) + \hat{\sigma}^2\bigl(\tilde{\mu}_{k_j}^{\mathrm{HR}}\bigr)\bigr), \tag{18}$$

just like we proposed for SSD estimation of [8], or its approximation

$$\bar{r}_3(\theta) = \sum_{j=1}^{M} \bigl(\bigl(\tilde{\mu}_{k_j}^{\mathrm{HR}} - \hat{\mu}_{k_j}^{\mathrm{H}}\bigr)^2 - \hat{v}_{k_j}^2 \bigr). \tag{19}$$

Given the sample, function (18) is minimized numerically with respect to θ. Here, in each iteration, for the fixed value of θ, the matrix of probabilities (15) is evaluated by independently simulating a large number of samples of size $2M$ from (14). That is, for each of these samples $\tilde{y}_1^{(t)}, \dots, \tilde{y}_{2M}^{(t)}$, $t = 1, \dots, T$, we find the ranks $\tilde{r}_1^{(t)}, \dots, \tilde{r}_{2M}^{(t)}$, and then we apply the approximation

$$\tilde{p}_{i;j} \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\bigl(\tilde{r}_i^{(t)} = j\bigr), \quad i, j = 1, \dots, 2M.$$

Denote the solution by $\hat{\theta}^* = \arg\min_{\theta \geqslant 0} r_3(\theta)$. The use of (19) instead of (18) lets us avoid repetitive calculations of possibly complex (for instance, bootstrap) estimates $\hat{\sigma}^2(\cdot)$. The obtained estimate $\hat{\theta}^*$ of the smoothing parameter θ is plugged into (16) to define the applicable estimators

$$\hat{\mu}_{k_j}^{\mathrm{HR}} = \tilde{\mu}_{k_j}^{\mathrm{HR}}(\hat{\theta}^*), \quad j = 1, \dots, M,$$

of the domain means, which are estimated generalized composite estimators (13) with the weights

$$\hat{a}_{k_i k_j} = \bigl(1 - \hat{\theta}^{*1/2}\bigr)\tilde{p}_{2i-1;2j}(\hat{\theta}^*) + \tilde{p}_{2i;2j}(\hat{\theta}^*), \quad i, j = 1, \dots, M,$$

and

$$\hat{b}_{k_i k_j} = \hat{\theta}^{*1/2}\tilde{p}_{2i-1;2j}(\hat{\theta}^*), \quad i, j = 1, \dots, M,$$

according to expression (17). It is easy to check that the identity $\sum_{i=1}^{M}(\hat{a}_{k_ik_j}+\hat{b}_{k_ik_j})=1$ holds.

Let us return back to the previous numeration of domains. To estimate MSEs of the proposed adaptive estimators $\hat{\mu}_j^{\mathrm{HR}}$, the general method of [20] can be applied, where squared design-biases of estimators are assumed approximately equal in the domains. The application leads to the estimators of $\mathrm{MSE}_{\mathrm{p}}(\hat{\mu}_j^{\mathrm{HR}})$, having the expressions

$$\mathrm{mse}_{\mathrm{M}}\left(\hat{\mu}_j^{\mathrm{HR}}\right)=\hat{\sigma}^2\left(\hat{\mu}_j^{\mathrm{HR}}\right)+\frac{1}{M}\sum_{i=1}^{M}\left(\left(\hat{\mu}_i^{\mathrm{HR}}-\hat{\mu}_i^{\mathrm{H}}\right)^2-\hat{\sigma}^2\left(\hat{\mu}_i^{\mathrm{HR}}-\hat{\mu}_i^{\mathrm{H}}\right)\right), \tag{20}$$

$j=1,\dots,M$. In the case of minimizer $\hat{\theta}^*$ obtained from (19), the formulas

$$\overline{\mathrm{mse}}_{\mathrm{M}}\left(\hat{\mu}_j^{\mathrm{HR}}\right)=\hat{\sigma}^2\left(\hat{\mu}_j^{\mathrm{HR}}\right)+\frac{1}{M}\sum_{i=1}^{M}\left(\left(\hat{\mu}_i^{\mathrm{HR}}-\hat{\mu}_i^{\mathrm{H}}\right)^2-\hat{v}_i^2-\hat{\sigma}^2\left(\hat{\mu}_i^{\mathrm{HR}}\right)\right), \tag{21}$$

$j=1,\dots,M$, would be more relevant. However, as noticed in [24, Sect. 3.2.5], such an estimation of MSE may be inefficient if the assumption on similar squared biases is not suitable or the variance term does not dominate the bias term in (20) and (21).

