Adaptive composite estimation in small population areas

Andrius Čiginas

Vilnius University

Small Area Estimation Conference
7–11 July 2025
Torino Italy

Design-based vs model-based estimation

Suppose there are estimation domains (areas) with sample sizes too small to yield sufficiently accurate direct estimates.

- That is, design-based direct estimation is inefficient for these domains.
- Traditional design-based (synthetic and) composite estimators improve efficiency and are desirable, though they present challenges in inference.
- By contrast, model-based estimators, such as EBLUPs, offer greater flexibility in estimation and have well-developed MSE estimation methods.

Aim: To re-evaluate classical design-based composite estimators and methods for estimating their MSEs.

Direct estimation in domains

- $\mathcal{U} = \{1, \ldots, N\}$ represents a finite survey population.
- There are M areas $\mathcal{U}_1, \ldots, \mathcal{U}_M$ with sizes $N_1, \ldots, N_M$, such that $\mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M = \mathcal{U}$ and $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$ for $i \neq j$.
- y is the study variable with fixed values $y_1, \ldots, y_N$ in $\mathcal{U}$.
- We aim to estimate the domain parameters θ_i, $i = 1, \ldots, M$.
- The sample $s \subset \mathcal{U}$, of size $n < N$, is drawn according to a sampling design $p(\cdot)$.
- Let $\hat{\theta}_i^{\mathrm{d}}$ be an 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, we get a large design variance $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$.
- Direct estimators $\hat{\psi}_i^{\mathrm{d}}$ of ψ_i also have high variances for small samples s_i.

Direct estimation in domains: an example

- 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 values N_i are assumed to be known.

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

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

are approximately unbiased direct estimators of θ_i.

- The direct estimators (Särndal et al., 1992)

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

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

Synthetic estimation

- The synthetic estimator $\hat{\theta}_i^{\mathrm{S}}$ of θ_i uses the sample of a larger area through an implicit linking model. This model relies on the ***synthetic assumption*** that the small domain $\mathcal{U}_i$ has the same characteristics as the large area (Rao and Molina, 2015). Consequently, $\hat{\theta}_i^{\mathrm{S}}$ has a smaller design variance than $\hat{\theta}_i^{\mathrm{d}}$, but it is biased.
- Similarly, the estimators $\hat{\psi}_i^{\mathrm{d}}$ of $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{d}})$ are smoothed using the generalized variance function approach (Wolter, 2007). This yields a smoothed (synthetic) estimator $\hat{\psi}_i^{\mathrm{s}}$.

Synthetic estimation: an example

- Let $\mathbf{z}_i = (z_{1i}, z_{2i}, \ldots, z_{Pi})'$ be auxiliary data for $\mathcal{U}_i$, and let $\hat{\psi}_i$ be any estimator of ψ_i. The regression-synthetic estimator

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

  of θ_i is derived from the basic area-level model for EBLUP, ignoring random area effects (Rao and Molina, 2015).

- It may be preferable to take $\hat{\psi}_i = \hat{\psi}_i^{\mathrm{s}}$ instead of the direct estimators $\hat{\psi}_i^{\mathrm{d}}$, where, for domain *proportions* θ_i, the quantities $\hat{\psi}_i^{\mathrm{s}}$ smooth $\hat{\psi}_i^{\mathrm{d}}$ using the assumption $\psi_i \approx K N_i^{\gamma}$. Here, $K > 0$ and $\gamma \in \mathbb{R}$ are estimated through a log-log regression model, similar to the approach in Dick (1995).

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

Question: How should we choose the weights λ_i, $i = 1, \ldots, M$?

By minimizing the function $\mathrm{MSE}_{\mathrm{p}}(\tilde{\theta}_i^{\mathrm{C}}(\lambda_i))$ with respect to λ_i, we obtain the optimal weight λ_i^*, which can be approximated by

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

as shown in Rao and Molina (2015).

