Direct, synthetic and composite estimation of parameters in population domains

Andrius Čiginas

Statistics Lithuania. State Data Agency

2023-02-14

On design-based estimation methods

Suppose there are estimation domains with too small sample sizes to get sufficiently accurate direct estimates.

- That is, design-based direct estimation is not efficient for that domains.
- Traditional design-based synthetic and composite estimators can be efficient and they are desirable due to their simplicity, but there are difficulties with inferences.

Direct, synthetic, and composite estimation is a necessary step before trying model-based estimation (Tzavidis et al., 2018). Moreover, sometimes it can be the final step if we are satisfied with the accuracy.

Preliminaries

- $\mathcal{U} = \{1, \ldots, N\}$ is a finite survey population.
- There are M domains (areas) $\mathcal{U}_1, \ldots, \mathcal{U}_M$ of sizes $N_1, \ldots, N_M$ such that $\mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M = \mathcal{U}$ and $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ as $i \neq j$.
- y is the study variable with the fixed values $y_1, \ldots, y_N$ in $\mathcal{U}$. This is a continuous, binary, or categorical variable.
- We estimate the domain parameters θ_i, $i = 1, \ldots, M$. That is domain sums, means, proportions, or more complex parameters.
- The sample $s \subset \mathcal{U}$ of size $n < N$ is drawn according to the sampling design $p(\cdot)$. Hereafter we use the symbols $\mathrm{P_p}$, $\mathrm{E_p}$, $\mathrm{var_p}$, and $\mathrm{MSE_p}$ to denote probability, expectation, variance, and MSE calculated according to $p(\cdot)$, respectively.

Direct estimation in domains

Direct estimates are the first thing needed to be calculated.

- Let $\hat{\theta}_i^{\mathrm{d}}$ be any approximately design unbiased direct estimator of θ_i based on the sample $s_i = s \cap \mathcal{U}_i$ of size n_i.
- If n_i is small, large design variance $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$ is obtained.
- Direct estimators $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i have high variances themselves for small samples s_i, too.

Let us estimate the domain means (or domain proportions)

$$\theta_i = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} y_k, \qquad i = 1, \ldots, M,$$

where the numbers N_i are assumed to be known.

Direct estimation. Example (I)
Weighted sample means

Let $\pi_k = \mathrm{P_p}\{k \in s\} > 0$. The weighted sample means

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

are approximately unbiased estimators of the domain means θ_i.

The direct estimators (Särndal et al., 1992)

$$\hat{\psi}_i^{\mathrm{d}} = \hat{\psi}_i^{\mathrm{H}} = \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} \sum_{l \in s_i} (1 - \pi_k \pi_l / \pi_{kl}) \frac{(y_k - \hat{\theta}_i^{\mathrm{d}})(y_l - \hat{\theta}_i^{\mathrm{d}})}{\pi_k \pi_l}$$

of the variances ψ_i, where $\pi_{kl} = \mathrm{P_p}\{k, l \in s\} > 0$, can have high variances as well.

Note: applied when there is no (useful) auxiliary information or it is available at the area level only.

Direct estimation. Example (II)
Generalized regression (GREG) estimators

Let $\mathbf{x}_k = (1, x_{2k}, \ldots, x_{Pk})'$ be the vector containing the values of auxiliary variables $x_2, \ldots, x_P$ for $k \in \mathcal{U}$. Assume that the data $\mathbf{x}_k$ are available for $k \in s$, and the vector $\boldsymbol{\theta}_{xi} = \sum_{k \in \mathcal{U}_i} \mathbf{x}_k / N_i$ of means is known for the ith area. Denote

$$\hat{\theta}_i^{\mathrm{HT}} = \frac{1}{N_i} \sum_{k \in s_i} \frac{y_k}{\pi_k} \quad \text{and} \quad \hat{\boldsymbol{\theta}}_{xi}^{\mathrm{HT}} = \frac{1}{N_i} \sum_{k \in s_i} \frac{\mathbf{x}_k}{\pi_k}.$$

