Nonlinear Analysis: Modelling and Control, Online First, 1–15
https://doi.org/10.15388/namc.2023.28.32197

Vilnius University Press

Design-based composite estimation of small proportions in small domains

Andrius Čiginas

Institute of Data Science and Digital Technologies,
Vilnius University,
Akademijos str. 4, LT-08412 Vilnius, Lithuania
andrius.ciginas@mif.vu.lt

Received: March 11, 2022 / **Revised:** April 27, 2023 / **Published online:** May 10, 2023

Abstract. Traditional direct estimation methods are inefficient for domains of a survey population with small sample sizes. To estimate the domain proportions, we combine the direct estimators and the regression-synthetic estimators based on domain-level auxiliary information. For the case of small true proportions, we propose the design-based linear combination that is a robust alternative to the empirical best linear unbiased predictor (EBLUP) based on the Fay–Herriot model.

We imitate the Lithuanian Labor Force Survey, where we estimate the proportions of the unemployed and employed in municipalities. We show where the proposed design-based composition and estimator of its mean square error are competitive for EBLUP and its accuracy estimation.

Keywords: small area estimation, area-level model, composite estimator, sample-size-dependent estimator, Labor Force Survey.

1 Introduction

In the classical survey statistics by Särndal et al. [26], design-based and model-assisted direct estimators of parameters rely only on the sample of the estimation domain (area). Therefore, after the sample is selected, their application for some unplanned domains leads to high variances of the estimators because of too small sample sizes. In the small area estimation theory by Rao and Molina [23], indirect estimators borrow sample information from neighbor domains through auxiliary information and linking models. These model-based estimators usually have lower variances than direct estimators, but their biases may be relatively large.

To estimate proportions in the domains, one can consider explicit linking models based on auxiliary data aggregated to the domain level. A popular model is the Fay–Herriot (FH) model, which is a separate case of linear mixed models, and Fay and Herriot

[9] derived from it the empirical best linear unbiased predictors (EBLUPs) of the domain means. This small area predictor is expressed as the linear combination of a regression-synthetic estimator and the direct estimator. While the former part accounts for a variation reflected in the auxiliary data, the direct component exploits the unbiasedness property. Compositions of the synthetic and the direct estimators constitute an important class of indirect estimators. Before the mixed models, traditional design-based composite estimators were often used [23, Chap. 3]. However, now it is accepted that the models including random area-specific effects are more useful. One of the reasons is that they are more convenient for handling complex data structures than the traditional estimators with only randomness induced by the sampling design. Some examples of complex models applied to the estimation of proportions are in [2,8,11,18,19], see also the book of Sugasawa and Kubokawa [27]. Another notable drawback of traditional estimators is the difficulty in estimating their precision. The problem is with bias estimation, while it is well developed for the estimators like EBLUP.

Small area estimation problems differ from classical survey statistics in that they require more advanced statistical techniques to produce precise estimates for small domains. It involves using more auxiliary data and complex models and evaluating potential biases in the estimates. For model-based estimators like EBLUP, the model is typically used to make inferences about the population. We focus on the design-based small area estimation approach, where the estimation task for inferences is similar to that in the classical theory: only the sample design is taken into account to produce estimates of parameters and evaluate their uncertainty. These estimators should be the first ones tested in any survey before applying more complex model-based estimation methods, as argued by Tzavidis et al. [28]. Moreover, the relatively simple design-based estimators may be the final choice in the survey if their accuracy meets the set requirements.

We use a conditional analysis to construct the design-based composite estimator, which is similar to EBLUP of [9] in some sense. According to the construction, it is a robust estimator suitable for small or large domain proportions. We compare the proposed estimator with the model-based EBLUP and the design-based sample-size-dependent (SSD) composition introduced by Drew et al. [7] and optimized with respect to its free parameter by Čiginas [3]. The MSEs of both the design-based compositions are estimated as suggested by Čiginas [4].

We compare the estimators and their MSE estimators in the simulation study using the Lithuanian Labor Force Survey (LFS) data, where fractions of the unemployed and employed are the proportions of interest estimated in municipalities. The applications of EBLUPs to LFS unemployment data are found, for example, in [1,12,13,16,17,21]. SSD compositions, with subjectively chosen values of the parameter, are used in [7, 29]. An adaptive selection of values of this parameter is applied to estimate the proportions of unemployed in [3].

We introduce the standard direct estimation and recall the famous EBLUP based on the domain-level model in Sections 2 and 3, respectively. We discuss the problems of design-based composite estimators in Section 4 and construct the new composition in Subection 4.2. We present the simulation study in Section 5 and conclude in Section 6.

2 Basic assumptions and direct estimation

The set $\mathcal{U} = \{1, \dots, N\}$ consists of the labels of elements of the survey population. Let y be a binary study variable with the fixed values $y_1, \dots, y_N$ assigned to the corresponding elements. To estimate the proportions in the population and its subsets, the sample $s \subset \mathcal{U}$ of size $n < N$ is drawn by the sampling design $\mathrm{p}(\cdot)$, and $\pi_k = \mathbf{P}_\mathrm{p}\{k \in s\} > 0$, $k \in \mathcal{U}$, are inclusion into the sample probabilities. Here the symbol $\mathbf{P}_\mathrm{p}$, and hereafter $\mathbf{E}_\mathrm{p}$, $\mathbf{Var}_\mathrm{p}$, and MSE_p denote probability, expectation, variance, and MSE according to $\mathrm{p}(\cdot)$, respectively. The characteristic $\mathbf{Var}_\mathrm{p}(\cdot)$ is called the sampling variance or design variance.