2.5 EBLUP under the Fay–Herriot model

Recall model (3) and its assumptions. Including domain-specific random effects a_i, $i=1,\dots,M$, assumed independent of the sampling errors, into that model, FH model [9]

$$Y_i^*=\mathbf{x}_i'\boldsymbol{\beta}+a_i+\varepsilon_i,\quad i=1,\dots,M, \tag{22}$$

is obtained, where the random variable Y_i^* models the direct estimate $\hat{\mu}_i^{\mathrm{H}}$. Here the random effects are supposed independent and identically distributed with $\mathrm{E}(a_i)=0$ and $\mathrm{var}(a_i)=\tau^2$ in respect of a distribution, different from that generated by the sampling design.

Then, treating the estimates $\hat{v}_i^2$ as given numbers, EBLUPs of means (1), according to [9], are the compositions

$$\hat{\mu}_i^{\mathrm{FH}}=\hat{\gamma}_i\hat{\mu}_i^{\mathrm{H}}+(1-\hat{\gamma}_i)\mathbf{x}_i'\hat{\boldsymbol{\beta}}\left(\hat{\tau}^2\right),\quad i=1,\dots,M, \tag{23}$$

with

$$\hat{\gamma}_i=\frac{\hat{\tau}^2}{\hat{v}_i^2+\hat{\tau}^2}\quad\text{and}\quad\hat{\boldsymbol{\beta}}(\hat{\tau}^2)=\left(\sum_{i=1}^{M}\frac{\mathbf{x}_i\mathbf{x}_i'}{\hat{v}_i^2+\hat{\tau}^2}\right)^{-1}\sum_{i=1}^{M}\frac{\mathbf{x}_i\hat{\mu}_i^{\mathrm{H}}}{\hat{v}_i^2+\hat{\tau}^2},$$

where $\hat{\tau}^2$ is an estimator of the variance τ^2 of the domain-specific random effects.

3 Applications in Labor Force Survey

To compare efficiencies of the discussed direct, synthetic, and composite estimators, also including EBLUPs (23), we carry out the simulation experiments using the Lithuanian

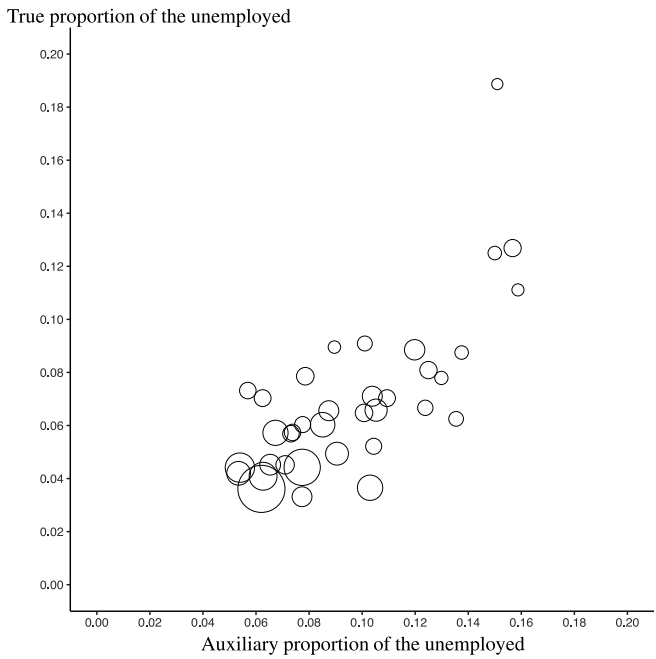

Figure 1. The true proportions of the unemployed μ_i, the auxiliary proportions of the unemployed x_{i2}, and the sizes N_i of municipalities, represented by the areas of bubbles.