The generalized regression estimators (Rao and Molina, 2015)

$$\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{GR}} = \hat{\theta}_i^{\mathrm{HT}} + (\boldsymbol{\theta}_{xi} - \hat{\boldsymbol{\theta}}_{xi}^{\mathrm{HT}})' \widehat{\mathbf{B}}_i, \qquad i = 1, \ldots, M,$$

of θ_i are approximately (asymptotically) design unbiased, where

$$\widehat{\mathbf{B}}_i = (\widehat{B}_{i1}, \ldots, \widehat{B}_{iP})' = \left(\sum_{k \in s_i} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k \pi_k} \right)^{-1} \sum_{k \in s_i} \frac{\mathbf{x}_k y_k}{c_k \pi_k}$$

with positive constants c_k.

Explanation: the quantity $\widehat{\mathbf{B}}_i$ estimates the characteristic

$$\mathbf{B}_i = (B_{i1}, \ldots, B_{iP})' = \left(\sum_{k \in \mathcal{U}_i} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_{k \in \mathcal{U}_i} \frac{\mathbf{x}_k y_k}{c_k}$$

of the ith domain, which is, in turn, the GLS estimator of the fixed effects $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iP})'$ of the assisting model $\mathrm{E_m}(y_k) = \mathbf{x}_k' \boldsymbol{\beta}_i$ defined for $k \in \mathcal{U}_i$, where $\mathrm{E_m}$ stands for the model expectation.

If it is supposed that the variances $\mathrm{Var_m}(y_k) = \sigma_k^2$ are unequal, it is natural to set $c_k = \sigma_k^2$, $k \in \mathcal{U}_i$, and estimate them. In some standard situations like $c_k = \sigma^2$, $k \in \mathcal{U}_i$, the GREG estimator reduces to

$$\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{GR}} = \boldsymbol{\theta}_{xi}' \widehat{\mathbf{B}}_i.$$

Note: GREG estimator $\hat{\theta}_i^{\mathrm{GR}}$ reduces to the weighted sample mean $\hat{\theta}_i^{\mathrm{H}}$ if c_k is a constant and there is no auxiliary information in the sense that $\mathbf{x}_k = 1$ for all $k \in \mathcal{U}_i$.

The direct estimators (Rao and Molina, 2015)

$$\hat{\psi}_i^{\mathrm{d}} = \hat{\psi}_i^{\mathrm{GR}} = \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} \sum_{l \in s_i} (1 - \pi_k \pi_l / \pi_{kl}) \frac{(y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_i)(y_l - \mathbf{x}_l' \widehat{\mathbf{B}}_i)}{\pi_k \pi_l}$$

of the variances ψ_i have high variances themselves for small n_i.

If the design $p(\cdot)$ is complex and the overall sampling fraction n/N is small ($\mathcal{U}$ is large), the assumption $\pi_{kl} \approx \pi_k \pi_l$, $k \neq l$, is often used, and then the approximation

$$\hat{\psi}_i^{\mathrm{GR}} \approx \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} \frac{1}{\pi_k} \left(\frac{1}{\pi_k} - 1 \right) (y_k - \mathbf{x}_k' \widehat{\mathbf{B}}_i)^2$$

is applied. The same is applied in the case of the weighted sample means.

Direct estimation. Example (III)
Modified GREG estimators

Let us replace the coefficients $\widehat{\mathbf{B}}_i$ used in the GREG estimators by the overall regression coefficient

$$\widehat{\mathbf{B}} = (\widehat{B}_1, \ldots, \widehat{B}_P)' = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k \pi_k} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k y_k}{c_k \pi_k},$$

which estimates the population characteristic

$$\mathbf{B} = (B_1, \ldots, B_P)' = \left(\sum_{k \in \mathcal{U}} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_{k \in \mathcal{U}} \frac{\mathbf{x}_k y_k}{c_k}.$$

Here the parameter $\mathbf{B}$ is taken from the assisting regression model

$$\mathrm{E_m}(y_k) = \mathbf{x}_k' \boldsymbol{\beta}, \qquad k \in \mathcal{U},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$ are fixed effects defined for the whole $\mathcal{U}$.

