Small area estimation in the survey of Lithuanian census

Andrius Čiginas
(joint work with Ieva Burakauskaitė)

Statistics Lithuania

IAOS 2022 Conference

Main objects of the survey

- $\mathcal{U} = \{1, \ldots, N\}$ is a finite census population of individuals.
- There are M domains $\mathcal{U}_1, \ldots, \mathcal{U}_M$ of known 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 $i \neq j$. For example, the domains are municipalities, $M = 60$.
- Categorical variables of the survey:
  1. *religion* (16 categories);
  2. *mother tongue* (more than 12 categories);
  3. *knowledge of other languages* (16 languages);
  4. *ethnicity* (mass imputation is used).

  It is sufficient to consider binary variables. Let y be one of these with the fixed values $y_1, \ldots, y_N$ in $\mathcal{U}$.
- We aim to estimate the domain proportions

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

  or totals $N_i \theta_i$.

Sample design and primary sampling weights

- The sample $s \subset \mathcal{U}$ of size $n < N$ was drawn according to the sampling design $p(\cdot)$ with inclusion into the sample probabilities $\pi_k = \mathrm{P_p}\{k \in s\} > 0$, $k \in \mathcal{U}$.
- We got the sample $s = s^{(1)} \cup s^{(2)} \cup s^{(3)}$, where
  1. the part $s^{(1)}$ contains individuals from the voluntary sample;
  2. $s^{(2)}$ consists of other units which cannot be included into the sampling frame (the part for imputation);
  3. the part $s^{(3)}$ is the probability sample drawn from the sampling frame $\mathcal{U}^{(3)} = \mathcal{U} \backslash \{s^{(1)} \cup s^{(2)}\}$.
- The primary sampling weights are $d_k = 1/\pi_k$, where $\pi_k = 1$ as $k \in s^{(1)} \cup s^{(2)}$, and, in the hth stratum of $\mathcal{U}^{(3)}$,

  $$\pi_k \approx m_k n'_h / N'_h, \qquad k \in s^{(3)},$$

  where N'_h is the stratum size, n'_h is the number of adresses selected, and m_k is the number of persons in the adress.

Regression (calibrated) estimators in domains

- The domain samples $s_i = s \cap \mathcal{U}_i$ are of sizes $n_i \leqslant N_i$.
- Let $\mathbf{x}_k = (1, x_{2k}, \ldots, x_{Pk})'$ be a P-dimensional vector containing the values of auxiliary variables $x_2, \ldots, x_P$ for $k \in \mathcal{U}$, and $\boldsymbol{\theta}_{xi} = \sum_{k \in \mathcal{U}_i} \mathbf{x}_k / N_i$ is the vector of means for each domain $i = 1, \ldots, M$.

The generalized regression estimators (Rao and Molina, 2015)

$$\hat{\theta}_i^{\mathrm{GR}} = \boldsymbol{\theta}'_{xi} \widehat{\mathbf{B}}_i \quad \text{with} \quad \widehat{\mathbf{B}}_i = \left(\sum_{k \in s_i} \frac{\mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \sum_{k \in s_i} \frac{\mathbf{x}_k y_k}{\pi_k}$$

of θ_i , $i = 1, \ldots, M$, are approximatelly design unbiased if n_i are not small.

The set of variables $x_2, \ldots, x_P$ includes binary variables on age groups, gender, and religions (2011 data) intersected with counties.

The problem

- The direct estimator $\hat{\theta}_i^{\mathrm{GR}}$ of the proportion θ_i is based only on the sample of the ith domain. The domain sample sizes n_i are small for some domains and there the design variances $\psi_i = \mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{GR}})$ are large.
- The direct estimators (Rao and Molina, 2015)

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

  of ψ_i, where $\pi_{kl} = \mathrm{P}_{\mathrm{p}}\{k, l \in s\} > 0$, have high variances themselves for small samples s_i.
- The true proportions are often very small in the estimation domains. For example, the five-number summary for 16 religions in 60 municipalities (2011 complete data) is

$$(0.000000, 0.000095, 0.000681, 0.007380, 0.922597).$$

Preliminaries for small area estimation

- $(\hat{\theta}_i^{\mathrm{GR}}, \hat{\psi}_i^{\mathrm{GR}})$ are the direct estimators for $i = 1, \ldots, M$.
- $\mathbf{z}_i = (1, z_i)'$ is auxiliary information for the ith domain, where z_i is the proportion of the corresponding variable from the previous complete census 2011.

