Design-based composite estimation rediscovered

Andrius Čiginas[1,2]

[1]Vilnius University
[2]Statistics Lithuania

Abstract

Small area estimation methods are used in surveys, where sample sizes are too small to get reliable direct estimates of parameters in some population domains. We consider design-based linear combinations of direct and synthetic estimators and propose a two-step procedure to approach the optimal combination. We construct the mean square error estimator suitable for this and for any other linear composition that estimates the optimal one. We will present a simulation study at the congress, where we use data from the Lithuanian Labor Force Survey to estimate proportions of the unemployed and employed in municipalities.

Keywords: small area estimation, composite estimator, mean square error, bias.

1 Introduction

Traditional direct estimation methods are not effective if there are additional needs to estimate parameters for unplanned domains of the survey population. This is because the direct estimators are based on the domain sample only and therefore the sample sizes obtained may be too small to provide accurate results in some of that domains. To solve such a problem, indirect estimators are used in the small area estimation theory [6], where the estimation domain (area) is called small, if the direct estimator has there an unacceptably large variance. The indirect estimators are based on linking models that help to borrow sample information from neighbor domains through auxiliary data available from registers or other surveys. That approach increases the effective sample size and hence reduces the variances of estimators in the small domain. The disadvantage of these estimators is their biases, while the direct estimators are unbiased or approximately so.

Synthetic estimators based on implicit linking models and their linear combinations with the direct estimators constitute an important subclass of the indirect estimators. They are considered in the traditional design-based estimation theory [6, Chapter 3], where estimators of parameters are based only on the randomness induced by the sampling design. The composite estimation is a good way to find a trade-off between large variances of the direct estimators and biases of the synthetic estimators. Even some modern indirect estimators, like the empirical best linear unbiased predictors (EBLUPs) [4, 1], built using linear mixed models, are expressed as the linear combinations of the direct and synthetic estimators. Nowadays, it is almost accepted that explicit small area models, like that including random area-specific effects for EBLUPs, are a more flexible tool in complex situations of estimation than the traditional estimators. On the other hand, the latter design-based approach is desirable in many sample surveys, and estimators are quite simple.

Despite the simplicity of the traditional design-based compositions, they are less attractive due to the difficulty in estimating their design mean square errors (MSEs) and especially the bias parts of MSEs. We derive the general MSE estimator for any composition that approximates the optimal one in the sense of the minimal MSE. We also propose a new two-step procedure to estimate the optimal linear combination for any pair of direct and synthetic estimators.

2 Design-based composite estimation

2.1 Preliminary concepts

The set $\mathcal{U} = \{1, \dots, N\}$ consists of the labels of elements of the finite survey population. The partition $\mathcal{U} = \mathcal{U}_1 \cup \dots \cup \mathcal{U}_M$ of the population describes the domains of interest, where $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ as $i \neq j$, and there are N_i elements in the domain $\mathcal{U}_i$. Let y be a study variable with the fixed values $y_1, \dots, y_N$ assigned to the elements of $\mathcal{U}$. To estimate the domain parameters θ_i, for instance, the domain means $\theta_i = \sum_{k \in \mathcal{U}_i} y_k / N_i$, $i = 1, \dots, M$, the sample $s \subset \mathcal{U}$ of size $n < N$ is drawn according to the sampling design $p(\cdot)$. If the design without replacement was not constructed to ensure the samples $s_i = s \cap \mathcal{U}_i$ of fixed sizes n_i in the domains, then small n_i can be obtained, and then the accuracy of any direct estimators $\hat{\theta}_i^{\mathrm{d}}$ of θ_i is questionable because of large design variances $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$. Hereafter we use the symbols E_{p}, $\mathrm{var}_{\mathrm{p}}$, and $\mathrm{MSE}_{\mathrm{p}}$ to denote expectation, variance, and MSE calculated according to $p(\cdot)$, respectively.

An alternative to the direct estimator $\hat{\theta}_i^{\mathrm{d}}$ is the synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$, which uses the sample of a larger area through the implicit linking model. A typical model stands on the synthetic assumption that the small domain has the same characteristics as the large area [6, Chapter 3]. Similarly, direct estimators $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i have large design variances themselves for small sample sizes. Therefore, applying the generalized variance function approach [7], the estimators $\hat{\psi}_i^{\mathrm{d}}$ are smoothed, and more stable estimators $\hat{\psi}_i^{\mathrm{s}}$ are further used.

