A solution in small area estimation problems

Andrius Ėginas and Tomas Rudys
Vilnius University Institute of Mathematics and Informatics, LT-08663 Vilnius, Lithuania

Abstract

We present a new method in problems where estimates are needed for finite population domains with small or even zero sample sizes. Our estimation approach can be interpreted as the so-called model-based with properties considered from the point of design-based estimation. In particular, we use the unit-level auxiliary information and regression models to introduce and formalize a new concept of a hidden randomness (HR) of a finite population. The approach presented can be also viewed as an additional characterization of regression models. It helps to describe a distance between the domain of interest and population elements outside the domain. Consequently, we construct estimators which use an information of the whole sample. With simple estimators of the domain total, we give estimators of their mean square errors (MSEs). We present also a regression type HR estimator which reduces an impact of design bias to the MSE. In contrast to other similar small area estimation (SAE) methods, the estimators introduced tend to be design unbiased as the estimation domain size increases, and thus they can be effective for any population domain with any sample size including the case of empty sample. An efficiency of the method proposed is illustrated by a simulation study.

Keywords: small area estimation, auxiliary information, linear regression, order statistics.
MSC classes: 62D05 (62J05)

1 Introduction

The topic SAE becomes more and more actual and thus popular in the last decades because of a need to get more inferences about survey populations than it seems possible with given samples or after a planning of them. The established term ‘SAE’ is not correct enough because, in reality, it means an estimation in the finite population domain where sample size is too small to get estimates of an admissible quality. It is a very common situation when we need to plan the sample to estimate in the number of the population domains which are determined by very various classifications of the population and which are intersecting in various ways. Clearly, a sampling design ensuring a sample in each of the intersections leads to the large total sample size and thus increases a cost of the survey. Next, if samples obtained in domains are small, applications of the classical estimation theory usually fail, in the sense of estimates quality, and such a failure not so much depends on a good auxiliary information availability. In SAE, auxiliary information plays a crucial role similarly as it is important in the traditional survey sampling where it is extensively used via ratio, regression, calibration estimators, etc., see, e.g., [2, 15, 6, 7]. The main difference is that, in SAE methods, the auxiliary data are used to extract an information about the variable (or parameter) of interest from the sample outside the domain (so-called indirect estimation), while, in the classical case, in fact, the information of the domain of interest is incorporated only into estimation (direct estimation).

The research of the first author is partially supported by European Union Structural Funds project "Postdoctoral Fellowship Implementation in Lithuania".
SAE methods can be classified into two large groups: design-based and model-based methods. The design-based methodologies are close to an understanding of the classical estimation. Data models, included into an estimation of this type, are advisedly used to keep the basic properties of the resulting estimators such as consistency and at least approximate unbiasedness with respect to the sampling design, i.e., here a randomness is considered by the distribution of different samples appearance. A validity of the underlying model is not so much important, i.e., if the model is an incorrect then the variance of the estimator can be left unreduced, but, in any case, consistency and an acceptable unbiasedness of the estimator still hold. Here the underlying model is therefore called assisting. See [10] and see also [12], for a wide review of the design-based methods. An unbiasedness of design-based estimators has its cost, i.e., for the population domains with markedly small sample size the variances of estimators are often large. Then estimators from the class of model-based estimators can be a better choice, as it is pointed in [14]. The number of model-based methods are collected in [13], see also [12] and [4] on a recent results. We also refer to review papers [8] and [11]. Differently from the design-based methods, the MSE of a model-based estimator is defined and estimated with a respect to the model. An estimation approach, which is presented in the paper, can be linked to both classes of SAE methods. A similar classification problem arises in the cases of popular synthetic and composite estimators, empirical best linear unbiased predictor (EBLUP) type estimators. This is because these and our estimators are usually design biased, and it is not a desirable feature in the design-based estimation. On the other hand, the estimators do not use explicit models, and it is not a typical situation in the model-based approach. The variances of these estimators can be small even for domains with very small sample sizes but the design biases can be relatively large as it is common in model-based estimation. Therefore, in the analysis of our estimators, we are oriented to have good estimators of biases, and we aim to control and reduce them. As in design-based estimation, we assume the randomness by the sampling design, i.e., in the paper, the operators \( \mathbf{P}, \mathbf{E}, \mathbf{Var} \) and \( \mathbf{Cov} \) mean the probability, expectation, variance and covariance according to the sampling design, respectively.

The auxiliary information can be of two types: aggregated and unit-level auxiliary data. Here, assuming unit-level auxiliary information, we suppose that, for instance, statistical business register, administrative register or human population census register data are available. Clearly, unit-level data can be used more flexible, see, e.g., [16] and [9] in the context of the design-based estimation.