Let $\mathcal{U} = \mathcal{U}_1 \cup \dots \cup \mathcal{U}_M$ be the partition of the population into the nonoverlapping domains, where the domain $\mathcal{U}_i$ contains N_i elements. Then the domain sample $s_i = s \cap \mathcal{U}_i$ is of size $n_i \leqslant N_i$. We aim to estimate the proportions

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

where the numbers N_i are assumed to be known. If the design $\mathrm{p}(\cdot)$ does not ensure the fixed sizes n_i, then they can be too small to get sufficiently accurate direct estimates $\hat{\theta}_i^\mathrm{d}$ of (1). The accuracy measure we use for any design-based estimator $\hat{\theta}_i$ of θ_i is

$$\mathrm{MSE}_\mathrm{p}(\hat{\theta}_i) = \big(\mathbf{E}_\mathrm{p}(\hat{\theta}_i) - \theta_i\big)^2 + \mathbf{Var}_\mathrm{p}(\hat{\theta}_i),$$

where the first term means the squared bias. While this term is typically negligible for direct estimators, it can be substantial for other small area estimators.

Assume that, for each domain $\mathcal{U}_i$, the auxiliary information is available as the vector of known characteristics $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iP})'$. This assumption narrows a choice of direct estimators $\hat{\theta}_i^\mathrm{d}$ to the classical design unbiased Horvitz–Thompson estimators $\hat{\theta}_i^\mathrm{HT} = N_i^{-1} \sum_{k \in s_i} y_k/\pi_k$ of θ_i or the weighted sample proportions

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

which are approximately unbiased. The latter estimators are also known as Hájek estimators. The approximate sampling variances of (2) and their estimators have the expressions [26, p. 185]

$$\mathbf{Var}_\mathrm{p}\big(\hat{\theta}_i^\mathrm{H}\big) \approx \psi_i^\mathrm{H} = \frac{1}{N_i^2} \sum_{k \in \mathcal{U}_i} \sum_{l \in \mathcal{U}_i} (\pi_{kl} - \pi_k \pi_l) \frac{(y_k - \theta_i)(y_l - \theta_i)}{\pi_k \pi_l} \tag{3}$$

and

$$\hat{\psi}_i^\mathrm{H} = \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} \sum_{l \in s_i} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{(y_k - \hat{\theta}_i^\mathrm{H})(y_l - \hat{\theta}_i^\mathrm{H})}{\pi_k \pi_l}, \quad i = 1, \dots, M, \tag{4}$$

respectively, where $\pi_{kl} = \mathbf{P}_\mathrm{p}\{k, l \in s\} > 0$ is the probability that both elements k and l will be included in the sample.

3 EBLUP under the Fay–Herriot model

The direct estimators $\hat{\theta}_i^{\mathrm{d}}$ of the domain proportions can be improved using the FH model [9]. The data for this domain-level model are the estimates $\hat{\theta}_i^{\mathrm{d}}$, their corresponding estimates $\hat{\psi}_i$ of the sampling variances $\psi_i = \mathbf{Var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$, and the covariates $\mathbf{z}_i$, $i = 1, \dots, M$. The basic FH model consists of two parts, see [23, Sect. 4.2], that are combined into the linear mixed model

$$\hat{\theta}_i^{\mathrm{d}} = \mathbf{z}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \dots, M, \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)'$ is the vector of fixed effects, the sampling errors ε_i are assumed independent with $\mathbf{E}_{\mathrm{p}}(\varepsilon_i) = 0$ and $\mathbf{Var}_{\mathrm{p}}(\varepsilon_i) = \psi_i$, and random domain effects v_i are assumed independent of these errors. The latter effects are supposed to be independent and identically distributed with $\mathbf{E}(v_i) = 0$ and $\mathbf{Var}(v_i) = \sigma_v^2 \geqslant 0$ with respect to a distribution, different from that generated by the design $\mathrm{p}(\cdot)$.

Treating the estimates $\hat{\psi}_i$ as given numbers, the method of EBLUP leads to the predictors of proportions (1) that are expressed as the linear combinations [9]

$$\hat{\theta}_i^{\mathrm{FH}} = \hat{\theta}_i^{\mathrm{FH}}(\hat{\psi}_i) = \hat{\gamma}_i\hat{\theta}_i^{\mathrm{d}} + (1 - \hat{\gamma}_i)\mathbf{z}_i'\hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\psi}_i + \hat{\sigma}_v^2}, \quad i = 1, \dots, M, \tag{6}$$

and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \frac{\mathbf{z}_i\mathbf{z}_i'}{\hat{\psi}_i + \hat{\sigma}_v^2} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{z}_i\hat{\theta}_i^{\mathrm{d}}}{\hat{\psi}_i + \hat{\sigma}_v^2},$$

where $\hat{\sigma}_v^2$ is an estimator of the variance σ_v^2. One of the ways to estimate σ_v^2 is the estimator $\hat{\sigma}_v^2$ based on the method of moments proposed by Fay and Herriot [9]. For this estimator, approximately unbiased estimators of $\mathrm{MSE}(\hat{\theta}_i^{\mathrm{FH}}) = \mathbf{E}(\hat{\theta}_i^{\mathrm{FH}} - \theta_i)^2$ were derived by Datta et al. [5]:

$$\begin{aligned} \mathrm{mse}\big(\hat{\theta}_i^{\mathrm{FH}}\big) = \hat{\gamma}_i\hat{\psi}_i + (1 - \hat{\gamma}_i)^2 \Bigg[& \mathbf{z}_i' \left(\sum_{j=1}^{M} \frac{\mathbf{z}_j\mathbf{z}_j'}{\hat{\psi}_j + \hat{\sigma}_v^2} \right)^{-1} \mathbf{z}_i \\ & + \frac{4M}{\hat{\psi}_i + \hat{\sigma}_v^2} \left(\sum_{j=1}^{M} \frac{1}{\hat{\psi}_j + \hat{\sigma}_v^2} \right)^{-2} \\ & - 2\hat{\sigma}_v^2 \left(\sum_{j=1}^{M} \hat{\gamma}_j \right)^{-3} \left\{ M \sum_{j=1}^{M} \hat{\gamma}_j^2 - \left(\sum_{j=1}^{M} \hat{\gamma}_j \right)^2 \right\} \Bigg], \quad i = 1, \dots, M. \end{aligned} \tag{7}$$