Approximations to optimal compositions

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

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

with 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 $\hat{\sigma}^2(\cdot)$ denotes an estimator of $\mathrm{var_p}(\cdot)$.

2. Use of a common weight across all domains by minimizing the total MSE (Purcell and Kish, 1979).
3. Application of the James–Stein method, as described in Rao and Molina (2015).
4. Sample-size-dependent estimation, where $\hat{\lambda}_i$ adjusts to the domain sample size (Drew et al., 1982).

Estimation of mean square errors

General method. Treating the composite estimator $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ as synthetic, one can use

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

to estimate $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}})$. This estimator is approximately unbiased but has a large variance and can yield negative values (Rao and Molina, 2015).

Alternative method. Assuming $\hat{\theta}_i^{\mathrm{C}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i)$ approximates the optimal combination $\hat{\theta}_i^{\mathrm{opt}} = \tilde{\theta}_i^{\mathrm{C}}(\lambda_i^*)$ well, we can use

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

to estimate $\mathrm{MSE_p}(\hat{\theta}_i^{\mathrm{C}})$, where $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{C}})$ estimates $\mathrm{var_p}(\hat{\theta}_i^{\mathrm{C}})$. This MSE estimator is non-negative but may be biased (Čiginas, 2023).

Sample-size-dependent estimation

According to Drew et al. (1982), the estimators of the weights λ_i are taken in 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 depend on a single, subjectively chosen parameter δ for all domains, with a default value of $\delta = 1$.

However, the choice of δ may vary by survey. If very small domains dominate, an appropriate δ is often significantly greater than 1 (Čiginas, 2020).

Adaptive sample-size-dependent estimator

To select the value of δ for the composition $\tilde{\theta}_i^{\mathrm{C}}(\delta) = \tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i(\delta))$, we numerically minimize the sample-based function

$$r(\delta) = \frac{1}{M} \sum_{i=1}^{M} \mathrm{mse_u}(\tilde{\theta}_i^{\mathrm{C}}(\delta))$$

with respect to δ. The adaptive composite estimators of the domain parameters θ_i are then defined by (Čiginas, 2020)

$$\hat{\theta}_i^{\mathrm{SSD}} = \tilde{\theta}_i^{\mathrm{C}}(\hat{\delta}^*) \quad \text{where} \quad \hat{\delta}^* = \arg\min_{\delta > 0} r(\delta).$$

To evaluate the MSEs of these compositions, we apply the estimators

$$\mathrm{mse_b}(\hat{\theta}_i^{\mathrm{SSD}}) = \hat{\lambda}_i(\hat{\delta}^*)(1 - \hat{\lambda}_i(\hat{\delta}^*))\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{SSD}}).$$

Self-adapting (two-step) composite estimator

The composition is built in two steps (Čiginas, 2023).

1. To estimate the optimal coefficient λ_i^*, use 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 $\hat{\lambda}_i^{(1)} < \lambda_i^*$ is expected, treat $\tilde{\theta}_i^{\mathrm{C}}(\hat{\lambda}_i^{(1)})$ as the synthetic estimator and construct the new composition

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

   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 the MSE estimator.

EBLUP based on the Fay–Herriot model

The EBLUP of the domain parameter θ_i is expressed as the linear combination (Fay and Herriot, 1979)

$$\hat{\theta}_i^{\mathrm{FH}} = \hat{\gamma}_i \hat{\theta}_i^{\mathrm{d}} + (1 - \hat{\gamma}_i)\mathbf{z}_i' \hat{\boldsymbol{\beta}} \quad \text{where} \quad \hat{\gamma}_i = \frac{\hat{\sigma}_v^2}{\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2},$$

and

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

where $\hat{\sigma}_v^2$ is an estimator of the variance σ_v^2 of random domain effects. Here, we use the estimator $\hat{\sigma}_v^2$ of σ_v^2 based on the method of moments, as originally proposed by Fay and Herriot (1979).