Let us introduce our technical framework. Consider a finite population \( \mathcal{U} = \{1, \ldots, N\} \) of size \( N \). Assume that, in order to estimate a parameter (characteristic) of the population, the sample \( s = \{i_1, \ldots, i_n\} \) of size \( n \) is drawn from \( \mathcal{U} \), according to a sampling design \( p(\cdot) \). Here \( p(s) \) is the probability to get the particular \( s \). Let \( \pi_i = \mathbf{P}\{i \in s\} \) and \( \pi_{ij} = \mathbf{P}\{i, j \in s\} \) be the inclusion into the sample probabilities for the population element \( i \) and for the pair of elements \( i \) and \( j \). We assume that these probabilities are strictly positive. It is usual situation when a construction of the sample design \( p(\cdot) \) is closely related to an auxiliary variable, say, \( z \) with values \{\( z_1, \ldots, z_N \)\} known for all units of \( \mathcal{U} \). In particular, e.g., for a stratified simple random sampling design and for probability proportional-to-size sampling the inclusion probability \( \pi_i \) represents an importance of the population unit \( i \) by the relative size of \( z_i \). We assume that, without loss of generality, \( z_1 \leq \cdots \leq z_N \). Let \( y \) be the variable of interest with the fixed values \{\( y_1, \ldots, y_N \)\} in the population, and, we aim to estimate the total

\[
t_{y;D} = \sum_{i \in D} y_i
\]

where \( D \subseteq \mathcal{U} \) is any non-empty set. If the particular estimation domain \( D \) is known before the sample selection, then, as it is common in practice, the sampling design \( p(\cdot) \) is constructed in order to get a sufficient (for quality requirements of estimates) sample size in that domain. But if we
are interested in $D$ after sample selection and collection of the data $y_{i1}, \ldots, y_{in}$, the sample size in $D$ can be too small or even equal zero and this leads to get bad estimates in the sense of their quality, if the estimators of (1) are usual direct Horvitz–Thompson (H–T), generalized regression (GREG) estimators, etc. Assume that at the estimation stage, we have an auxiliary variable $x$ and all its values $\{x_1, \ldots, x_N\}$ are known. Let these values be the realization of independent random variables $X_1, \ldots, X_N$ modeled by the linear regression model

$$X_i = \alpha_1 + \alpha_2 z_i + \delta_i, \quad i = 1, \ldots, N,$$

which we call $\eta$. Here $X_i$ has the distribution function $F_{\eta}(\cdot)$ with $E_{\eta} X_i = \alpha_1 + \alpha_2 z_i$ and $\text{Var}_{\eta} X_i = \tau^2_i$. Here $E_{\eta}$ and $\text{Var}_{\eta}$ denote expectation and variance with respect to the model $\eta$. Thus we assume a superpopulation model, where $\alpha_1, \alpha_2$ and $\tau_1, \ldots, \tau_N$ are unknown model parameters. A less general and more applicable specification of (2) would be, for instance, $\tau^2_i = \tau^2 z_i^\gamma$, $\tau > 0$, $z_i > 0$, with known parameter $\gamma \geq 0$, and with $F_{\eta}(\cdot)$ from the family of Gaussian distributions. In practice, the meaning of variable $x$ can be, for instance:

(i) $x$ is simply a better predictor of $y$ than $z$, particularly in cases where the sample survey has more variables of interest and $z$ was important for all of them at the sample planning moment;

(ii) $x$ is observed in the time period between the sample selection and estimation, e.g., we get it from administrative data sources.

Then it is also meaningful to incorporate a relation between $y$ and $x$ into the estimation process. But first, in Section 2 we assume only relation (2) where, for instance:

(iii) $x$ can have the same definition as $z$ but it is observed at the different moment of time than $z$, and thus their values are different;

(iv) definitions of $z$ and $x$ slightly differ;

(v) $x$ is an alternative measure of size of the population units which was not used at the sample planning stage;

(vi) $z$ can be measured roughly (quickly by a different way) and so we get $x$, which can be modeled by $X_i = z_i + \delta_i, i = 1, \ldots, N$, where $\delta_i$ represents a measurement error.

Thus, an origin of $x$ can be various, but we keep in mind that both $z$ and $x$ more or less represent a size of $y$. Our idea is to describe how the sizes of the population units ‘like’ to change and thus how the study variable $y$ ‘likes’ to change. To be more specific, the population values of $y$ remain fixed but, with a help of the model $\eta$, a dynamics of the sampling design is imitated through sizes of the population units. One can also ask: if $p_z(\cdot)$ and $p_x(\cdot)$ are two the same type sampling designs constructed by the variables $z$ and $x$, so which of them is better if $z$ and $x$ explain $y$ similarly. It seems that neither of them, and therefore we are motivated to think about an interplay between the actual design $p_z(\cdot)$ and the artificial design $p_x(\cdot)$. In Section 2 we introduce the HR method, construct the HR estimator and its quality estimators. Next, in Section 3 we assume that a relation between $y$ and $x$ is significant, and apply it to minimize the bias of HR estimator, without the loss of the accuracy in the sense of the HR estimator variance. In Section 4 we present a limited simulation study, where we compare the constructed HR estimators with the direct H–T, direct regression, synthetic and EBLUP estimators. We also show how the HR estimators behave where the estimation domain $D$ and its sample size are sufficiently large. In Section 5, we summarize our findings and discuss advantages and disadvantages of the estimation method proposed.
2 Method of hidden randomness and simple estimators

Let \( X_{(1)} \leq \cdots \leq X_{(N)} \) be the order statistics of random variables \( X_1, \ldots, X_N \) which are generated by model \( \mathbb{2} \). Denote the probabilities

\[
p_{ij} = P_{\eta} \{ X_i = X(j) \}, \quad i, j = 1, \ldots, N.
\]