Predictors (6) and their MSE estimators (7) also depend on the estimators $\hat{\psi}_i$ of the sampling variances ψ_i of $\hat{\theta}_i^{\mathrm{d}}$. However, direct estimators $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i, as, for example, approximately design unbiased estimators (4) of (3) for $\hat{\theta}_i^{\mathrm{H}}$, have large variances themselves for small sample sizes. Therefore, the direct estimates $\hat{\psi}_i^{\mathrm{d}}$ are smoothed, and new more stable estimates $\hat{\psi}_i^{\mathrm{s}}$ are used in (6) and (7). According to Wolter [30], it is called the

generalized variance function (GVF) approach. The specific example of the GVF method, similar to that used for estimation of census undercounts by Dick [6], is to assume that $\psi_i \approx KN_i^{\gamma}$ and estimate the parameters $K > 0$ and $\gamma \in \mathbb{R}$ using the regression model

$$\log\bigl(\hat{\psi}_i^{\mathrm{d}}\bigr) = \log(K) + \gamma \log(N_i) + \eta_i, \quad i = 1, \dots, M,$$

where errors η_i are independent and identically distributed. That is, the smoothed estimates

$$\hat{\psi}_i^{\mathrm{sD}} = \widehat{K} N_i^{\hat{\gamma}}, \quad i = 1, \dots, M, \tag{8}$$

of ψ_i are based on the ordinary least squares estimates of the regression parameters. A similar smoothing is considered in [31]. Other smoothing examples are pooled variance estimator [1] and a nonparametric smoothing like in [12]. Despite the smoothing, estimators (7) tend to underestimate MSEs of (6) because the estimation of the sampling variances ψ_i is ignored in the derivation of (7).

4 Design-based composite estimation

4.1 Evaluation of optimal compositions and their accuracy estimation

Let us exclude the random effects v_i from FH model (5). Then this model, formulated for the Horvitz–Thompson estimators $\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{HT}}$ or the weighted sample proportions $\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{H}}$ specified by (2), becomes

$$\hat{\theta}_i^{\mathrm{d}} = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, M, \tag{9}$$

and, using the estimates $\hat{\psi}_i$ of the variances ψ_i, we arrive to the regression-synthetic estimators [23, Sect. 4.2]

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{S}}(\hat{\psi}_i) = \mathbf{z}_i'\hat{\boldsymbol{\beta}}, \quad i = 1, \dots, M, \tag{10}$$

of the domain proportions θ_i, where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \frac{\mathbf{z}_i \mathbf{z}_i'}{\hat{\psi}_i} \right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{z}_i \hat{\theta}_i^{\mathrm{d}}}{\hat{\psi}_i} \tag{11}$$

is the generalized least squares estimate of $\boldsymbol{\beta}$. Here, as for EBLUPs, the use of smoothed estimates $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$ instead of $\hat{\psi}_i^{\mathrm{d}}$ stabilizes synthetic estimators (10).

Estimators (10) rely on a synthetic assumption that the parameter $\boldsymbol{\beta}$ is the same across all domains. Therefore, having a good regression model, their design variances may be low compared to that of chosen direct estimators $\hat{\theta}_i^{\mathrm{d}}$ or even the EBLUPs $\hat{\theta}_i^{\mathrm{FH}}$. However, the design biases of (10) can be relatively large if the synthetic assumption is not realistic. To find a trade-off between larger variances of $\hat{\theta}_i^{\mathrm{d}}$ and biases of the synthetic estimators $\hat{\theta}_i^{\mathrm{S}}$, we consider their linear combinations

$$\tilde{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i) = \lambda_i \hat{\theta}_i^{\mathrm{d}} + (1 - \lambda_i)\hat{\theta}_i^{\mathrm{S}}, \quad i = 1, \dots, M, \tag{12}$$

with weights $0 \leqslant \lambda_i \leqslant 1$. Minimizing the function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\theta}_i^{\mathrm{C}}(\lambda_i))$ with respect to λ_i, the optimal weight for the domain $\mathcal{U}_i$ is the population parameter [23, Sect. 3.3]

$$\lambda_i^* = \frac{\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}}) - C_i}{\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}}) + \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}}) - 2C_i} \quad \text{with} \quad C_i = \mathbf{E}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{d}} - \theta_i\big)\big(\hat{\theta}_i^{\mathrm{S}} - \theta_i\big), \tag{13}$$

and then the theoretically optimal design-based linear combination is

$$\hat{\theta}_i^{\mathrm{opt}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i^*). \tag{14}$$

Assuming that $|C_i| \ll \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$, the approximation

$$\lambda_i^* \approx \frac{\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})}{\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}}) + \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})}$$

is applied, but the further difficulty is to evaluate the quantities $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$. A common approach to this is to use the representation [23, Sect. 3.2.5]

$$\mathrm{MSE}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{S}}\big) = \mathbf{E}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}\big)^2 - \mathbf{Var}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}\big) + \mathbf{Var}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{S}}\big), \tag{15}$$

where $\hat{\theta}_i^{\mathrm{d}}$ is assumed to be unbiased, and then to build an approximately design unbiased estimator

$$\mathrm{mse}_{\mathrm{u}}\big(\hat{\theta}_i^{\mathrm{S}}\big) = \big(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}\big)^2 - \hat{\sigma}^2\big(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}\big) + \hat{\sigma}^2\big(\hat{\theta}_i^{\mathrm{S}}\big) \tag{16}$$

of (15), where $\hat{\sigma}^2(\cdot)$ is an estimator of the design variance $\mathbf{Var}_{\mathrm{p}}(\cdot)$. Unfortunately, estimator (16) is typically unstable and can take negative values for individual small domains. Therefore, the straightforward estimation of optimal weights (13) is avoided.