2.2 Approximations to optimal compositions

Since the synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$ of θ_i uses larger sample, its design variance is smaller compared to that of the direct estimator $\hat{\theta}_i^{\mathrm{d}}$. However, a contribution of its design bias to MSE can be substantial if the synthetic assumption is not realistic. To find a balance between larger variances ψ_i of $\hat{\theta}_i^{\mathrm{d}}$ and the biases of $\hat{\theta}_i^{\mathrm{S}}$, we consider the linear compositions

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

with coefficients $0 \leqslant \lambda_i \leqslant 1$. Minimizing the function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\theta}_i^{\mathrm{C}}(\lambda_i))$ with respect to λ_i, the optimal weight for the ith domain is the population characteristic

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

Applying the assumption $|C_i| \ll \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$ and knowing that the estimator $\hat{\theta}_i^{\mathrm{d}}$ is nearly unbiased, a standard approximation used to optimal parameter (2) is [6, Section 3.3]

$$\lambda_i^* \approx \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})/(\psi_i + \mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})). \tag{3}$$

However, further evaluation from sample data is still complicated because of difficulties to estimate $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$. The best general method known in the literature, which does not require any additional synthetic assumptions, is to use the representation [6, Section 3.2.5]

$$\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{S}}) = \mathrm{E_p}(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}})^2 - \mathrm{var_p}(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}) + \mathrm{var_p}(\hat{\theta}_i^{\mathrm{S}}), \tag{4}$$

which includes the unbiased direct estimator $\hat{\theta}_i^{\mathrm{d}}$, and then to build an approximately design unbiased estimator

$$\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{S}}) = (\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}})^2 - \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}) \tag{5}$$

of (4), where $\hat{\sigma}^2(\cdot)$ stands for an estimator of the design variance $\mathrm{var_p}(\cdot)$. However, estimator (5) can be very unstable for individual small domains, and thus it is not efficient to use it for estimation of weight (2) or its approximation (3).

Therefore, less straightforward ways are used to approximate and estimate the optimal coefficients for compositions (1). One of the ideas is to set a common weight for all domains (or groups of them) and then minimize a total MSE with respect to that weight [5]. A similar but more sophisticated composite estimation is to apply James–Stein method [6, Section 3.4]. A flexible proposal is sample-size-dependent estimation [3], where estimators of the weights λ_i in (1) are taken to be dependent on the sample sizes in the domains.

2.3 Estimation of mean square errors

Estimation of MSEs of the design-based composite estimators is a difficult task, as pointed several times in [6, Chapter 3]. That is due to estimation of the component $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{S}})$, and estimated weights $\hat{\lambda}_i$ add more complexity. The main problem here is to estimate biases of the estimators, while we can always apply at least resampling methods to evaluate the design variances.

The general method used for the synthetic estimators can be applied to the compositions as well, see [6, Example 3.3.1] and [2]. That is, treating the composition $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ as a synthetic estimator, one can use the estimator

$$\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{C}}) = (\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}})^2 - \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}} - \hat{\theta}_i^{\mathrm{d}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}}) \tag{6}$$

of $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}})$. However, this estimator has the same drawbacks as (5) including undesirable property to take negative values.

We construct the estimator of MSE for any composite estimator $\hat{\theta}_i^{\mathrm{C}}$ defined by (1) that is close to the optimal combination $\hat{\theta}_i^{\mathrm{opt}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i^*)$. First consider general composition (1) with a fixed weight. Assuming that its direct component $\hat{\theta}_i^{\mathrm{d}}$ is nearly unbiased, we have

$$\widetilde{D}_i = \mathrm{E_p}(\tilde{\theta}_i^{\mathrm{C}}) - \theta_i \approx (1-\lambda_i)B_i, \quad \text{where} \quad B_i = \mathrm{E_p}(\hat{\theta}_i^{\mathrm{S}}) - \theta_i \tag{7}$$