Then the so-called modified GREG estimators of the domain means θ_i are (Rao and Molina, 2015)

$$\hat{\theta}_i^{\mathrm{d}} = \hat{\theta}_i^{\mathrm{GRm}} = \hat{\theta}_i^{\mathrm{HT}} + (\boldsymbol{\theta}_{xi} - \hat{\boldsymbol{\theta}}_{xi}^{\mathrm{HT}})'\widehat{\mathbf{B}}, \qquad i = 1, \ldots, M.$$

The estimator $\hat{\theta}_i^{\mathrm{GRm}}$ for the ith domain is still approximately (asymptotically) design unbiased despite that the coefficient $\widehat{\mathbf{B}}$ uses the sample outside the domain.

The direct estimators (Rao and Molina, 2015)

$$\hat{\psi}_i^{\mathrm{d}} = \hat{\psi}_i^{\mathrm{GRm}} = \frac{1}{\widehat{N}_i^2} \sum_{k \in s_i} \sum_{l \in s_i} (1 - \pi_k \pi_l / \pi_{kl}) \frac{(y_k - \mathbf{x}_k' \widehat{\mathbf{B}})(y_l - \mathbf{x}_l' \widehat{\mathbf{B}})}{\pi_k \pi_l}$$

of the variances ψ_i might have high variances for small n_i, too.

Synthetic estimation and smoothing

- The synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$ of θ_i uses the sample of a larger area through an implicit linking model. The model stands on the ***synthetic assumption*** that the small domain $\mathcal{U}_i$ has the same characteristics as the large area (Rao and Molina, 2015). That $\hat{\theta}_i^{\mathrm{S}}$ has a smaller design variance than $\hat{\theta}_i^{\mathrm{d}}$ but is biased.
- Similarly, the estimators $\hat{\psi}_i^{\mathrm{d}}$ of $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$ are smoothed applying the generalized variance function approach (Wolter, 2007). That smoothed (synthetic) $\hat{\psi}_i^{\mathrm{s}}$ provides more realistic information on the accuracy of the direct estimator $\hat{\theta}_i^{\mathrm{d}}$ than $\hat{\psi}_i^{\mathrm{d}}$ with small n_i and is used in further estimation procedures.

Note on mean squared errors

In the design mean squared error (MSE)

$$\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i) = \mathrm{E}_{\mathrm{p}}(\hat{\theta}_i - \theta_i)^2 = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i) + [\mathrm{E}_{\mathrm{p}}(\hat{\theta}_i) - \theta_i]^2$$

measuring the accuracy of the estimator $\hat{\theta}_i$ of the parameter θ_i, the squared bias part

$$\mathrm{B}^2_{\mathrm{p}}(\hat{\theta}_i) = [\mathrm{E}_{\mathrm{p}}(\hat{\theta}_i) - \theta_i]^2$$

is typically assumed negligible for the direct estimators $\hat{\theta}_i = \hat{\theta}^{\mathrm{d}}_i$ but $\mathrm{B}^2_{\mathrm{p}}(\hat{\theta}_i)$ cannot be ignored for the synthetic estimators $\hat{\theta}_i = \hat{\theta}^{\mathrm{S}}_i$.

Synthetic estimation. Example (I)
Weighted sample mean over population

If there is no auxiliary information, the design unbiased weighted sample mean

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}^{\mathrm{H}} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{y_k}{\pi_k}, \quad \text{where} \quad \widehat{N} = \sum_{k \in s} \frac{1}{\pi_k}, \qquad i = 1, \ldots, M,$$

estimating the whole population mean $\theta = \sum_{k \in \mathcal{U}} y_k / N$ can be used as estimators of the domain means θ_i.

Explanation: the true means θ_i and θ are characteristics of the definition above, and the synthetic assumption on their closeness is used to derive the synthetic estimates $\hat{\theta}^{\mathrm{H}}$.