LFS data of the first quarter of 2018. A part of these sample data is public [17], and Statistics Lithuania granted the permission to use some other variables according to the agreement No. STAT-132(2019). We have narrowed the sample to individuals from age interval 15–70, and we also removed a part of municipalities from the data because of too small fractions of the unemployed.

The fixed artificial population, created by replicating the remaining sample data 6 times, consists of $N = 51226$ persons. The study variable y is binary: $y_k = 1$ if the kth individual is unemployed according to the survey definition, and $y_k = 0$ otherwise, $k \in \mathcal{U}$. Therefore, the target means μ_i, $i = 1, \dots, M$, are proportions of the unemployed in $M = 36$ municipalities. To create model (3), auxiliary variables were selected using a stepwise regression, similarly as in [14], where the true values μ_i were taken as the values of a dependent variable. The obtained model has the coefficient of determination equal to 0.52, and it is based on the known auxiliary vectors $\mathbf{x}_i = (1, x_{i2}, x_{i3}, x_{i4}, x_{i5})'$, where x_{i2} is the proportion of unemployed (registered unemployed individuals) derived from the administrative Lithuanian Labor Exchange data, x_{i3} is the proportion of urban population, and x_{i4} and x_{i5} are the proportions of individuals from age intervals 15–25 and 41–55, respectively. Figure 1 shows the relation between the true proportion of the unemployed μ_i and the auxiliary proportion x_{i2}, which is the best regressor in model (3), and it might be used alone for the prediction. It is seen that both proportions of the unemployed tend to be larger for smaller municipalities.

In the real survey, the sample of households is drawn with probabilities proportional to the number of their members, and then entire households are surveyed. For simplicity, we consider the simple random sampling without replacement $p(\cdot)$, which is also sufficient to get reliable direct estimates of the proportions for the subdivision of $\mathcal{U}$ by counties according to the primary aim of the survey, but the expected sample sizes are too low for a part of municipalities. For the given sampling design, weighted sample means (2) reduce to the unbiased estimators

$$\hat{\mu}_i^{\mathrm{H}} = \frac{1}{n_i} \sum_{k \in s_i} y_k, \quad i = 1, \dots, M,$$

of μ_i, with the sampling variances [25]

$$v_i^2 = \frac{N_i - n_i}{N_i - 1} \frac{\mu_i(1-\mu_i)}{n_i}, \quad i = 1, \dots, M.$$

The design-unbiased estimators

$$\hat{v}_i^{2(\mathrm{H})} = \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{\mu}_i^{\mathrm{H}}(1-\hat{\mu}_i^{\mathrm{H}})}{n_i - 1}, \quad i = 1, \dots, M, \tag{24}$$

of the parameters v_i^2 are consistent as the population and sample sizes approach the infinity in some way. However, they are unstable for small sample sizes as well as their components. We apply a generalized variance function method [27] in order to smooth estimators (24). More specifically, since the direct estimators $\hat{\mu}_i^{\mathrm{H}}$ of domains with smaller sizes N_i have larger sampling variances v_i^2 because of smaller sample sizes, and due to the mentioned allocation of the domains by their sizes N_i in Fig. 1, we choose the assumption $v_i^2 \approx K N_i^{\gamma}$, such as in [7]. We estimate the parameters $K > 0$ and $\gamma \in \mathbb{R}$ using the regression equation

$$\log V_i^2 = \log K + \gamma \log N_i + \eta_i \quad \text{with } \eta_i \sim \mathcal{N}(0, \zeta^2),$$

assumed as independent errors, where the random variables V_i^2 model estimates (24). In practice, from the pairs of data $(\hat{v}_i^{2(\mathrm{H})}, N_i)$, satisfying $\hat{v}_i^{2(\mathrm{H})} > 0$, we obtain ordinary least squares estimates of the regression parameters, and then

$$\hat{v}_i^2 = \widehat{K} N_i^{\hat{\gamma}}, \quad i = 1, \dots, M,$$

are smoothed estimates of v_i^2, which we apply to derive regression-synthetic estimators (4). The estimators $\hat{\mu}_i^{\mathrm{H}}$ and $\hat{\mu}_i^{\mathrm{S}}$ are components of the variants of compositions (5) and (16), compared in the numerical survey:

(i) $\hat{\mu}_i^{\mathrm{opt}} = \tilde{\mu}_i^{\mathrm{C}}(b_i^*)$, based on optimal weights (6);
(ii) $\hat{\mu}_i^{\mathrm{uns}} = \tilde{\mu}_i^{\mathrm{C}}(\hat{b}_i^*)$ with possibly unstable coefficients (9);
(iii) $\hat{\mu}_i^{\mathrm{PKopt}} = \tilde{\mu}_i^{\mathrm{C}}(\phi^*)$, where ϕ^* is a minimizer of the true risk $R_1(\phi)$;
(iv) $\hat{\mu}_i^{\mathrm{PK}} = \tilde{\mu}_i^{\mathrm{C}}(\hat{\phi}^*)$ with sample-based minimizer (10);

(v) $\hat{\mu}_i^{\mathrm{DSCopt}} = \tilde{\mu}_i^{\mathrm{C}}(\phi_i(\delta^*))$, where δ^* is a minimizer of the risk $R_2(\delta)$;
(vi) $\hat{\mu}_i^{\mathrm{DSCfix}} = \tilde{\mu}_i^{\mathrm{C}}(\phi_i(1.5))$ at the fixed value $\delta = 1.5$;
(vii) $\hat{\mu}_i^{\mathrm{DSC}} = \tilde{\mu}_i^{\mathrm{C}}(\phi_i(\hat{\delta}^*))$ with the minimizer $\hat{\delta}^*$ of empirical risk (12);
(viii) $\hat{\mu}_i^{\mathrm{HRopt}} = \tilde{\mu}_i^{\mathrm{HR}}(\theta^*)$, where θ^* is a minimizer of the function $R_3(\theta)$;
(ix) $\hat{\mu}_i^{\mathrm{HR}} = \tilde{\mu}_i^{\mathrm{HR}}(\hat{\theta}^*)$ with the minimizer $\hat{\theta}^*$ of simplified empirical risk (19).

Here the optimal values b_i^*, ϕ^*, δ^*, and θ^* of theoretical interest are obtained by evaluating the corresponding estimators for a large number of samples drawn from the population, and searching for the minimum of the respective true risk function. To compute the variances $\hat{\sigma}^2(\cdot)$ in the empirical risk of case (vii), we apply a resampling procedure from [3]. That is, the number of artificial populations of size N is reconstructed from the given sample, and a larger number of bootstrap samples is drawn from each of them using original $p(\cdot)$. Then the bootstrap variance is the variance of values of the required estimator, that are calculated from all the samples, multiplied by the finite population correction factor $(N-n)/(N-1)$. If risk (18) were used in (ix), the bootstrap is applicable as well.

In formula (23) of the estimators $\hat{\mu}_i^{\mathrm{FH}}$, we use the iterative FH moment estimator $\hat{\tau}^2$ of the variance τ^2 as proposed in [9]. The choice is based on the analysis made in [6].

We compare the efficiencies of estimators by evaluating them for each of $R = 10^4$ independent samples, selected from the survey population, and then estimating their root mean squared errors (RMSEs) and absolute biases (ABs). That is, we use the accuracy measures

$$\mathrm{RMSE}(\hat{\mu}_i) = \left(\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_i^{(r)} - \mu_i)^2\right)^{1/2} \quad \text{and} \quad \mathrm{AB}(\hat{\mu}_i) = \left|\frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_i^{(r)} - \mu_i\right|, \tag{25}$$

where $\hat{\mu}_i^{(r)}$ is a realization of the specific estimator $\hat{\mu}_i$ of μ_i, based on the rth Monte Carlo sample.