Mean square error estimation for EBLUP

An approximately unbiased estimator of the MSE of $\hat{\theta}_i^{\mathrm{FH}}$ was derived in Datta et al. (2005):

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

For *comparison*, if we ignore the covariance term, we obtain

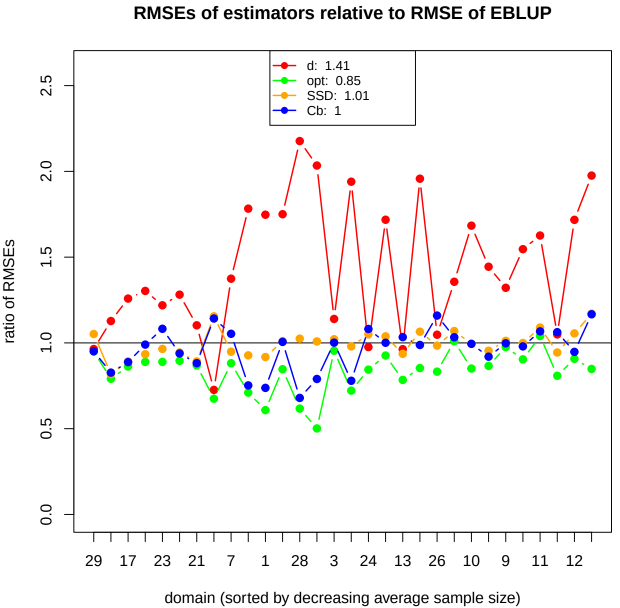
$$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{C}}) \approx \hat{\lambda}_i \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\lambda}_i)^2 \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})$$

for the design-based composite estimators.

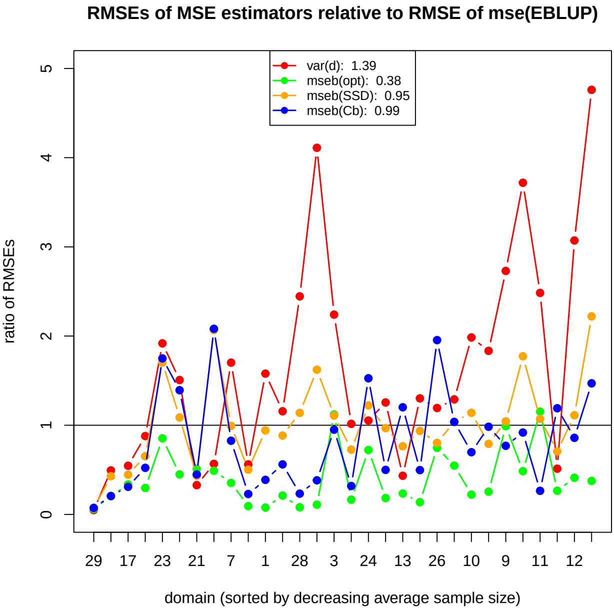
Simulation using the Labor Force Survey data

- The parameters θ_i are the *proportions* of unemployed and employed individuals in Lithuanian municipalities.
- The artificial population $\mathcal{U}$ has $N = 1\,396\,763$ individuals and $M = 30$ domains.
- The proportions in $\mathbf{z}_i = (1, z_{2i}, z_{3i}, z_{4i}, z_{5i}, z_{6i})'$ are defined as follows: z_{2i} is registered unemployment, z_{3i} represents individuals paying social contributions, z_{4i} denotes males, and z_{5i} and z_{6i} correspond to age groups 26–40 and 41–55.
- The estimators $\hat{\theta}_i^{\mathrm{d}}$ and $\hat{\theta}_i^{\mathrm{S}}$ are used as in the examples.
- We draw $R = 1\,000$ samples, each with $n \approx 7\,667$ individuals, by selecting $n' = 3\,700$ households with unequal probabilities.
- We calculate the ratios between RMSEs for $\hat{\theta}_i^{\mathrm{d}}$, $\hat{\theta}_i^{\mathrm{opt}}$, $\hat{\theta}_i^{\mathrm{SSD}}$, $\hat{\theta}_i^{\mathrm{Cb}}$, and the RMSE of EBLUP $\hat{\theta}_i^{\mathrm{FH}}$.
- We evaluate the ratios between RMSEs for $\hat{\psi}_i^{\mathrm{d}}$, $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{opt}})$, $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{SSD}})$, $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{Cb}})$, and the RMSE of $\mathrm{mse}(\hat{\theta}_i^{\mathrm{FH}})$.