As $n_i \ll n$, one can expect that $\mathrm{var_p}(\hat{\theta}^{\mathrm{H}}) \ll \mathrm{var_p}(\hat{\theta}_i^{\mathrm{H}})$. However, the bias part $\mathrm{B_p^2}(\hat{\theta}^{\mathrm{H}})$ can dominate $\mathrm{MSE_p}(\hat{\theta}^{\mathrm{H}})$ for domains where the means θ_i differ significantly from the population mean θ.

Let us use a similar idea to smooth the direct estimators $\hat{\psi}_i^{\mathrm{H}}$ of the variances $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{H}})$. We apply the pooled variance estimator

$$\hat{\psi}_i^{\mathrm{s}} = \hat{\psi}_i^{\mathrm{sP}} = \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{s}^2}{n_i} \quad \text{with} \quad \hat{s}^2 = \frac{1}{\widehat{N} - 1} \sum_{k \in s} \frac{(y_k - \hat{\theta}^{\mathrm{H}})^2}{\pi_k}$$

of ψ_i for each $i = 1, \ldots, M$.

Explanation: the smoothing (synthetic estimators $\hat{\psi}_i^{\mathrm{sP}}$) is based on the synthetic assumption that the characteristics

$$s_i^2 = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} (y_k - \theta_i)^2 \quad \text{and} \quad s^2 = \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \theta)^2$$

are close according to the definition of synthetic estimation.

Synthetic estimation. Example (II)
Regression-synthetic estimators

Consider the weighted sample means $\hat{\theta}_i^{\mathrm{H}}$ as the direct estimators of θ_i. Let $\mathbf{z}_i = (1, z_{2i}, \ldots, z_{Pi})'$ be auxiliary data available for the ith domain, and $\hat{\psi}_i^{\mathrm{s}}$ are smoothed estimators of $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{H}})$, for example, $\hat{\psi}_i^{\mathrm{s}} = \hat{\psi}_i^{\mathrm{sP}}$. The regression-synthetic estimator

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{RS}} = \mathbf{z}_i'\hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{M} \frac{\mathbf{z}_i\mathbf{z}_i'}{\hat{\psi}_i^{\mathrm{s}}}\right)^{-1} \sum_{i=1}^{M} \frac{\mathbf{z}_i\hat{\theta}_i^{\mathrm{H}}}{\hat{\psi}_i^{\mathrm{s}}}$$

of θ_i is obtained from the area-level model

$$\hat{\theta}_i^{\mathrm{H}} = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, M,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)'$ are fixed effects and the sampling errors ε_i are assumed independent with $\mathrm{E}_{\mathrm{p}}(\varepsilon_i) = 0$ and $\mathrm{var}_{\mathrm{p}}(\varepsilon_i) = \psi_i$.

Explanation: the estimators $\hat{\theta}_i^{\mathrm{RS}}$ rely on the synthetic assumption that the parameter $\boldsymbol{\beta}$ is the same across all domains.

To smooth the direct estimators $\hat{\psi}_i^{\mathrm{H}}$ of the variances $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{H}})$ for the domain *proportions* θ_i, another simple method can be used. Assume that $\psi_i \approx K N_i^{\gamma}$ and estimate the parameters $K > 0$ and $\gamma \in \mathbb{R}$ applying the regression model (Dick, 1995)

$$\log(\hat{\psi}_i^{\mathrm{H}}) = \log(K) + \gamma \log(N_i) + \eta_i, \qquad i = 1, \ldots, M,$$

where errors η_i are independent and identically distributed. Then the smoothed variances are

$$\hat{\psi}_i^{\mathrm{sD}} = \widehat{K} N_i^{\hat{\gamma}}, \qquad i = 1, \ldots, M,$$

which can be next multiplied by a bias correction as suggested in Hidiroglou et al. (2019).

Synthetic estimation. Example (III)
GREG-synthetic estimators

Assume that unit-level auxiliary information is at our disposal so that we can build the direct GREG estimators. Recall their simplified (projection) form $\hat{\theta}_i^{\mathrm{GR}} = \boldsymbol{\theta}'_{xi}\widehat{\mathbf{B}}_i$.