The estimators $\hat{\mu}_i^{\mathrm{H}}$, $\hat{\mu}_i^{\mathrm{S}}$, (i)–(ix), and $\hat{\mu}_i^{\mathrm{FH}}$ of interest are denoted hereafter by their respective superscripts. We compute their characteristics (25) for two different overall sample sizes: $n = 2000$ and $n = 5000$. In each of these two situations, the domains are classified by the expected domain sample size into three classes of equal size, and the average of RMSEs as well as ABs is calculated over domains of each class. These results for different n are reported in Tables 1 and 2, respectively, where the averages of (25) over all domains are also presented as common measures of accuracy.

In the tables, estimator H acts as a reference. It is seen that such a direct estimation is not reliable at the municipality level of the survey. In contrast, regression-synthetic estimator S is much more efficient in both tables, despite that its ABs are comparatively large. Treating RMSE as a principal measure of accuracy, optimal composition opt would be the best choice in all the classes of domains. Then the respective errors of other theoretical estimators PKopt, DSCopt, and HRopt inform us what we lose due to reduction of the number of parameters to the single unknown parameter. If to compare the errors of empirical analogues PK, DSC, and HR with respect to that of synthetic

Table 1. Average RMSEs and ABs of estimators in domain size classes as $n = 2000$.

Estimator	Average RMSE ($\times 10^2$)				Average AB ($\times 10^2$)			
	Expected domain sample size				Expected domain sample size			
	All	< 33	$[33, 51]$	> 51	All	< 33	$[33, 51]$	> 51
H	4.0276	5.9865	3.7721	2.3241	0.2260	0.3808	0.1903	0.1069
S	1.8480	2.6500	1.7314	1.1627	1.4619	2.2122	1.3343	0.8391
opt	1.6311	2.3591	1.5085	1.0257	1.0678	1.6836	0.9262	0.5937
uns	2.7552	4.0172	2.5458	1.7028	0.7516	1.1202	0.6796	0.4551
PKopt	1.7953	2.5708	1.6790	1.1360	1.1937	1.8252	1.0839	0.6720
PK	2.0290	3.0094	1.8336	1.2442	1.1373	1.6728	1.0603	0.6789
DSCopt	1.7987	2.5705	1.6889	1.1367	1.1858	1.8076	1.0815	0.6684
DSCfix	2.8260	4.1180	2.6794	1.6807	0.5883	0.8952	0.5314	0.3384
DSC	2.1774	3.1783	2.0246	1.3294	0.9697	1.4208	0.9147	0.5736
HRopt	1.8175	2.6248	1.6821	1.1457	1.4382	2.1987	1.2974	0.8184
HR	1.8203	2.6043	1.6039	1.2525	1.4119	2.1596	1.2279	0.8483
FH	1.8909	2.6112	1.7105	1.3509	1.2069	1.9840	1.0663	0.5704

Table 2. Average RMSEs and ABs of estimators in domain size classes as $n = 5000$.

Estimator	Average RMSE ($\times 10^2$)				Average AB ($\times 10^2$)			
	Expected domain sample size				Expected domain sample size			
	All	< 84	$[84, 130]$	> 130	All	< 84	$[84, 130]$	> 130
H	2.4999	3.6542	2.4902	1.3553	0.1336	0.2080	0.1118	0.0811
S	1.6162	2.3696	1.5120	0.9672	1.4624	2.2141	1.3389	0.8342
opt	1.2029	1.7572	1.1109	0.7405	0.7826	1.2103	0.6766	0.4607
uns	1.9410	2.8784	1.9267	1.0180	0.5863	0.7938	0.5918	0.3735
PKopt	1.5026	2.1599	1.4699	0.8779	0.9070	1.3517	0.8419	0.5276
PK	1.6667	2.4585	1.6149	0.9266	0.8183	1.1877	0.7843	0.4830
DSCopt	1.5004	2.1514	1.4744	0.8755	0.9062	1.3571	0.8333	0.5282
DSCfix	1.8067	2.5911	1.8159	1.0132	0.5423	0.7961	0.5066	0.3242
DSC	1.7573	2.5937	1.7192	0.9590	0.6712	0.9652	0.6431	0.4052
HRopt	1.5717	2.3131	1.4581	0.9439	1.4159	2.1512	1.2870	0.8096
HR	1.5883	2.3347	1.4060	1.0243	1.3682	2.0690	1.2348	0.8009
FH	1.5949	2.2530	1.5324	0.9994	0.8915	1.6003	0.7689	0.3053