denotes the design bias of the synthetic part. Assuming additionally that $\max\{|C_i|, \mathrm{var_p}(\hat{\theta}_i^{\mathrm{S}})\} \ll \psi_i$, optimal parameter (2) is approximated by the quantity $\tilde{\lambda}_i^* = B_i^2/(\psi_i + B_i^2)$. Assume next that the number λ_i in (1) is chosen so that it is close to the optimal λ_i^*. Then, solving the approximate equation $\tilde{\lambda}_i^* \approx \lambda_i$, we obtain $B_i^2 \approx \lambda_i\psi_i/(1-\lambda_i)$. Inserting the latter relation into the square of (7), we arrive to

$$\widetilde{D}_i^2 \approx \lambda_i(1-\lambda_i)\psi_i. \tag{8}$$

For any design-based composite estimator $\hat{\theta}_i^{\mathrm{C}} \approx \hat{\theta}_i^{\mathrm{opt}}$, we derive the squared estimator $\widehat{D}_i^2$ of the bias $D_i = \mathrm{E_p}(\hat{\theta}_i^{\mathrm{C}}) - \theta_i$ by letting $D_i^2 \approx \widetilde{D}_i^2$ and then replacing the unknown parameters in (8) by their empirical analogs. Finally, we get the estimators

$$\mathrm{mse_b}(\hat{\theta}_i^{\mathrm{C}}) = \hat{\lambda}_i(1-\hat{\lambda}_i)\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}}), \qquad i = 1, \dots, M, \tag{9}$$

of $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}})$, where the term $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$ is an estimator of the design variance $\mathrm{var_p}(\hat{\theta}_i^{\mathrm{C}})$.

2.4 Procedure of composite estimation

We propose a straightforward procedure to estimate optimal weight (2) through approximation (3) and employing MSE estimation by (9). In the first step, let us assume that the squared bias B_i^2 is negligible. Then $\hat{\lambda}_i^{(1)} = \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}))$ is a good estimator of (2), and $\hat{m}_i^{(1)} = \mathrm{mse}_{\mathrm{b}}(\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)}))$ is the MSE estimator of the obtained composition. However, the assumption $B_i \approx 0$ may be incorrect, and then the empirical weight $\hat{\lambda}_i^{(1)}$ is very likely to underestimate λ_i^*. Therefore, in the second step, we treat the composition $\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)})$ as the synthetic estimator and build the new composition

$$\hat{\theta}_i^{\mathrm{Cb}} = \hat{\lambda}_i^{(2)}\hat{\theta}_i^{\mathrm{d}} + (1 - \hat{\lambda}_i^{(2)})\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)}), \quad \text{where} \quad \hat{\lambda}_i^{(2)} = \hat{m}_i^{(1)}/(\hat{\psi}_i^{\mathrm{s}} + \hat{m}_i^{(1)}), \tag{10}$$

and $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{Cb}}) = \hat{\lambda}_i^{(2)}(1 - \hat{\lambda}_i^{(2)})\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{Cb}})$ is the estimator of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{Cb}})$ according to (9).

3 Concluding remarks

The proposed MSE estimation supports the idea that if the optimal linear composition cannot be evaluated well, one should not expect to get an accurate MSE estimate. The estimator of MSE is very general and takes only non-negative values.

The direct and traditional synthetic estimators should always be combined in small domains. There are many ways to estimate the optimal combination, but we construct another design-based composite estimator that adapts to the MSE estimation construction principle.

References

[1] G.E. Battese, R.M. Harter, W.A. Fuller, An error-components model for prediction of county crop areas using survey and satellite data, *J. Amer. Statist. Assoc.*, **83**(401):28–36, 1988.

[2] A. Čiginas, Adaptive composite estimation in small domains, *Nonlinear Anal. Model. Control*, **25**(3):341–357, 2020.

[3] 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.

[4] R.E. Fay, R.A. Herriot, Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.*, **74**(366):269–277, 1979.

[5] N.J. Purcell, L. Kish, Estimation for small domains, *Biometrics*, **35**:365–384, 1979.

[6] J.N.K. Rao, I. Molina, *Small Area Estimation*, John Wiley, New Jersey, 2 edition, 2015.

[7] K.M. Wolter, *Introduction to Variance Estimation*, Springer-Verlag, New York, 2 edition, 2007.