Replacing the quantities $\widehat{\mathbf{B}}_i$ by the overall regression coefficient $\widehat{\mathbf{B}}$ as we did to derive the modified GREG estimators, we get the synthetic estimators

$$\hat{\theta}_i^{\mathrm{S}} = \hat{\theta}_i^{\mathrm{GRS}} = \boldsymbol{\theta}'_{xi}\widehat{\mathbf{B}}, \qquad i = 1, \ldots, M,$$

called GREG-synthetic estimators or indirect GREG estimators.

Explanation: the GREG-synthetic estimator $\hat{\theta}_i^{\mathrm{GRS}}$ is supported by the synthetic assumption that the domain-specific parameter $\widehat{\mathbf{B}}_i$ is close to the global characteristic $\widehat{\mathbf{B}}$.

To smooth the direct estimators $\hat{\psi}_i^{\mathrm{GR}}$ of $\psi_i = \mathrm{var_p}(\hat{\theta}_i^{\mathrm{GR}})$, one can apply ANOVA-type pooling (Boonstra et al., 2008)

$$\hat{\psi}_i^{\mathrm{s}} = \hat{\psi}_i^{\mathrm{sAP}} = \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{s}_p^2}{n_i} \quad \text{with} \quad \hat{s}_p^2 = \frac{1}{n - M} \sum_{i=1}^{M} (n_i - 1) \hat{\psi}_i^{\mathrm{GR}}$$

similar to the one considered above.

Estimation of MSE of the synthetic estimators

For any synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$, the main problem is its bias $\mathrm{B_p}(\hat{\theta}_i^{\mathrm{S}})$ estimation, while the variance part $\mathrm{var_p}(\hat{\theta}_i^{\mathrm{S}})$ of the MSE can be approximated analytically or evaluated at least by applying resampling methods such as a bootstrap or jackknife.

One choice is to apply an approximately design unbiased estimator (Gonzalez and Waksberg, 1973)

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

of $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{S}})$, where $\hat{\sigma}^2(\cdot)$ stands for an estimator of $\mathrm{var_p}(\cdot)$. A good approximation to this estimator is

$$\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{S}}) \approx (\hat{\theta}_i^{\mathrm{S}} - \hat{\theta}_i^{\mathrm{d}})^2 - \hat{\psi}_i^{\mathrm{d}}.$$

However, both the estimators have large design variances and can even take negative values (Rao and Molina, 2015).

If the aim is to compare different synthetic estimation methods, one can average the MSE estimator over domains to get a stable accuracy characteristic (Gonzalez and Waksberg, 1973). That is, we can take

$$\overline{\mathrm{mse}}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{S}}) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_m^{\mathrm{S}}).$$

However, this estimator is not area-specific.

There are no straightforward ways to smooth unstable estimators $\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{S}})$ of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{S}})$ for individual domains, therefore additional assumptions should be introduced to derive more stable estimators of MSE. For example, the synthetic assumption (Marker, 1995)

$$\mathrm{B}_{\mathrm{p}}^2(\hat{\theta}_i^{\mathrm{S}}) \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{B}_{\mathrm{p}}^2(\hat{\theta}_m^{\mathrm{S}})$$

allows to derive

$$\mathrm{mse}_{\mathrm{M}}(\hat{\theta}_i^{\mathrm{S}}) = \overline{\mathrm{mse}}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{S}}) + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}) - \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^2(\hat{\theta}_m^{\mathrm{S}}).$$

Summary for the synthetic estimators

Advantages:

- Their variances are small compared to that of the direct estimators;
- Estimation is possible for domains where there is no sample.

Disadvantages:

- They can be seriously biased due to strong synthetic assumptions;
- Their biases persist as the sample size increases;
- Benchmarking adjustments are needed;
- There are no stable MSE estimators.