estimator S, composition HR improves S in two classes of smaller domains in both tables, while compositions PK and DSC are worse than S in almost all the classes of domains. Adaptive composition DSC is a significantly better choice than simple case DSCfix, but with alternative PK more stable results are obtained, despite that all the accuracy characteristics of corresponding DSCopt and PKopt are very similar. On the other hand, ABs of DSC are lower than that of PK in both tables. Naive composition uns improves direct estimator H, but yields similar results to that of composition DSCfix. Comparing more reliable and applicable estimators S, PK, DSC, HR, and FH, compositions PK and DSC appear relatively good only in the largest domains, and estimator FH is better than S in smaller domains. Generalized composition HR outperforms FH as $n = 2000$, and the accuracy of both these compositions is similar for a larger overall sample size (Table 2).

In the case of smaller n, a worse performance of compositions PK and DSC may be explained by a relatively small contribution of the direct components as seen from

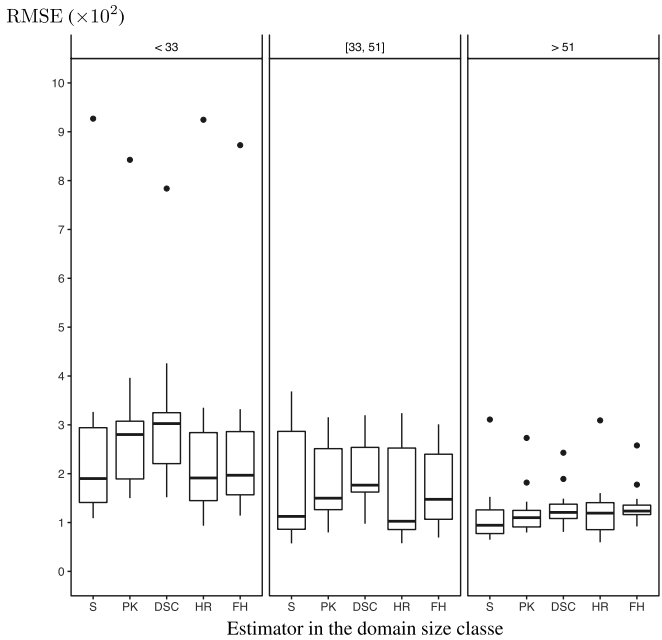

Figure 2. Boxplots of RMSEs of estimators in domain size classes as $n = 2000$.

Table 3. Values of optimal parameters.

n	b^*	ϕ^*	δ^*	θ^*
2000	0.15	0.20	4.88	0.01
5000	0.26	0.40	2.50	0.03

the values of the summary characteristic $b^* = M^{-1}\sum_{i=1}^{M} b_i^*$ and the values of optimal parameters ϕ^* and δ^* in Table 3. In this case, errors of the estimated optimal parameters are large compared to the sizes of the true values. The values of δ^* are larger than the blindly chosen value $\delta = 1.5$ in DSCfix.

To get a deeper insight into differences among reliable estimators S, PK, DSC, HR, and FH, we also present the distributions of their RMSEs in the classes of domains. The situations $n = 2000$ and $n = 5000$ are shown in Fig. 2 and 3, respectively. One can conclude from both figures that RMSEs of estimators PK, DSC, and FH are more homogeneous across the domains than that in the case of estimators S and HR. On the other hand, the latter two estimators tend to yield good estimation results for a part of domains in their different classes. That is, the prediction by regression-synthetic estimator S is efficient for large domains, and it is well explained by the scattering in Fig. 1, and adaptive generalized composition HR is effective for a large part of domains from the middle class because there is a larger number of similarly small domains (Fig. 1), and therefore sample information is flexibly shared among them according to the construction of the estimator.