Using the approximation $\boxed{\psi_i \approx D_i \theta_i (1 - \theta_i)/n_i}$ by Kish (1995) and assuming that the design effects $D_i = c$ for all $i = 1, \ldots, M$,

$$\hat{\psi}_i^{\mathrm{s}} = \hat{c} z_i (1 - z_i)/n_i, \quad \text{where} \quad \hat{c} = \frac{N^2 \hat{\psi}^{\mathrm{s}}}{\sum_{i=1}^{M} \widetilde{N}_i^2 z_i (1 - z_i)/n_i},$$

are smoothed versions of the variances $\hat{\psi}_i^{\mathrm{GR}}$, $i = 1, \ldots, M$.

Here $\hat{\psi}^{\mathrm{s}}$ smooths the direct estimator $\hat{\psi}^{\mathrm{GR}}$ of the variance of the calibrated estimator for the whole population proportion, and $\widetilde{N}_i$ is the size of the ith domain in census 2011.

Synthetic estimation

1. In the case of a *not very small* proportion θ_i, we apply the regression-synthetic estimator

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

   which is obtained from the basic domain-level model for EBLUP ignoring random area effects (Rao and Molina, 2015). Here we take

$$\hat{\psi}_i = \hat{\psi}_i^{\mathrm{c}} = \max\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{GR}}\}$$

   according to Čiginas (2022).

2. For a *very small* proportion θ_i, we apply the synthetic estimator

$$\hat{\theta}_i^{\mathrm{S}} = z_i,$$

   which is a constant.

Design-based composite estimation

To estimate the domain proportions θ_i, we apply the composite (shrinkage) estimators (Čiginas, 2022)

$$\hat{\theta}_i^{\mathrm{C}} = \hat{\lambda}_i \hat{\theta}_i^{\mathrm{GR}} + (1 - \hat{\lambda}_i)\hat{\theta}_i^{\mathrm{S}} \quad \text{with} \quad \hat{\lambda}_i = \frac{\min\{\hat{\psi}_i^{\mathrm{s}}, \hat{\psi}_i^{\mathrm{GR}}\}}{\hat{\psi}_i^{\mathrm{c}}}.$$

That estimation is based on the monotonicity of the function $\psi_i \approx \psi(\theta_i) := D_i \theta_i (1 - \theta_i)/n_i$. That is if the direct estimator $\hat{\theta}_i^{\mathrm{GR}}$ is an outlier by its small or large value, then relatively more weight is attached to the synthetic part $\hat{\theta}_i^{\mathrm{S}}$.

Assuming that $\hat{\theta}_i^{\mathrm{C}}$ approximates an optimal linear combination of $\hat{\theta}_i^{\mathrm{GR}}$ and $\hat{\theta}_i^{\mathrm{S}}$ quite well, we 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 $\mathrm{var}_{\mathrm{p}}(\hat{\theta}_i^{\mathrm{C}})$. We use Rao et al. (1992) bootstrap to estimate the latter variance.

Benchmarking

Let $\hat{\theta}^{\mathrm{C}}_{ij}$, $i = 1, \ldots, M$, be the estimates for $j = 1, \ldots, J$ categories ($J = 16$ for religions). Let $\hat{\theta}^{\mathrm{c}}_{j}$ be the final estimate of the whole population proportion θ_j for the jth category.