Design-based composite estimation

The design-based linear composition

$$\tilde{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i) = \lambda_i \hat{\theta}_i^{\mathrm{d}} + (1 - \lambda_i)\hat{\theta}_i^{\mathrm{S}}$$

with weight $0 \leqslant \lambda_i \leqslant 1$ is a trade-off between the larger variance of the direct estimator $\hat{\theta}_i^{\mathrm{d}}$ and the bias of the synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$. Or, in other words, it is a balance between unbiasedness of $\hat{\theta}_i^{\mathrm{d}}$ and smaller variance of $\hat{\theta}_i^{\mathrm{S}}$.

Question: how to properly choose the weights λ_i, $i = 1, \ldots, M$?

Minimizing the function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\theta}_i^{\mathrm{C}}(\lambda_i))$ with respect to λ_i, the optimal weight λ_i^* is obtained and then approximated using

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

see Rao and Molina (2015).

Approximations to optimal compositions

1. A straightforward estimation of the optimal weight λ_i^* leads to estimators like

$$\hat{\lambda}_i = \frac{\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{S}})}{\hat{\psi}_i^{\mathrm{s}} + \mathrm{mse_u}(\hat{\theta}_i^{\mathrm{S}})}$$

with the considered estimator $\mathrm{mse_u}(\hat{\theta}_i^{\mathrm{S}})$ of $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{S}})$ from Gonzalez and Waksberg (1973). This is too unstable to be used in practice.

2. Purcell and Kish (1979) propose to assume a common weight $\lambda_i = \lambda$ in the composition $\tilde{\theta}_i^{\mathrm{C}}(\lambda_i)$ and minimize the function $\sum_{m=1}^{M} \mathrm{MSE_p}(\tilde{\theta}_m^{\mathrm{C}}(\lambda))$ with respect to λ. It leads to more stable estimators

$$\hat{\lambda} = \sum_{m=1}^{M} \mathrm{mse_u}(\hat{\theta}_m^{\mathrm{S}}) \Bigg/ \left\{ \sum_{m=1}^{M} [\hat{\psi}_m^{\mathrm{s}} + \mathrm{mse_u}(\hat{\theta}_m^{\mathrm{S}})] \right\},$$

but that pooling over domains may not be reasonable for some of them.

3. In a sample-size-dependent estimation in Drew et al. (1982), $\hat{\lambda}_i$ is set to be dependent on the sample size in the domain. The estimators of the weights λ_i are taken to be of the form

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

These weights are dependent on the single subjectively chosen parameter δ for all domains with default value $\delta = 1$.

However, a choice of δ may vary from survey to survey. To choose the value of δ for the composition $\tilde{\theta}_i^{\mathrm{C}}(\delta) = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i(\delta))$, one can minimize numerically the sample based function

$$r(\delta) = \overline{\mathrm{mse}}_{\mathrm{u}}(\tilde{\theta}_i^{\mathrm{C}}(\delta)) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{mse}_{\mathrm{u}}(\tilde{\theta}_m^{\mathrm{C}}(\delta))$$

with respect to δ, where $\tilde{\theta}_i^{\mathrm{C}}(\delta)$ is treated as the synthetic estimator (Čiginas, 2020).

Estimation of MSE of the composite estimators

Treating any composition $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ as the synthetic estimator, we can apply the same formulas to estimate $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$ introduced for the synthetic estimators:

- approximately unbiased but unstable $\mathrm{mse}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$;
- stable but not area-specific $\overline{\mathrm{mse}}_{\mathrm{u}}(\hat{\theta}_i^{\mathrm{C}})$;
- area-specific but biased $\mathrm{mse}_{\mathrm{M}}(\hat{\theta}_i^{\mathrm{C}})$.

Alternative method. Assuming that $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ approximates the optimal combination $\hat{\theta}_i^{\mathrm{opt}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i^*)$ well, one can apply the estimator (Čiginas, 2021)

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

of $\mathrm{MSE}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$, where the term $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$ is an estimator of $\mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$. The estimator takes only non-negative values but can be biased.