To evaluate the optimal coefficients for compositions (12), one can set a common weight for all domains and then minimize a total MSE with respect to that weight by Purcell and Kish [22]. A similar approach is to apply James–Stein method [23, Sect. 3.4]. One more idea is SSD estimation by Drew et al. [7], where estimators of the weights in (12) are taken to be of the form

$$\hat{\lambda}_i = \hat{\lambda}_i(\delta) = \begin{cases} 1 & \text{if } \widehat{N}_i/N_i \geqslant \delta, \\ \widehat{N}_i/(\delta N_i) & \text{otherwise.} \end{cases} \tag{17}$$

These weights depend on the single subjectively chosen parameter δ for all domains. According to [23], a general-purpose choice of δ in (17) is $\delta = 1$. Särndal and Hidiroglou [25] derived similar SSD estimators by applying a conditional analysis.

Estimation of MSEs of the design-based composite estimators like these is treated as a difficult problem in the literature [23, Chap. 3]. One general solution is to consider the composition $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ as a synthetic estimator and use the estimator

$$\mathrm{mse}_{\mathrm{u}}\big(\hat{\theta}_i^{\mathrm{C}}\big) = \big(\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}}\big)^2 - \hat{\sigma}^2\big(\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}}\big) + \hat{\sigma}^2\big(\hat{\theta}_i^{\mathrm{C}}\big) \tag{18}$$

of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$, see [23, Ex. 3.3.1] and [3]. However, this estimator has the same drawbacks as (16). Another general method is to assume that the estimator $\hat{\theta}_i^{\mathrm{C}}$ defined by (12) approximates the optimal combination $\hat{\theta}_i^{\mathrm{opt}}$ quite well and derive the approximation [4]:

$$\mathrm{MSE}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{C}}\big) \approx \lambda_i\big(1-\lambda_i\big)\psi_i + \mathbf{Var}_{\mathrm{p}}\big(\hat{\theta}_i^{\mathrm{C}}\big)$$

with the empirical version

$$\mathrm{mse}_{\mathrm{b}}\big(\hat{\theta}_i^{\mathrm{C}}\big) = \hat{\lambda}_i\big(1-\hat{\lambda}_i\big)\hat{\psi}_i + \hat{\sigma}^2\big(\hat{\theta}_i^{\mathrm{C}}\big), \tag{19}$$

where we would set $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$ to have MSE estimators, which are less sensitive to the outliers (due to small sample sizes) than those using the direct estimators $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{d}}$. Estimator (19) takes only nonnegative values.

4.2 New composite estimation

The sampling variance ψ_i is approximately proportional to the product $\theta_i(1-\theta_i)$. That is, one can use the approximation

$$\psi_i \approx \frac{D_i\theta_i(1-\theta_i)}{n_i}, \tag{20}$$

where D_i is the design effect reflecting the sample efficiency of the complex sampling design, according to Kish [14]. Then, inserting $\hat{\theta}_i^{\mathrm{d}}$ and an appropriate estimator $\widehat{D}_i$ of D_i into (20), we would approximate the direct estimator $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i.

Let us first suppose that the domain proportions θ_i are small, say $\theta_i < 0.1$. In this case, it is even more complicated to get reliable direct estimates and estimates of their accuracy [10, 15]. Because the smaller the true proportions, the larger the samples are needed to maintain the same accuracy of estimators. For example, the direct estimator of the proportion of the unemployed can take zero value even for a sample of moderate size in the municipality.

Consider two candidate estimators $\hat{\psi}_i^{\mathrm{d}}$ and $\hat{\psi}_i^{\mathrm{s}}$ of ψ_i used in regression-synthetic estimator (10). Assume that we got too small estimate $\hat{\theta}_i^{\mathrm{d}}$ of θ_i for the specific sample s. The direct estimate $\hat{\psi}_i^{\mathrm{d}}$ then underestimates the sampling variance ψ_i. Therefore, the inequality $\hat{\psi}_i^{\mathrm{s}} > \hat{\psi}_i^{\mathrm{d}}$ should often hold, that is, the smoothed variance $\hat{\psi}_i^{\mathrm{s}}$ could be a better choice than $\hat{\psi}_i^{\mathrm{d}}$. Now suppose that $\hat{\theta}_i^{\mathrm{d}}$ overestimated the parameter θ_i. Then $\hat{\psi}_i^{\mathrm{d}}$ overestimates ψ_i as well, and the inequality $\hat{\psi}_i^{\mathrm{s}} < \hat{\psi}_i^{\mathrm{d}}$ should hold if $\hat{\theta}_i^{\mathrm{d}}$ is an outlier. This larger estimate $\hat{\psi}_i^{\mathrm{d}}$ can be employed to down-weight the outlying observation $\hat{\theta}_i^{\mathrm{d}}$ used in (11) and thus synthetic estimators (10) are less sensitive to the outliers. From these considerations, we derive the combined estimators

$$\hat{\psi}_i^{\mathrm{c}} = \max\big\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{d}}\big\}, \quad i = 1, \dots, M,$$

of the sampling variances ψ_i that should improve the regression-synthetic estimators. Next, in line with the same ideas, we define the design-based composite estimators

$$\hat{\theta}_i^{\mathrm{C}} = \hat{\lambda}_i\hat{\theta}_i^{\mathrm{d}} + (1-\hat{\lambda}_i)\hat{\theta}_i^{\mathrm{S}}\big(\hat{\psi}_i^{\mathrm{c}}\big) \quad \text{with} \quad \hat{\lambda}_i = \frac{\min\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{d}}\}}{\hat{\psi}_i^{\mathrm{c}}}, \quad i = 1, \dots, M, \tag{21}$$

of domain proportions (1). If the estimate $\hat{\theta}_i^{\mathrm{d}}$ is an outlier by its too small or too large value, then relatively more weight is attached to the synthetic part of composition (21). The composition is a shrinkage estimator because it shrinks the direct estimator toward the synthetic one.