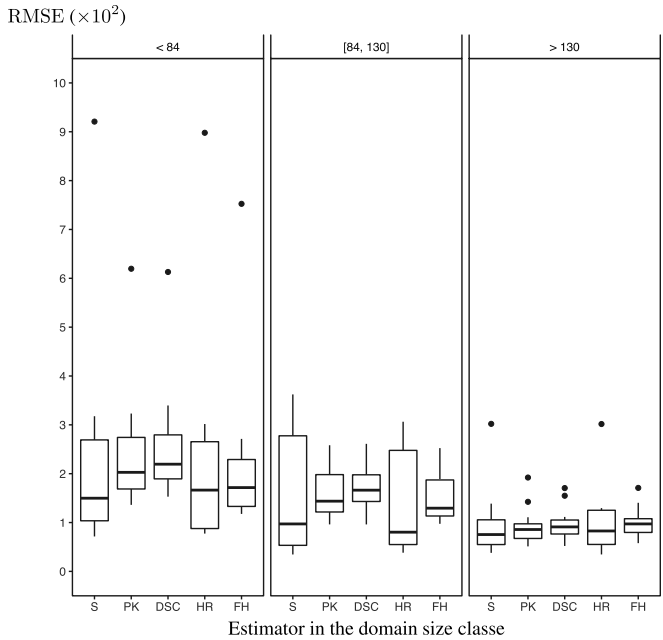

Figure 3. Boxplots of RMSEs of estimators in domain size classes as $n = 5000$.

4 Conclusions

The adaptive version of the composite estimation of [8] may be useful for routine applications, where there is a larger number of survey variables. For repeated surveys, non-adaptive alternatives to the proposed composite estimator could be constructed. In that case, the unknown parameter would be estimated treating a sample of the previous period as an independent data source. Comparing our simulation results with that of [8, 26], the present conclusion is that there is no universal prior choice of the parameter value even for similar surveys.

The new adaptive composite estimators, introduced in Section 2.4, strongly depend on the underlying regression model. Therefore, the accuracy of these generalized compositions should be similar to that of the regression-synthetic estimators. Indeed, considering the simulation results, we observe the improvement of the synthetic estimators, but a common gain does not become significant if the synthetic estimators are relatively inefficient.

From additional simulation tests we conclude that both empirical risk functions give similar results in the case of the generalized composition, and thus, the use of the simplified risk is sufficient, while the evaluation of the empirical risk function is computationally intensive in the adaptive SSD estimation. On the other hand, the construction of weights of the generalized compositions is complex itself.

The choices of (empirical) risk functions are not unique. The functions should be consistent with the accuracy characteristics of estimators we are interested in. For instance, if

inferences are based on the relative RMSEs, then the considered empirical risk functions should be properly normalized by the corresponding estimators.

Applying the composite estimators in Lithuanian LFS, their success depends on the overall sample size. The results of the proposed generalized composite estimator are very good for smaller domains and where other composite estimators, including EBLUP, are less efficient due to the small overall sample size. In some situations, EBLUP of [9] yields quite reliable estimates for smaller domains, while the synthetic estimation appears efficient for large domains.

An efficient estimation of MSEs of the considered synthetic and composite estimators is complicated in general. There exist only several universal proposals and it seems that there are no new ones in recent literature, except the case of popular EBLUP. The simulation of LFS shows that the biases of estimators are neither negligible nor similar across the domains. Therefore, the empirical MSEs like (20) and (21) may be insufficient for applications. We believe that the estimation of errors can be improved by applying data-specific modeling, but it is outside the scope of the paper.

Acknowledgment. Additional knowledge on the Lithuanian LFS data were obtained thanks to Tomas Rudys at Statistics Lithuania. I am also grateful to three anonymous reviewers for their comments and suggestions.