These numbers are the parameters of the linear regression model \( \eta \). In particular, it is an additional (to the parameters \( \alpha_1, \alpha_2 \) and \( \tau_1, \ldots, \tau_N \)) characterization of the model. Characteristics \( \mathbb{3} \) describe a distribution of the given values of \( z \), while the basic model parameters, i.e., their (for instance, least-squares) estimates are, in fact, moments of \( z \). In our context, probabilities \( p_{ij}, j = 1, \ldots, N \), show how \( z_i \) ‘likes’ (tends) to variate as a size. Define the numbers

\[
t_j = \sum_{i \in \mathcal{U}} p_{ij} y_i, \quad j = 1, \ldots, N.
\]

Let \( \mathcal{D} = \{j_1, \ldots, j_m\} \subseteq \mathcal{U} \) be the domain of interest where sample size can be small or even equal zero, but not necessarily. Since for \( \mathcal{D} = \mathcal{U} \), we have \( t_{y;\mathcal{D}} = \sum_{i \in \mathcal{U}} y_i = \sum_{j \in \mathcal{U}} t_j \), we assume that

\[
t_{y;\mathcal{D}} \approx \sum_{j \in \mathcal{D}} t_j = \sum_{i \in \mathcal{U}} \theta_{i;\mathcal{D}} y_i, \quad \text{where} \quad \theta_{i;\mathcal{D}} = \sum_{j \in \mathcal{D}} p_{ij}.
\]

Here \( \theta_{i;\mathcal{D}} \) can be interpreted as the probability that the population element \( i \) should ‘belong’ to the domain \( \mathcal{D} \). Moreover, the function \( \rho(\theta_{i;\mathcal{D}}) = \theta_{i;\mathcal{D}}^{-1} - 1, \theta_{i;\mathcal{D}} \in (0, 1] \) is a distance measure between the unit \( i \) and the domain \( \mathcal{D} \) (one can additionally set \( \rho(0) = \infty \)). We define a new estimator of \( t_{y;\mathcal{D}} \) by

\[
t_{y;\mathcal{D}}^{HR} = \sum_{i \in \mathcal{S}} \frac{\theta_{i;\mathcal{D}} y_i}{\sum_{j \in \mathcal{D}} \theta_{i;\mathcal{D}}} \quad \text{with} \quad \hat{\theta}_{i;\mathcal{D}} = \sum_{j \in \mathcal{D}} \hat{p}_{ij},
\]

which coincides with usual design unbiased H–T estimator if \( \mathcal{D} = \mathcal{U} \). Here \( \hat{p}_{ij} \) are estimates of \( \mathbb{3} \), because the parameters \( \alpha_1, \alpha_2 \) and \( \tau_1, \ldots, \tau_N \) of model \( \mathbb{2} \) are assumed not known as well as \( F_1(\cdot), \ldots, F_N(\cdot) \) are not specified. We note that an exact calculation of \( \mathbb{3} \) is complicated even the mentioned model characteristics are known, since, except trivial cases, \( X_1, \ldots, X_N \) are non-identically distributed random variables. Calculations of the distributions of order statistics of these random variables require very intensive computing in similar situations, see \( \mathbb{5} \) and see also \( \mathbb{4}, \mathbb{3} \), therefore here, first, we propose an alternative Monte–Carlo (M–C) approach. Second, to evaluate the estimator \( t_{y;\mathcal{D}}^{HR} \), we do not need to estimate probabilities \( \mathbb{3} \) separately, thus next, we give simple empirical approximations to the numbers \( \theta_{i;\mathcal{D}}, i = 1, \ldots, N \).

Monte–Carlo estimates of \( p_{ij} \). Depending on the specification of model \( \mathbb{2} \), first, we use the data \( (x_i, z_i), i = 1, \ldots, N \) to get estimates \( \hat{\alpha}_1, \hat{\alpha}_2 \) and \( \hat{\tau}_1, \ldots, \hat{\tau}_N \) of the model parameters, and also to decide on the class of \( F_i(\cdot) \) in order to obtain estimates \( \hat{F}_i(\cdot) \). Second, we apply the estimated model \( \hat{\eta} \):

\[
\hat{X}_i = \hat{\alpha}_1 + \hat{\alpha}_2 z_i + \hat{\tau}_i, \quad i = 1, \ldots, N,
\]

where \( \hat{X}_i \) has the distribution function \( \hat{F}_i(\cdot) \) with \( \mathbb{E}_{\hat{\eta}} \hat{X}_i = \hat{\alpha}_1 + \hat{\alpha}_2 z_i \) and \( \mathbb{V}_{\hat{\eta}} \hat{X}_i = \hat{\tau}_i^2 \), to generate the independent collections \( (x_{1r}, \ldots, x_{Nr}), r = 1, \ldots, R \), where \( R \) is a number of M–C iterations. Finally, for each \( r \), we order the numbers: \( x_{(1)r} \leq \cdots \leq x_{(N)r} \), and take

\[
\hat{p}_{ij} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I} \{ x_{ir} = x_{(j)r} \}, \quad i, j = 1, \ldots, N,
\]

5
where $\mathbb{I}\{\cdot\}$ is the indicator function. The relative frequencies (7) are consistent estimates of probabilities (3), as $N \to \infty$, if model (2) assumptions are sufficient to have the consistency of its parameter estimators, and because of the law of large numbers (as $R = R(N) \to \infty$). In practice, $R$ should be at least of the order $O(N^2)$.