We require that

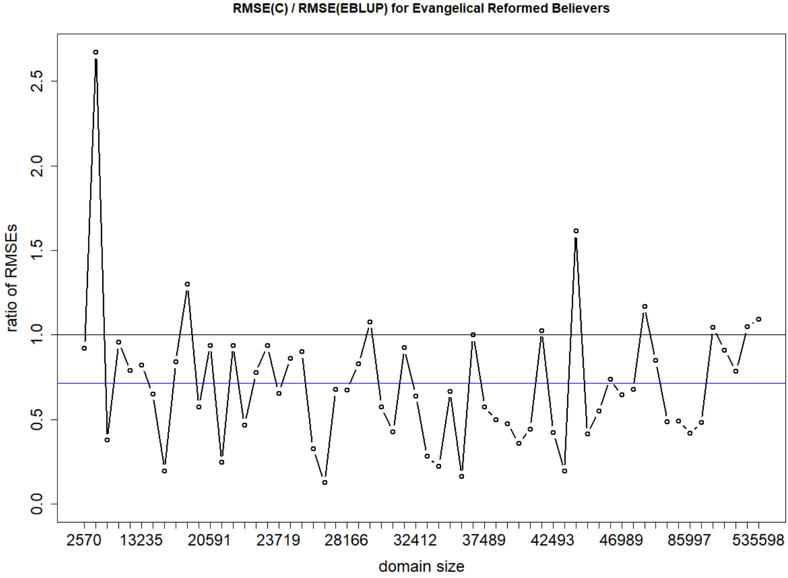
$$\frac{1}{N}\sum_{i=1}^{M} N_i \hat{\theta}^{\mathrm{C}}_{ij} = \hat{\theta}^{\mathrm{c}}_{j} \quad \text{for} \quad j = 1, \ldots, J$$

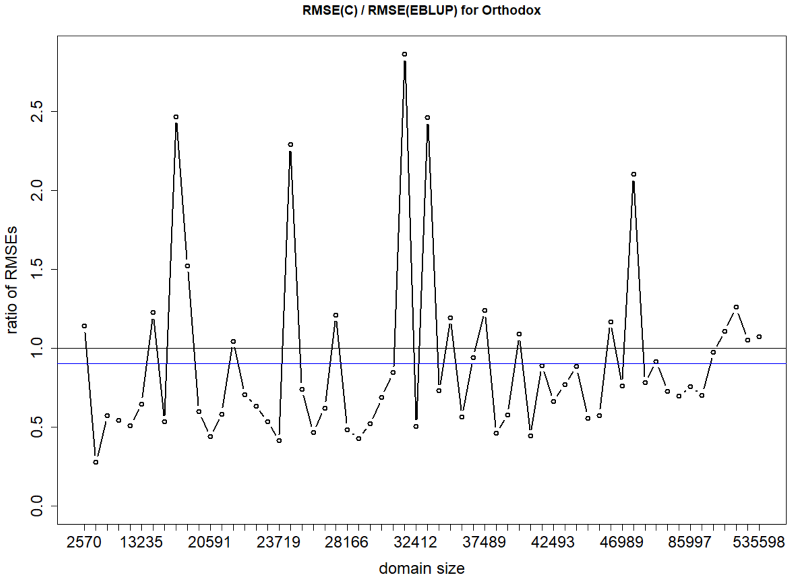
and

$$\sum_{j=1}^{J} \hat{\theta}^{\mathrm{C}}_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, M.$$

The estimates $\hat{\theta}^{\mathrm{C}}_{ij}$, $i = 1, \ldots, M$, $j = 1, \ldots, J$, are benchmarked to satisfy the above conditions using the criterion of weighted least squares with the inverse MSE estimates $\mathrm{mse}_{\mathrm{b}}(\hat{\theta}^{\mathrm{C}}_{i})$ as the weights (Boonstra et al., 2008).

Simulation with 2011 data. Composition vs EBLUP




RMSE(C) / RMSE(EBLUP) for Orthodox
ratio of RMSEs
0.0
0.5
1.0
1.5
2.0
2.5
2570
13235
20591
23719
28166
32412
37489
42493
46989
85997
535598
domain size

Summary

- Due to the very small true proportions, the small area estimation may be relevant for any division of the population.
- Applied design-based composite shrinkage estimation supported by domain-level models is robust compared to unit-level alternatives.
- Domain-level information available from the previous full census is crucial for the efficiency of the estimators. That means more challenges in the next census.

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

Čiginas, A. (2022). Design-based composite estimation of small proportions in small domains. `arXiv:2202.13085 [stat.ME]`.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics* 11:55–77.

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

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* 18:209–217.