We apply the same arguments to create (21) if the parameters θ_i are not small, but then the inequalities $\max\{\theta_i, \hat{\theta}_i^{\mathrm{d}}\} < 1/2$ or $\min\{\theta_i, \hat{\theta}_i^{\mathrm{d}}\} > 1/2$ must be satisfied. If these inequalities are not valid, the composite estimator is still applicable, but it can be less efficient. The worst scenario here would be a large difference $\theta_i - \hat{\theta}_i^{\mathrm{d}}$ and the relation $\theta_i \approx 1 - \hat{\theta}_i^{\mathrm{d}}$ but such events are rare.

To estimate the MSE of composition (21), we suggest applying general estimator (19). We study the accuracy of both these estimators in Section 5.

4.3 Sample-size-dependent estimation

A choice of the parameter δ in (17) varies from survey to survey. That is, the values 2/3 and 1 are good for LFS in [7], the authors of [29] try the larger points 1.5 and 2 for their data, and optimal values of δ are even higher in [3]. Therefore, to select the value of the parameter for the composition $\tilde{\theta}_i^{\mathrm{C}}(\delta) = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i(\delta))$ defined by (12), the sample-based function

$$r(\delta) = \frac{1}{M} \sum_{i=1}^{M} \mathrm{mse}_{\mathrm{u}}\big(\tilde{\theta}_i^{\mathrm{C}}(\delta)\big) \tag{22}$$

is minimized numerically with respect to δ in [3]. The minimization is implemented by applying any univariate optimization algorithm. Function (22) is the average of individual MSE estimators (18) over domains and, therefore, it is stable, unlike the individual ones. Then the adaptive design-based composite estimators of the domain proportions are [3]

$$\hat{\theta}_i^{\mathrm{SSD}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\delta}^*) \quad \text{with} \quad \hat{\delta}^* = \arg\min_{\delta>0} r(\delta), \quad i = 1, \ldots, M. \tag{23}$$

We apply estimators (19) to evaluate the MSEs of these compositions.

5 Simulations using the Labor Force Survey data

The main LFS variable is the categorical one that indicates an individual's participation in the labor market. This variable is decomposed into three binary variables: the person is unemployed, employed, and not in the labor force. We estimate the proportions of the former two variables in the municipalities of Lithuania. To imitate the real survey, we construct the artificial population from the sample data of the fourth quarter of 2018 as follows: we remove municipalities with too small fractions of observed unemployed persons and then replicate the data of each individual the number of times equal to the rounded survey weight. The size of that population $\mathcal{U}$ is $N = 1\,396\,763$, and it contains $M = 30$ municipalities. In LFS, the sample of households is drawn without replacement with probabilities proportional to the number of their members, and then the selected households are surveyed entirely. We use the same sampling design to draw $R = 10^3$

independent samples of households of size $n' = 3\,700$. It yields the samples of persons of sizes close to $n = 7\,667$. Then, for the kth individual that belongs to the lth household of size h_l, we apply the approximation $\pi_k \approx h_l n'/N$, $k \in \mathcal{U}$.

We compare the following estimators of the domain proportion θ_i:

- the direct estimator $\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{H}}$ from (2);
- the regression-synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$ given by (10);
- EBLUP $\hat{\theta}_i^{\mathrm{FH}}$ in (6) calculated using the package `sae` for `R` by [20];
- the new design-based composition $\hat{\theta}_i^{\mathrm{C}}$ given by (21);
- SSD composite estimator $\hat{\theta}_i^{\mathrm{SSD}}$ from (23);
- the optimal combination $\hat{\theta}_i^{\mathrm{opt}}$ by (14).

Moreover, we compare the accuracy of these MSE estimators:

- the estimator $\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$ of the parameter $\mathrm{MSE}(\hat{\theta}_i^{\mathrm{FH}})$ from (7);
- the estimator $\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$ by (18) applied to composition (21);
- the estimator $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$ by (19) for estimator (21);
- the estimator $\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{SSD}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{SSD}})$ applying formula (18);
- the estimator $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{SSD}})$ of the parameter $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{SSD}})$ using formula (19);
- the estimator $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{opt}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{opt}})$ calculated by (19).

To model the direct estimates of the proportions of interest by (5) and (9), we use the municipality characteristics $\mathbf{z}_i = (1, z_{i2}, z_{i3}, z_{i4}, z_{i5}, z_{i6})'$, where z_{i2} is the proportion of registered unemployed individuals derived from the administrative Lithuanian Labor Exchange data, z_{i3} is the proportion of persons who, according to the register of the State Social Insurance Fund Board, paid the social contribution one month before they participated in the survey, z_{i4} is the proportion of males, and z_{i5} and z_{i6} are the proportions of individuals from age intervals 26–40 and 41–55, respectively.

Since the sampling fractions are small in the municipalities, we take $\pi_{kl} \approx \pi_k \pi_l$, $k \neq l$, and so approximate direct estimators (4) of sampling variances (3) by

$$\hat{\psi}_i^{\mathrm{H}} \approx \hat{\psi}_i^{\mathrm{d}} = \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} w_k (w_k - 1) \big(y_k - \hat{\theta}_i^{\mathrm{d}}\big)^2, \quad i = 1, \ldots, M,$$

where we write $w_k = 1/\pi_k$. Then we smooth these $\hat{\psi}_i^{\mathrm{d}}$ to obtain $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{sD}}$ according to (8) and use the smoothed estimates for (6), (7), (10), (14), (19), and in the synthetic parts of (23).

We apply the bootstrap method by Rao et al. [24] to evaluate the estimators of the design variances used in (18), (19), and (22). Let us estimate the variance for any estimator $\hat{\theta}_i$. The bootstrap procedure works as follows: (i) Draw a simple random sample of $m = n' - 1$ households with replacement from the sample of n' households. Let m_l^* be the number of times the lth household is selected, and then $\sum_{l=1}^{n'} m_l^* = m$. Define the bootstrap weights $w_l^* = n' m_l^* w_l / m$, $l = 1, \ldots, n'$. Calculate the bootstrap estimate $\hat{\theta}_i^*$ using the weights w_l^* in the formula for $\hat{\theta}_i$. (ii) Repeat step (i) B times independently to

obtain the estimates $\hat{\theta}_i^{*(b)}$, $b = 1, \dots, B$. Then

$$\hat{\sigma}^2(\hat{\theta}_i) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_i^{*(b)} - \bar{\theta}_i^*\right)^2, \quad \text{where} \quad \bar{\theta}_i^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_i^{*(b)},$$

is the bootstrap estimator of the design variance $\mathbf{Var}_{\mathrm{p}}(\hat{\theta}_i)$. We take $B = 200$.