**Approximations to $\theta_{i;D}$**. We propose approximations of the form: for $\hat{\alpha}_2 > 0$,

$$\hat{\theta}_{i;D} = cm \sum_{j \in \mathcal{D}} \mathbb{I}\{|z_j - z_i| \leq d_0 \hat{\alpha}_2^{-1}(\hat{\tau}_j^2 + \hat{\tau}_i^2)^{1/2}\}, \quad i = 1, \ldots, N,$$

(8)

where $c$ is the normalizing constant such that $\sum_{i \in \mathcal{U}} \hat{\theta}_{i;D} = m$, and $d_0$ is a number chosen from the interval $[1, 3]$, and $\hat{\alpha}_2$ and $\hat{\tau}_1, \ldots, \hat{\tau}_n$ are estimates of model (2) parameters $\alpha_2$ and $\tau_1, \ldots, \tau_N$ obtained from the data $(x_i, z_i)$, $i = 1, \ldots, N$. One can choose simply $d_0 = 2$.

**Remark 1.** Note that, for suitable application of estimator (6), the linear dependence between the variables $x$ and $z$ should be not too strong as well as not too weak. In particular, if $\text{Var}_y X_i = 0$ in (2), then $p_{ij} = \mathbb{I}\{i = j\}$, and then transform (4) becomes $t_j = y_j$. Thus we have nothing new if the relation between $x$ and $z$ is exact, i.e., from (4), we get the H–T estimator $\hat{\psi}_{HR} = \frac{1}{N} \sum_{i \in \mathcal{U}} \pi_i^{-1} y_i$, and return to the initial problem of the paper. The second marginal case in (2) is $\alpha_2 = 0$, $\text{Var}_y X_i = \tau^2 > 0$ and $F_i(\cdot) \equiv F(\cdot)$. Then $p_{ij} = N^{-1}$ for all $i, j = 1, \ldots, N$, and we get $t_j = N^{-1} \sum_{i \in \mathcal{U}} y_i$. Thus, if there is no linear dependence between $x$ and $z$, transform (1) is an uninformative, and, from (3), we obtain the estimator $\hat{\psi}_{HR} = m N^{-1} \sum_{i \in \mathcal{U}} \pi_i^{-1} y_i$ of (11) which is sensitive to an arbitrarily chosen domain $\mathcal{D}$ (homogeneous or not, with respect to the population).

**Bias and MSE of $\hat{\psi}_{HR}$ and their estimators.** Typically, in SAE model-based methods, the bias is a weighty component of MSE. In our case, assumption (5) also implies it. The bias of estimator (6) is

$$B_{yHR} = \text{BIAS}(\hat{\psi}_{HR}) = \sum_{i \in \mathcal{U}} \hat{\theta}_{i;D} y_i - \sum_{i \in \mathcal{D}} y_i.$$

(9)

Clearly, an unbiased estimator of (9) is not suitable because $B_{yHR}$ includes the parameter $t_{y;D}$. To get the estimator of $B_{yHR}$, a natural choice is to apply usual H–T estimation to the first sum in (9) and replace the second sum by (6). The result $\hat{B}_{yHR}$ seems paradoxical and also not reliable. Let us introduce the estimator

$$\hat{B}_{y2} = \sum_{i \in \mathcal{U}} \hat{\theta}_{i;D} \hat{y}_{i;\mathcal{D}} y_i \pi_i$$

(10)

and consider its linear combination with the estimator-zero: $\hat{B}_{yHR}(\psi) = \psi \hat{B}_{y2}$, $\psi \in \mathbb{R}$, with the aim to evaluate $\psi$ minimizing the MSE of $\hat{B}_{yHR}(\psi)$. In particular, the function $\text{MSE}(\hat{B}_{yHR}(\psi)) = \psi^2 \text{Var} \hat{B}_{y2} + (\psi \text{E} \hat{B}_{y2} - \hat{B}_{yHR})^2$ attains its unique minimum at the point

$$\psi_{0y} = \frac{\text{E} \hat{B}_{y2}}{\text{Var} \hat{B}_{y2} + (\text{E} \hat{B}_{y2})^2}$$.  

(11)

Here

$$\text{E} \hat{B}_{y2} = \sum_{i \in \mathcal{U}} \hat{\theta}_{i;D} \hat{y}_{i;\mathcal{D}}$$

and

$$\text{Var} \hat{B}_{y2} = \sum_{i, j \in \mathcal{U}} \hat{\theta}_{i;D} \hat{\theta}_{j;D} \hat{y}_{i;\mathcal{D}} \hat{y}_{j;\mathcal{D}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) y_i y_j$$

with $\pi_{ii} = \pi_i$. Thus, to estimate the bias $B_{yHR}$ of estimator (6), our suggestion is the estimator $\hat{B}_{yHR}(\hat{\psi}_{0y}) = \hat{\psi}_{0y} \hat{B}_{y2}$ with an estimate $\hat{\psi}_{0y}$ of $\psi_{0y}$. We note also that $\text{E} \hat{B}_{yHR}(\hat{\psi}_{0y}) = B_{yHR}(1 +$
Thus, in some situations, estimator (16) can reduce the MSE of the HR estimator.

Remark 2. Here \( \hat{\rho} \) is the unbiased estimator of \( \rho \). The size of the domain \( \mathcal{D} \) is given by (9).

Proposition 1. The MSE of the estimator \( \hat{t}_{HR} \) of the sum \( t_{y;D} \) in the domain \( \mathcal{D} \) is

\[
\text{MSE}(\hat{t}_{HR}) = \text{Var}(\hat{t}_{HR}) + ( \hat{B}_{y}^{HR} )^2,
\]

where \( \text{Var}(\hat{t}_{HR}) = \sum_{i,j \in \mathcal{D}} \hat{\theta}_{i;D} \hat{\theta}_{j;D}(\pi_i^{-1} \pi_j^{-1} - \pi_{ij}^{-1})y_i y_j \), and \( \hat{B}_{y}^{HR} \) is given by (6).