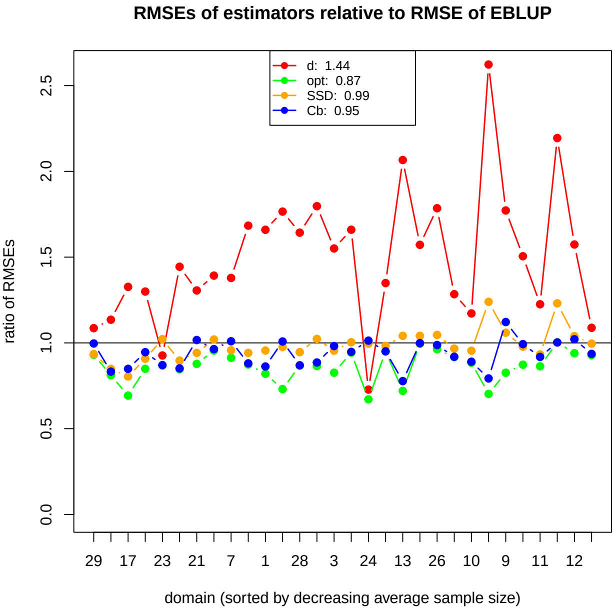
Estimation of unemployed



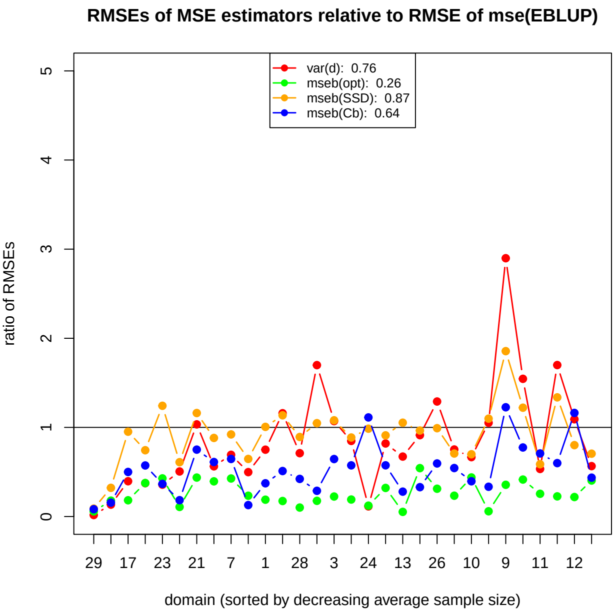
Estimation of MSEs for unemployed



Estimation of employed



Estimation of MSEs for employed



Conclusions

- The proposed adaptive design-based composite estimators provide viable alternatives to the model-based EBLUP, as demonstrated in the presented simulation study and observed in other experiments.
- These compositions are applicable to direct and synthetic estimators based on unit-level auxiliary data and can be used to estimate various domain parameters.
- The proposed design-based MSE estimator is well-suited for any design-based composition aimed at estimating the optimal one. This estimator is simple and ensures non-negative values.

References

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

Čiginas, A. (2023). Design-based composite estimation rediscovered. *Stat* 12:1–8.

Datta, G.S., Rao, J.N.K., Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* 92:183–196.

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.

Fay, R.E., Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74:269–277.

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.

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.

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