References

1. E. Bartoloni, Small area estimation and the labour market in Lombardy's industrial districts: A methodological approach, *Scienze Regionali*, **7**(2):27–54, 2008.
2. H.J. Boonstra, J.A. van den Brakel, B. Buelens, S. Krieg, M. Smeets, Towards small area estimation at Statistics Netherlands, *Metron*, **66**(1):21–49, 2008.
3. J.G. Booth, R.W. Butler, P. Hall, Bootstrap methods for finite populations, *J. Am. Stat. Assoc.*, **89**(428):1282–1289, 1994.
4. A. Čiginas, Domains similarity models in synthetic estimation, *Stat. J. IAOS*, **31**(1):51–57, 2015.
5. G.S. Datta, M. Ghosh, Small area shrinkage estimation, *Stat. Sci.*, **27**(1):95–114, 2012.
6. G.S. Datta, J.N.K. Rao, D.D. Smith, On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**(1):183–196, 2005.
7. P. Dick, Modelling net undercoverage in the 1991 Canadian census, *Surv. Methodol.*, **21**(1): 45–54, 1995.
8. J.D. Drew, M.P. Singh, G.H. Choudhry, Evaluation of small area estimation techniques for the Canadian Labour Force Survey, *Surv. Methodol.*, **8**:17–47, 1982.
9. R.E. Fay, R.A. Herriot, Estimates of income for small places: an application of James–Stein procedures to census data, *J. Am. Stat. Assoc.*, **74**(366):269–277, 1979.
10. M.E. Gonzalez, Use and evaluation of synthetic estimators, in *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, 1973, pp. 33–36.

11. M.E. Gonzalez, C. Hoza, Small-area estimation with application to unemployment and housing estimates, *J. Am. Stat. Assoc.*, **73**(361):7–15, 1978.
12. B. Hulliger, *Nonparametric M-Estimation of a Population Mean*, PhD thesis, ETH, Zürich, 1991.
13. A.-R. Khoshgooyanfard, M.T. Monazzah, A cost-effective strategy for provincial unemployment estimation: A small area approach, *Surv. Methodol.*, **32**(1):105–114, 2006.
14. D. Krapavickaitė, T. Rudys, Small area estimates for the fraction of the unemployed, *Lith. Math. J.*, **55**(2):243–254, 2015.
15. R. Lehtonen, C.-E. Särndal, A. Veijanen, The effect of model choice in estimation for domains, including small domains, *Surv. Methodol.*, **29**(1):33–44, 2003.
16. P.S. Levy, The use of mortality data in evaluating synthetic estimates, in *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, 1971, pp. 328–331.
17. Statistics Lithuania, *Public data files. Labour Force Survey*, `https://osp.stat.gov.lt/viesos-duomenu-rinkmenos`. [Online; accessed 2019-06-27]
18. N.T. Longford, *Missing Data and Small-Area Estimation*, Springer-Verlag, New York, 2005.
19. E. López-Vizcaíno, M.J. Lombardía, D. Morales, Small area estimation of labour force indicators under a multinomial model with correlated time and area effects, *J. R. Stat. Soc., Ser. A*, **178**(3):535–565, 2015.
20. D.A. Marker, *Small Area Estimation: A Bayesian Perspective*, PhD thesis, University of Michigan, Ann Arbor, 1995. Unpublished
21. I. Molina, E. Strzalkowska-Kominiak, Estimation of proportions in small areas: application to the labour force using the Swiss Census Structural Survey, *J. R. Stat. Soc., Ser. A*, **183**(1):281–310, 2020.
22. N.J. Purcell, L. Kish, Estimation for small domains, *Biometrics*, **35**:365–384, 1979.
23. T. Randrianasolo, Y. Tillé, Small area estimation by splitting the sampling weights, *Electron. J. Stat.*, **7**:1835–1855, 2013.
24. J.N.K. Rao, I. Molina, *Small Area Estimation,* 2nd ed., John Wiley, New Jersey, 2015.
25. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.
26. M.D. Ugarte, T. Goicoa, A.F. Militino, M. Sagaseta-López, Estimating unemployment in very small areas, *SORT*, **33**(1):49–70, 2009.
27. K.M. Wolter, *Introduction to Variance Estimation,* 2nd ed., Springer-Verlag, New York, 2007.
28. L.-C. Zhang, R.L. Chambers, Small area estimates for cross-classifications, *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, **66**(2):479–496, 2004.