The estimator of (13) is

\[
\text{MSE}(\hat{t}_{HR}) = \text{Var}(\hat{t}_{HR}) + \hat{\psi}_{0y} \left( \hat{B}_{y}^{HR} \right)^2,
\]

where \( \text{Var}(\hat{t}_{HR}) = \sum_{i,j \in \mathcal{D}} \hat{\theta}_{i;D} \hat{\theta}_{j;D}(\pi_i^{-1} \pi_j^{-1} - \pi_{ij}^{-1})y_i y_j \) is the unbiased estimator of \( \text{Var}(\hat{t}_{HR}) \), the estimator \( \hat{B}_{y}^{HR} \) is given by (10) and

\[
\hat{\psi}_{0y} = \frac{\hat{B}_{y}^{HR} \hat{B}_{y}^{HR}}{\text{Var}(\hat{t}_{HR}) + ( \hat{B}_{y}^{HR} )^2}.
\]

Remark 2. Estimator (12) is also a candidate to estimate the bias \( B_{y}^{HR} \) in (13).

Remark 3. If the total sample size \( n \) and the size \( m \) of the domain \( \mathcal{D} \) are small then the sampling variance of the random variable \( \hat{t}_{HR} = \sum_{i \in \mathcal{S}} \hat{\theta}_{i;D} \pi_i^{-1} \) can be large because of a randomness of the set \( \{ \hat{\theta}_{i;D}, i \in \mathcal{S} \} \). Clearly, it means an additional error source for the HR estimator (6). Thus it is reasonable to apply the ratio estimator

\[
\hat{t}_{HR}^{1;D} = m \hat{t}_{HR}^{1;D},
\]

with the aim to reduce the variance of \( \hat{t}_{HR}^{1;D} \). Note that, for estimates such as (7) or (8), \( \hat{t}_{HR}^{1;D} \) is the unbiased estimator of \( m \). Therefore, roughly speaking, the ratio estimator \( \hat{t}_{HR}^{1;D} \) does not produce an additional bias compared to the HR estimator \( \hat{t}_{HR}^{1;D} \). Indeed, denote by \( \hat{B}_{y}^{HR1} = E \hat{t}_{HR1} - t_{y;D} \) the bias of the ratio estimator. Then, from \( \text{Cov}(\hat{t}_{HR1}^{1;D}, \hat{t}_{HR1}^{1;D}) = m(\text{E} \hat{t}_{HR1}^{1;D} - \hat{E} \hat{t}_{HR1}^{1;D}) \), we have

\[
\hat{B}_{y}^{HR1} = B_{y}^{HR} - m \text{Cov}(\hat{t}_{HR1}^{1;D}, \hat{t}_{HR1}^{1;D}).
\]

From here, it is easy to get the bound \( |\hat{B}_{y}^{HR1} - B_{y}^{HR}| \leq (\text{Var}(\hat{t}_{HR1}^{1;D}))^{1/2} \text{CV}(\hat{t}_{HR1}^{1;D}) \) where, generally, \( \text{CV}(\hat{t}_{HR1}^{1;D}) = m^{-1}(\text{Var}(\hat{t}_{HR1}^{1;D}))^{1/2} = O(n^{-1/2}) \) as \( n \to \infty \). Thus, in some situations, estimator (16) can reduce the MSE of the HR estimator.
3 Regression type estimator

Assume that the variable \( x \) explains the study variable \( y \) well, e.g., we have situations (i) and (ii) mentioned in Section 1. Note that, by Proposition 1, the variance of HR estimator (9) does not exceed the variance of H-T estimator of the whole population total (the case \( D = U \)). Therefore, if the variance of the unbiased estimator \( \hat{\theta}_{y:D} = \frac{\sum_{i\in s} \pi_i y_i}{\sum_{i\in s} \pi_i} \) is sufficient for quality requirements (and it is usually the case), then it is meaningful to use the auxiliary variable \( x \) to reduce the bias \( B_y^{HR} \) of (6). Thus we are motivated to consider a regression type estimator of the form \( \hat{\theta}_{y:D} = b(t_x:D - \hat{\theta}_{y:D}) \) with a properly chosen characteristic \( b \). In particular, minimizing the MSE of the proposed form, we get

\[
b_0 = \frac{C_{yx} + B_y^{HR} B_x^{HR}}{V_x + (B_x^{HR})^2} \tag{17}
\]

where \( C_{yx} = \text{Cov}(\hat{\theta}_{y:D}, \hat{\theta}_{y:D}) = \sum_{i,j\in U} \hat{\theta}_{i:D} \hat{\theta}_{j:D} (\pi_i^{-1} \pi_j^{-1} \pi_{ij}^{-1} - 1) y_i x_j \) and \( V_x = \text{Var} \hat{\theta}_{x:D} = C_{xx} \). Therefore, we introduce the following regression type estimator:

\[
\hat{\theta}_{y:D} = \hat{\theta}_{y:D} + b_0 (t_x:D - \hat{\theta}_{y:D}) \tag{18}
\]