One more (self-adapting) composite estimator

The composition is built in two steps (Čiginas, 2021).

1. To estimate the optimal coefficient λ_i^*, take the estimator

$$\hat{\lambda}_i^{(1)} = \frac{\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})}{\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})},$$

and $\hat{m}_i^{(1)} = \mathrm{mse}_{\mathrm{b}}(\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)}))$ is the MSE estimator for the composition $\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)})$.

2. Since it is expected that $\hat{\lambda}_i^{(1)} < \lambda_i^*$, treat $\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)})$ as the synthetic estimator and build the new composition

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

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 for the MSE.

Direct and synthetic estimators to combine

- If there is no auxiliary information, combine the weighted sample means $\hat{\theta}_i^{\mathrm{H}}$ with the weighted sample mean over population $\hat{\theta}^{\mathrm{H}}$.
- If auxiliary information is available at the area level only, combine the weighted sample means $\hat{\theta}_i^{\mathrm{H}}$ with the regression-synthetic estimators $\hat{\theta}_i^{\mathrm{RS}}$.
- If auxiliary information is available at the unit level, combine GREG estimators $\hat{\theta}_i^{\mathrm{GR}}$ or modified GREG estimators $\hat{\theta}_i^{\mathrm{GRm}}$ with GREG-synthetic estimators $\hat{\theta}_i^{\mathrm{GRS}}$.

Benchmarking

The direct estimator $\hat{\theta}^{\mathrm{d}}$ of the whole population mean θ is usually reliable, but the indirect estimators $\hat{\theta}_i = \hat{\theta}_i^{\mathrm{S}}$ or $\hat{\theta}_i^{\mathrm{C}}$ of the domain means θ_i do not necessarily satisfy

$$\frac{1}{N}\sum_{m=1}^{M} N_m \hat{\theta}_m = \hat{\theta}^{\mathrm{d}}.$$

Therefore, the synthetic or composite estimators are benchmarked. A simple ratio adjustment is

$$\hat{\theta}_i^{*} = \hat{\theta}_i \hat{\theta}^{\mathrm{d}} \Big/ \frac{1}{N}\sum_{m=1}^{M} N_m \hat{\theta}_m,$$

for $i = 1, \ldots, M$.

Some final notes

- The design-based direct, synthetic, and composite estimators for the domains of interest are the initial triplet of estimators that needed to be calculated before proceeding to model-based estimation.
- The choice of the particular triplet or triplets should be based on the availability of auxiliary information. From these relatively simple estimators, we get information on the predictive power of the auxiliary variables and heterogeneity across the estimation domains, which is useful for further modeling.
- If we are satisfied with the accuracy of the indirect estimators, the tested composite estimator can be one that is used to solve the small area estimation problem of the survey.

References

Boonstra, H.J., van den Brakel, J.A., Buelens, B., Krieg, S., Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *Metron* 66:21–49.

Čiginas, A. (2020). Adaptive composite estimation in small domains. *Nonlinear Analysis: Modelling and Control* 25:341–357.

Čiginas, A. (2021). Design-based composite estimation rediscovered. `arXiv:2108.05052 [stat.ME]`.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology* 21:45–54.

Drew, J.D., Singh, M.P., Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology* 8:17–47.

Gonzalez, M.E., Waksberg, J. (1973). Estimation of the error of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Hidiroglou, M.A., Beaumont, J.-F., Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology* 45:101–126.

Marker, D.A. (1995). *Small Area Estimation: A Bayesian Perspective.* PhD thesis, University of Michigan, Ann Arbor. Unpublished

Purcell, N.J., Kish, L. (1979). Estimation for small domains. *Biometrics* 35:365–384.

Rao, J.N.K., Molina, I. (2015). *Small Area Estimation.* 2nd edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Tzavidis, N., Zhang, L.-Ch., Luna, A., Schmid, T., Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society, Series A* 181:927–979.

Wolter, K.M. (2007). *Introduction to Variance Estimation.* 2nd edition, Springer-Verlag, New York.