We evaluate all estimators for each of the R samples and calculate approximations to their root mean squared errors (RMSEs) and absolute biases (ABs). It means we use the accuracy measures

$$\mathrm{RMSE}(\hat{\mu}_i) = \left(\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\mu}_i^{(r)} - \mu_i\right)^2\right)^{1/2} \quad \text{and} \quad \mathrm{AB}(\hat{\mu}_i) = \left|\frac{1}{R}\sum_{r=1}^{R}\hat{\mu}_i^{(r)} - \mu_i\right|, \quad (24)$$

where $\hat{\mu}_i^{(r)}$ is a realization of the specific estimator $\hat{\mu}_i$ of the parameter μ_i, based on the rth sample. We classify the municipalities by the expected domain sample size into three classes of equal size and calculate the averages of RMSEs and ABs over the domains of each class. We also present the averages of (24) over all municipalities as common accuracy indicators.

The results for the proportions of the unemployed and employed are presented in Tables 1 and 2, respectively. Let us use the superscripts of estimators to discuss the output. In both the tables, any indirect estimator of the proportions improves the direct one in the sense of RMSE, and theoretical composition opt is the best estimator. Among the indirect estimators, regression-synthetic estimator S has much larger design biases than compositions FH, C, and SSD. In Table 1, the averages of RMSEs over all domains for design-based composite estimators C and SSD are smaller than that for EBLUP FH. It is not valid for estimator C in Table 2 because the proportions of the employed are distributed near

Table 1. Average RMSEs and ABs of estimators for the unemployed proportions in domain size classes as $n \approx 7\,667$. The domain is small if its expected sample size $\bar{n}_i = \mathbf{E}_{\mathrm{p}}(n_i) < 116$, is medium for $116 \leqslant \bar{n}_i < 159$, and is large as $\bar{n}_i \geqslant 159$.

Estimator	Average RMSE ($\times 10^2$) Domain size class by $\bar{n}_i$				Average AB ($\times 10^2$) Domain size class by $\bar{n}_i$			
	any	small	medium	large	any	small	medium	large
$\hat{\theta}_i^{\mathrm{d}}$	2.4793	3.8540	2.4578	1.1259	0.0636	0.1200	0.0485	0.0223
$\hat{\theta}_i^{\mathrm{S}}$	1.8174	2.8950	1.5632	0.9940	1.3461	2.3656	1.0677	0.6050
$\hat{\theta}_i^{\mathrm{FH}}$	1.7857	2.6707	1.7156	0.9707	0.7349	1.4738	0.5496	0.1811
$\hat{\theta}_i^{\mathrm{C}}$	1.7511	2.6798	1.6838	0.8897	0.7951	1.4777	0.6130	0.2946
$\hat{\theta}_i^{\mathrm{SSD}}$	1.7529	2.7228	1.6162	0.9196	0.8649	1.4974	0.6928	0.4045
$\hat{\theta}_i^{\mathrm{opt}}$	1.4712	2.3804	1.2486	0.7846	0.7301	1.3978	0.5206	0.2720
$\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$	0.0223	0.0445	0.0173	0.0051	0.0180	0.0373	0.0128	0.0039
$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$	0.0708	0.1540	0.0491	0.0094	0.0263	0.0532	0.0215	0.0041
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$	0.0173	0.0371	0.0119	0.0030	0.0135	0.0296	0.0087	0.0021
$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{SSD}})$	0.0593	0.1164	0.0494	0.0120	0.0115	0.0290	0.0051	0.0005
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{SSD}})$	0.0257	0.0497	0.0210	0.0063	0.0172	0.0314	0.0153	0.0050
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{opt}})$	0.0098	0.0206	0.0064	0.0023	0.0050	0.0110	0.0027	0.0012

Table 2. Average RMSEs and ABs of estimators for the employed proportions in domain size classes as $n \approx 7\,667$. The domain is small if its expected sample size $\bar{n}_i = \mathbf{E}_{\mathrm{p}}(n_i) < 116$, is medium for $116 \leqslant \bar{n}_i < 159$, and is large as $\bar{n}_i \geqslant 159$.

Estimator	Average RMSE ($\times 10^2$) Domain size class by $\bar{n}_i$				Average AB ($\times 10^2$) Domain size class by $\bar{n}_i$			
	any	small	medium	large	any	small	medium	large
$\hat{\theta}_i^{\mathrm{d}}$	4.7718	6.9201	4.7104	2.6848	0.1516	0.2577	0.1395	0.0575
$\hat{\theta}_i^{\mathrm{S}}$	3.4061	4.9905	3.2215	2.0061	2.6481	4.1247	2.5006	1.3188
$\hat{\theta}_i^{\mathrm{FH}}$	3.3054	4.6679	3.1768	2.0716	1.7276	2.8992	1.6535	0.6302
$\hat{\theta}_i^{\mathrm{C}}$	4.2265	6.0532	4.1539	2.4724	0.4024	0.6893	0.4213	0.0967
$\hat{\theta}_i^{\mathrm{SSD}}$	3.2747	4.7502	3.1425	1.9314	1.6996	2.6130	1.6237	0.8622
$\hat{\theta}_i^{\mathrm{opt}}$	2.8602	4.1800	2.6261	1.7746	1.6026	2.5092	1.4717	0.8269
$\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$	0.0724	0.1281	0.0670	0.0221	0.0561	0.1070	0.0469	0.0143
$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$	0.0666	0.1371	0.0481	0.0144	0.0297	0.0590	0.0216	0.0086
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$	0.0535	0.1093	0.0406	0.0107	0.0102	0.0170	0.0120	0.0016
$\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{SSD}})$	0.1992	0.3628	0.1784	0.0563	0.0297	0.0674	0.0201	0.0014
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{SSD}})$	0.0655	0.1091	0.0697	0.0178	0.0442	0.0715	0.0491	0.0119
$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{opt}})$	0.0181	0.0374	0.0110	0.0059	0.0082	0.0155	0.0049	0.0043

the point $1/2$ if we look at the five-number summary $(0.379, 0.585, 0.634, 0.668, 0.766)$ for the true proportions.