where

\[
\hat{b_0} = \frac{\hat{C}_{yx} + \hat{V}_x \hat{\theta}_{x:D} \hat{B}_{y2} \hat{B}_{x2}^{HR}}{V_x + \hat{V}_x (\hat{B}_{x2}^{HR})^2} \tag{19}
\]

is the estimator of the parameter \( b_0 \) with \( \hat{\psi}_{0y} \) and \( \hat{\psi}_{0x} \) given by (15), \( \hat{B}_{y2}^{HR} \) and \( \hat{B}_{x2}^{HR} \) as in (10), and

\[
\hat{C}_{yx} = \sum_{i,j\in s} \hat{\theta}_{i:D} \hat{\theta}_{j:D} \left( \frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right) y_i x_j \quad \text{and} \quad \hat{V}_x = \sum_{i,j\in s} \hat{\theta}_{i:D} \hat{\theta}_{j:D} \left( \frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right) x_i x_j
\]

are unbiased estimators of \( C_{yx} \) and \( V_x \), respectively.

**Bias and MSE of \( \hat{\theta}_{y:D} \) and their estimators.** To evaluate the error of the regression type estimator, we start from the assumption that estimator (13) remains almost the same if \( \hat{\psi}_{0y} \) and \( \hat{\psi}_{0x} \) in \( \hat{b_0} \) are replaced back into their true values \( \psi_{0y} \) and \( \psi_{0x} \), i.e., we write

\[
\hat{\theta}_{y:D} \approx \hat{\theta}_{y:D} := \hat{\theta}_{y:D} + \hat{b_0} (t_x:D - \hat{\theta}_{y:D}) \quad \text{where} \quad \hat{b_0} = \frac{\hat{C}_{yx} + \hat{B}_y^{HR}(\psi_{0y})\hat{B}_x^{HR}(\psi_{0x})}{\hat{V}_x + (\hat{B}_x^{HR}(\psi_{0x}))^2}
\]

with the notation of Section 2. Next, the Taylor linearization of the estimators function \( \hat{\theta}_{y:D} \) at the point \( (E \hat{\theta}_{y:D}, E \hat{\theta}_{x:D}, \hat{C}_{yx}, \hat{V}_x, \hat{B}_y^{HR}, \hat{B}_x^{HR}) \) yields

\[
\hat{\theta}_{y:D} \approx \hat{\theta}_{y:D} := \hat{\theta}_{y:D} + b_0 (t_x:D - \hat{\theta}_{y:D}) - B_x^{HR} \left( V_x + (B_x^{HR})^2 \right)^{-1} \left\{ \hat{C}_{yx} - C_{yx} - b_0 \left( \hat{V}_x - V_x \right) 
\right.
\]

\[
+ B_x^{HR} \left( B_y^{HR}(\psi_{0y}) - B_y^{HR} \right) - \left( 2b_0 B_x^{HR} - B_y^{HR} \right) \left( B_x^{HR}(\psi_{0x}) - B_x^{HR} \right) \right\}
\]

Assume that \( \max\{\text{CV}(\hat{B}_y^{HR}), \text{CV}(\hat{B}_x^{HR})\} = o(1) \) as \( N, m, n \to \infty \). It is a correct condition by the same arguments as in Section 2. Then it follows, from the one-term Taylor expansion \( \hat{\theta}_{y:D} \), that for the approximation of the bias of estimator (13), holds \( |\text{BIAS}(\hat{\theta}_{y:D})| \leq |B_y^{HR} - b_0 B_y^{HR}|(1 + o(1)) \) as \( N, m, n \to \infty \). We conclude from this bound that the regression type HR estimator can reduce the
bias $B^R_y$ of estimator (6) if the variables $y$ and $x$ have a significant correlation. Next, cumbersome analysis shows that the variance of the term $i^H_{y;D} + b_0(t_{x;D} - i^H_{x;D})$ dominates in the variance of $i^R_{y;D}$. Thus we formulate the following proposition on the accuracy of estimator (18).

**Proposition 2.** The MSE approximation of the estimator $i^R_{y;D}$ of the sum $t_{y;D}$ in the domain $D$ is

$$\text{MSE}(i^R_{y;D}) \approx \text{Var}(i^H_{y;D} + \bar{b}_0V_x - 2\bar{b}_0C_{yx} + (B^R_y - \bar{b}_0B^H_x)^2),$$

where $\text{Var}(i^H_{y;D})$ is the same as in (13), $C_{yx}$ and $V_x$ are from expression (17) of $b_0$, and $B^R_y$ and $B^H_x$ are by formula (9).

The estimator of approximation (20) is

$$\text{MSE}(i^R_{y;D}) = \text{Var}(i^H_{y;D} + \bar{b}_0V_x - 2\bar{b}_0C_{yx} + (\hat{\psi}_{y\nu}B^H_{y\nu} - \hat{\psi}_{y\nu}\hat{B}^H_{x\nu})^2),$$

where $\text{Var}(i^H_{y;D})$ is the same as in (14), $\hat{C}_{yx}$ and $\hat{V}_x$ are from expression (19) of $b_0$, and $\hat{\psi}_{y\nu}$, $\hat{\psi}_{0\nu}$ and $\hat{B}^H_{y\nu}$, $\hat{B}^H_{x\nu}$ are given by (15) and (10), respectively.

4 Simulation study

The aim of the numerical modeling is to compare HR estimator (6) and regression type HR estimator (18) with direct design-based H–T and regression estimators, and synthetic and EBLUP estimators. Denote these estimators by $i^H_{y;D}$, $i^R_{y;D}$, $i^{HT}_{y;D}$, $i^{GREG}$, $i^{SYN}$ and $i^{EBLUP}$, respectively. In particular, we calculate their MSEs, variances and biases. To imitate a real situation, we generate the artificial population $U$ and the domain $D$ of interest as follows: $m$ values of the size variable $z$ is obtained from the particular distribution (superpopulation) and we say that it is the size values of the domain $D$. Next, with a different distribution, we simulate the remaining $N - m$ values of the variable $z$. Then we apply the specified regression models to get the population values of the variable $x$ from the obtained values of $z$, and, similarly, to simulate the values of the study variable $y$ from $x$. The probabilities $\theta_{i;D}$, $i = 1, \ldots, N$, are estimated by M–C method. The sampling design $\rho(\cdot)$ is chosen to be stratified with simple random samples without replacement inside the strata, where the total sample size $n$ is distributed by the values of the size variable $z$ using Neyman’s allocation. To evaluate the quality characteristics of the estimators, we apply M–C simulations by drawing independently the large number of stratified samples from the population. In Simulation 1 an expected sample size in the domain $D$ is small. In Simulation 2 the conditions of Simulation 1 are slightly modified in order to see a comparable behaviour of the estimators for a large sample in the domain.

**Simulation 1.** We get $m = 50$ values of the variable $z$ from the normal distribution $N(6.5, 0.64)$, and $N - m = 450$ values from the gamma distribution $\Gamma(2, 1)$ shifted to the right by 5 units. Here 2 and 1 is the shape and scale parameters, respectively. Regression model (2) is $X_i = z_i + \delta_i$ with the independent errors $\delta_i$ distributed by $\mathcal{N}(0, \tau^2)$ where $\tau$ is chosen so that the correlation coefficient $\rho_{xz}$ between $x$ and $z$ is close to 0.9. From this particular model we get and fix the population values of the variable $x$. The variable of interest $y$ is simulated by the model $Y_i = 2 + 5x_i + \varepsilon_i$ with the independent errors $\varepsilon_i$ distributed by $\mathcal{N}(0, \sigma^2x_i)$ where $\sigma$ is such that $\rho_{yx}$ is close to 0.7. The total sample size $n = 75$ is allocated to 4 strata. The simulation results are shown in Table 4.

The true value of the parameter of interest (11) is $t_{y;D} = 1694$ and the expected sample size in the domain $D$ is about 7. It is seen from Table 1 that the unbiased H–T estimator $i^{HT}_{y;D}$ has comparatively large variance. The direct regression estimator $i^{GREG}_{y;D}$ is almost (approximately)
Table 1: Simulation 1. Accuracies of the estimators.

|       | $\hat{t}_{HR_{y,D}}$ | $\hat{t}_{RHR_{y,D}}$ | $\hat{t}_{HT_{y,D}}$ | $\hat{t}_{GREG_{y,D}}$ | $\hat{t}_{SYN_{y,D}}$ | $\hat{t}_{EBLUP_{y,D}}$ |
|-------|----------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|
| MSE   | 7521                 | 4278                   | 316386                | 13269                  | 3922                  | 6221                   |
| Var   | 3473                 | 2058                   | 316386                | 13261                  | 2410                  | 5488                   |
| BIAS  | 64                   | 47                     | 0                     | 3                      | 39                    | 27                     |

unbiased and its variance is much smaller. Small-area estimators improve both of them. Their MSEs are similar but the synthetic estimator $\hat{t}_{SYN_{y,D}}$ has the smallest one. As we expected, the regression type HR estimator $\hat{t}_{RHR_{y,D}}$ reduces the bias of the HR estimator $\hat{t}_{HR_{y,D}}$.

Simulation 2. We take $m = 400$ and $N - m = 100$, and generate the population values of the variable $z$ from the same superpopulations as in Simulation 1. The variables $x$ and $y$ are simulated by the same regression models but with different constants $\tau$ and $\sigma$ to have similar correlations between the variables as in Simulation 1. The total sample size $n = 250$ is allocated to 4 strata. The simulation results are shown in Table 2.

Table 2: Simulation 2. Accuracies of the estimators.

|       | $\hat{t}_{HR_{y,D}}$ | $\hat{t}_{RHR_{y,D}}$ | $\hat{t}_{HT_{y,D}}$ | $\hat{t}_{GREG_{y,D}}$ | $\hat{t}_{SYN_{y,D}}$ | $\hat{t}_{EBLUP_{y,D}}$ |
|-------|----------------------|------------------------|-----------------------|------------------------|-----------------------|------------------------|
| MSE   | 14837                | 9330                   | 112079                | 10225                  | 12064                 | 9382                   |
| Var   | 14555                | 8444                   | 112079                | 10224                  | 9504                  | 9286                   |
| BIAS  | 17                   | 30                     | 0                     | -1                     | 51                    | 10                     |

The true value of the parameter of interest (1) is $t_{y,D} = 13840$ and the expected sample size in the domain $D$ is about 200. By Table 2 the direct H–T estimator $\hat{t}_{HT_{y,D}}$ is the worst one. The best practical choice would be the direct regression estimator $\hat{t}_{GREG_{y,D}}$ because of its approximate unbiasedness, but the estimators $\hat{t}_{RHR_{y,D}}$ and $\hat{t}_{EBLUP_{y,D}}$ are better in terms of MSE. Compared to Simulation 1 the increase of the domain sample considerably reduced the impact of the bias square to the MSE for all small-area estimators. For the HR estimator $\hat{t}_{HR_{y,D}}$ this decrease is the largest.