MSE estimators (19) for design-based compositions C and SSD evidently improve estimators (18) and yield similar or even better results than MSE estimator (7) for FH. The best MSE estimation using (19) is obtained for optimal composition opt. Since composite estimators C and SSD only approximate the optimal one, their MSE estimators have larger errors. On the other hand, these errors are acceptable if compared with the results for FH.

Detailed information about RMSEs of estimators d, FH, C, SSD, and some selected MSE estimators presented in Figs. 1–2 supports conclusions derived from the tables.

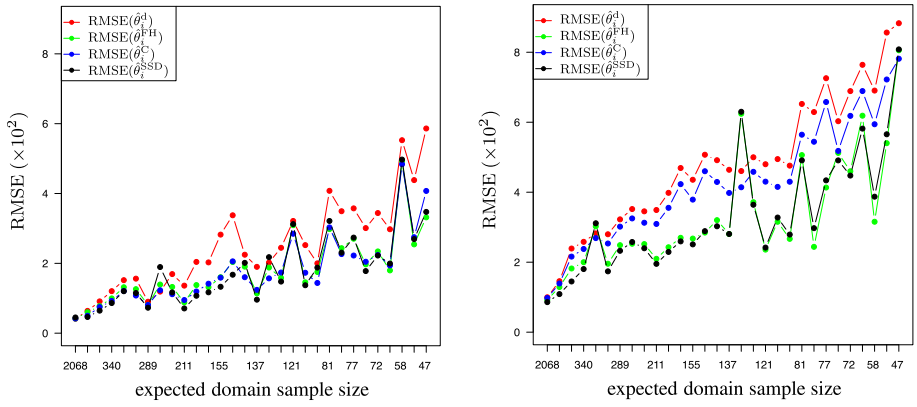

Figure 1. RMSEs of the estimators $\hat{\theta}_i^{\mathrm{d}}$, $\hat{\theta}_i^{\mathrm{FH}}$, $\hat{\theta}_i^{\mathrm{C}}$, and $\hat{\theta}_i^{\mathrm{SSD}}$ for the proportions of unemployed (a) and employed (b).

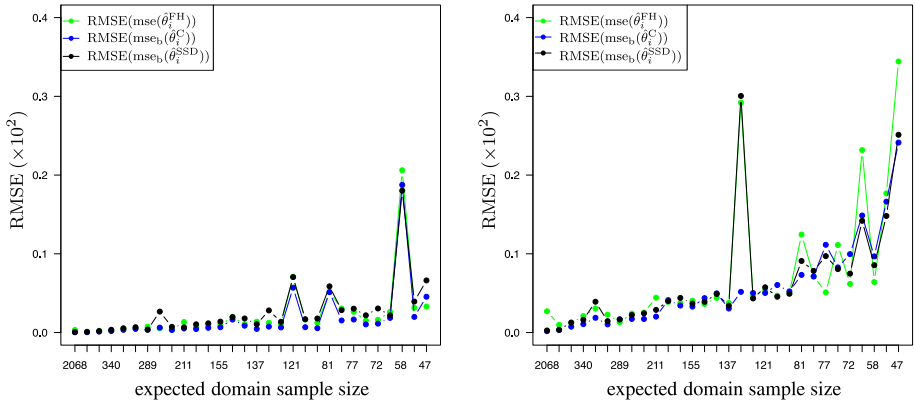

Figure 2. RMSEs of the MSE estimators $\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$, $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}})$, and $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{SSD}})$ for the proportions of unemployed (a) and employed (b).

The same experiment but with the twice smaller sample size $n' = 1\,850$ leads to similar conclusions. In this case, the proposed design-based composition C improves EBLUP FH more for small proportions.

6 Conclusions

The construction of new composite estimator (21) is based on the monotonicity of the variance of the direct estimator as the function of the proportion. Approximation (20) is the monotone function in two separate parts of the interval $[0, 1]$. Therefore, the composition loses its efficiency for the proportions close to the turning point $1/2$, where the monotonicity changes.

In general, the sampling variance of any direct estimator of the domain mean is not the monotone function of the target parameter. On the other hand, some GVF models from [30, p. 274] suggest that this function might be treated as an approximately monotonic one. Therefore, if we can find the GVF model that fits the data well and is the monotonic function, then estimator (21) could be applied to the domain means with this fitted model used instead of smoothed sampling variances (8).

The simulation study shows that the design-based compositions might be a good alternative to the classical EBLUP estimating proportions in small domains. Adaptive SSD composite estimator (23) works well for both unemployment and employment cases, while simpler new composition (21) is efficient for the unemployment fractions that are small proportions.

Design-based estimators and estimators of MSE under the design-based approach are desirable in practice [21]. That design MSE estimator (19) works well in our simulations, and its formula is simple compared to that of model MSE estimator (7) for EBLUP.

Acknowledgment. The confidential data used in the study were obtained from Statistics Lithuania (the State Data Agency), where the author is employed. The author is also grateful for two anonymous reviewers whose suggestions helped to improve the presentation of the article.