5 Conclusions

The hidden randomness estimators (6) and (18) have a robustness feature. That is, for any chosen estimation domain, an effective form of the estimators is the same. It is not necessarily a case for, e.g., synthetic and other estimators where procedures, to deal with an inhomogeneity between different domains in an optimal way, are not so clearly defined. HR estimator (6) can be also viewed as a robust version of its special case $\hat{t}_{HR_{y,D}} = m N^{-1} \sum_{i \in s} \pi_i^{-1} y_i$, see Remark 1 so, with (18), we have no care about a structure of the domain $D$ inside the population.

The HR estimators implicit dependence on underlying model (2) is quite week. In fact, model (2) assists in a use of the distribution of the size variable $z$, and only correctness of assumptions on the variance structure, which is generally defined by the parameters $\tau_1, \ldots, \tau_N$, and its estimation can affect a reliability of the estimators. Note that if we have knowledge about approximately equal variabilities of the population elements sizes, then it is not necessarily to look for an alternative size variable $x$. Then, simply, instead of (2), one can define the working model: $X_i = z_i + \delta_i$, $i = 1, \ldots, N$, with, say, independent errors $\delta_i$ distributed by $N(0, \tau^2)$ where $\tau > 0$ is chosen so that
the correlation coefficient \( \rho_{xz} \) is similar to the estimated from the sample coefficient \( \rho_{yz} \) between the variables \( y \) and \( z \).

The variance of HR estimator (6) in the domain \( D \) does not exceed the variance of the corresponding (H–T) estimator for the whole population. This feature remains to be valid, as the population and domain sizes grow. A study of an asymptotic behaviour of the design bias of estimator (6) is complicated because of insufficient knowledge about the probabilities (3). Therefore a consistency of the HR estimators is an open question. It is only clear that, by the construction of HR estimator (6), the bias tends to zero if the size of the domain approaches the size of the population; and then also the domain sample size increases accordingly. Moreover, one can expect that for a large estimation domain the bias becomes relatively small, as it is seen from the simulations in Section 4 and thus the HR estimators can be efficient for any domain of the population. It is known that it is not necessarily a case for other model-based estimators.

The estimation of the design bias of (6) also needs a further research because properties of the estimator \( \hat{B}_{y}^{HR} \), given by (12) and used in the proposed optimal form \( \psi \hat{B}_{y}^{HR} \), in general, are difficult verifiable. With a good estimator of the bias, say \( \hat{B}_{y}^{HR} \), one can expect to reduce the design bias of the domain sum HR estimator by taking the difference \( \hat{h}_{y}^{HR;D} - \hat{B}_{y}^{HR} \) instead of (6).

The HR estimator (6) has the additivity feature: if the population is partitioned into non-overlapping domains then the sum of the HR estimators over the domains coincides with H–T estimator of the whole population. Regression type HR estimator (18), however, is not additive.

A good feature of HR estimators (6) and (18) is that, in the case of no sample in the domain, the evaluation of the estimators quality characteristics is not a different, while for typical design-based estimators it is impossible.

The simulation study shows that, in the sense of the quality, the presented HR estimators are similar to the synthetic and EBLUP estimators.

References

[1] N. Balakrishnan. Permanents, order statistics, outliers, and robustness. *Revista Matemática Complutense*, 20:7–107, 2007.
[2] W.G. Cochran. *Sampling Techniques*. John Wiley & Sons, Inc., New York, 3rd edition, 1977.
[3] E. Cramer, K. Herle, and N. Balakrishnan. Permanent expansions and distributions of order statistics in the INID case. *Communications in Statistics – Theory and Methods*, 38:2078–2088, 2009.
[4] G.S. Datta. Model-based approach to small area estimation. In D. Pfeffermann and C.R. Rao, editors, *Sample Surveys: Inference and Analysis*, volume 29B of *Handbook of Statistics*, pages 251–288. North-Holland, Amsterdam, 2009.
[5] H.A. David and H.N. Nagaraja. *Order Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 3rd edition, 2003.
[6] J.C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.
[7] W.A. Fuller. *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.
[8] M. Ghosh and J.N.K. Rao. Small area estimation: an appraisal. *Statistical Science*, 9:55–76, 1994.
[9] M.A. Hidiroglou and Z. Patak. Domain estimation using linear regression. *Survey Methodology*, 30:67–78, 2004.
[10] R. Lehtonen and A. Veijanen. Design-based methods of estimation for domains and small areas. In D. Pfeffermann and C.R. Rao, editors, *Sample Surveys: Inference and Analysis*, volume 29B of *Handbook of Statistics*, pages 219–249. North-Holland, Amsterdam, 2009.
[11] D. Pfeffermann. Small area estimation–new developments and directions. *International Statistical Review, 70:*125–143, 2002.

[12] D. Pfeffermann. New important developments in small area estimation. *Statistical Science, 28:*40–68, 2013.

[13] J.N.K. Rao. *Small Area Estimation.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

[14] J.N.K. Rao. Inferential issues in small area estimation: some new developments. *Statistics in Transition, 7:*513–526, 2005.

[15] C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling.* Springer-Verlag, New York, 1992.

[16] C. Wu and R.R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association,* 96:185–193, 2001.