References

1. 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, https://www.researchgate.net/profile/Jan-Brakel/publication/227458249_Towards_small_area_estimation_at_Statistics_Netherlands/links/0c96052f8fdda8fdda8aedd000000/Towards-small-area-estimation-at-Statistics-Netherlands.pdf.
2. R. Chambers, N. Salvati, N. Tzavidis, Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK, *J. R. Stat. Soc., Ser. A, Stat. Soc.*, **179**(2):453–479, 2016, https://doi.org/10.1111/rssa.12123.
3. A. Čiginas, Adaptive composite estimation in small domains, *Nonlinear Anal. Model. Control*, **25**(3):341–357, 2020, https://doi.org/10.15388/namc.2020.25.16773.
4. A. Čiginas, Design-based composite estimation rediscovered, *Stat*, 2023, https://doi.org/10.1002/sta4.579.
5. 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, https://doi.org/10.1093/biomet/92.1.183.
6. P. Dick, Modelling net undercoverage in the 1991 Canadian census, *Surv. Methodol.*, **21**(1):45–54, 1995, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1995001/article/14411-eng.pdf?st=PM5ps205.
7. 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, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1982001/article/14328-eng.pdf?st=i7XnItQz.
8. M.D. Esteban, M.J. Lombardía, E. López-Vizcaíno, D. Morales, A. Pérez, Small area estimation of proportions under area-level compositional mixed models, *Test*, **29**(3):793–818, 2020, https://doi.org/10.1007/s11749-019-00688-w.
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, https://doi.org/10.1080/01621459.1979.10482505.
10. C. Franco, R.J.A. Little, T.A. Louis, E.V. Slud, Comparative study of confidence intervals for proportions in complex sample surveys, *J. Surv. Stat. Methodol.*, **7**(3):334–364, 2019, https://doi.org/10.1093/jssam/smy019.
11. C. Gonçalves, L. Hidalgo, D. Silva, J.A. van den Brakel, Single-month unemployment rate estimates for the Brazilian Labour Force Survey using state-space models, *J. R. Stat. Soc., Ser. A, Stat. Soc.*, **185**(4):1707–1732, 2022, https://doi.org/10.1111/rssa.12914.

12. W. González-Manteiga, M.J. Lombardía, I. Molina, D. Morales, L. Santamaría, Small area estimation under Fay–Herriot models with non-parametric estimation of heteroscedasticity, *Stat. Model.*, **10**(2):215–239, 2010, https://doi.org/10.1177/1471082X0801000206.

13. M.A. Hidiroglou, J.-F. Beaumont, W. Yung, Development of a small area estimation system at Statistics Canada, *Surv. Methodol.*, **45**(1):101–126, 2019, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf?st=Z2b0IXsL.

14. L. Kish, Methods for design effects, *J. Off. Stat.*, **11**(1):55–77, 1995, https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/methods-for-design-effects.pdf.

15. E.L. Korn, B.I. Graubard, Confidence intervals for proportions with small expected number of positive counts estimated from survey data, *Surv. Methodol.*, **24**(2):193–201, 1998, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1998002/article/4356-eng.pdf?st=0FMGYdWr.

16. D. Krapavickaitė, T. Rudys, Small area estimates for the fraction of the unemployed, *Lith. Math. J.*, **55**(2):243–254, 2015, https://doi.org/10.1007/s10986-015-9277-9.

17. 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, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6605-eng.pdf?st=mNBnbq6f.

18. 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, Stat. Soc.*, **178**(3):535–565, 2015, https://doi.org/10.1111/rssa.12085.

19. M.F. Marino, M.G. Ranalli, N. Salvati, M. Alf, Semiparametric empirical best prediction for small area estimation of unemployment indicators, *Ann. Appl. Stat.*, **13**(2):1166–1197, 2019, https://doi.org/10.1214/18-AOAS1226.

20. I. Molina, Y. Marhuenda, sae: An R package for small area estimation, *R J.*, **7**(1):81–98, 2015, https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf.

21. I. Molina, E. Strzalkowska-Kominiak, Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey, *J. Roy. Statist. Soc. Ser. A*, **183**(1):281–310, 2020, https://doi.org/10.1111/rssa.12498.

22. N.J. Purcell, L. Kish, Estimation for small domains, *Biometrics*, **35**:365–384, 1979, https://doi.org/10.2307/2530340.

23. J.N.K. Rao, I. Molina, *Small Area Estimation,* 2nd ed., John Wiley & Sons, Hoboken, NJ, 2015, https://doi.org/10.1002/9781118735855.

24. J.N.K. Rao, C.F.J. Wu, K. Yue, Some recent work on resampling methods for complex surveys, *Surv. Methodol.*, **18**(2):209–217, 1992, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf?st=TmCBmezi.

25. C.-E. Särndal, M.A. Hidiroglou, Small domain estimation: A conditional analysis, *J. Am. Stat. Assoc.*, **84**(405):266–275, 1989, https://doi.org/10.2307/2289873.

26. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer, New York, 1992, https://link.springer.com/book/9780387406206.

27. S. Sugasawa, T. Kubokawa, *Mixed-Effects Models and Small Area Estimation*, Springer, Singapore, 2023, https://doi.org/10.1007/978-981-19-9486-9.

28. N. Tzavidis, L.-Ch. Zhang, A. Luna, T. Schmid, N. Rojas-Perilla, From start to finish: A framework for the production of small area official statistics, *J. R. Stat. Soc., Ser. A, Stat. Soc.*, **181**(4): 927–979, 2018, https://doi.org/10.1111/rssa.12364.

29. M.D. Ugarte, T. Goicoa, A.F. Militino, M. Sagaseta-López, Estimating unemployment in very small areas, *SORT*, **33**(1):49–70, 2009, https://academica-e.unavarra.es/xmui/bitstream/handle/2454/10871/716.pdf?sequence=1&isAllowed=y.

30. K.M. Wolter, *Introduction to Variance Estimation,* 2nd ed., Springer, New York, 2007, https://doi.org/10.1007/978-0-387-35099-8.

31. Y. You, Small area estimation using Fay–Herriot area level model with sampling variance smoothing and modeling, *Surv. Methodol.*, **47**(2):361–370, 2021, https://www150.statcan.gc.ca/n1/pub/12-001-x/2021002/article/00007-eng.pdf.