State data agency Statistics Lithuania

Small area estimation using incomplete auxiliary information

Donatas Šlevinskas[1,2], Andrius Čiginas[1,2], Ieva Burakauskaitė[1,2]

[1]State Data Agency (Statistics Lithuania)

[2]Vilnius University

Motivating examples and the problem

To better estimate important population parameters in surveys with probability samples, we may want to exploit

- closely related
- but incomplete auxiliary data (non-probability samples).

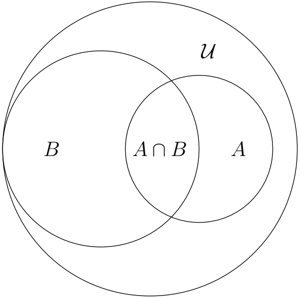
Examples:

1. *Administrative* VAT turnover data for monthly Short-term statistics surveys on the turnover of business enterprises.
2. Investment in tangible assets data from a *cut-off sample* of the Structural Business Statistics for the annual investment survey.
3. *Scraped* online job advertisement data for job vacancies measured in the quarterly statistical survey on earnings.

Problem: to properly integrate samples of non-probability origin for estimation in population domains (areas).

Basic setup

Data structure:
$\mathcal{U} = \{1, \dots, N\}$ is a finite population,
$A \subset \mathcal{U}$ is a probability sample of size n,
$B \subset \mathcal{U}$ is a non-probability sample of size N_B, and the set $A \cap B$ is abundant.



Data model:
The values y_i, $i \in A$, and contaminated values y_i^*, $i \in B$, are observed. Auxiliary vector values $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, $p \geq 1$, $i \in \mathcal{U}$, are available.

A model for a similarity of y to y^*:

$$y_i = f(y_i^*, \mathbf{x}_i) + \epsilon_i, \quad i \in B. \qquad (\mathcal{M})$$

Choices of measurement error model $(\mathcal{M})$ can be:

- linear regression;
- non-linear parametric model;
- non-parametric model.

Direct estimation in population domains

- Let $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M$ be the partition of the population into M non-overlapping domains, where $\mathcal{U}_m$ is of size N_m.
- We aim to estimate the domain totals

$$t_m = \sum_{i \in \mathcal{U}_m} y_i, \quad m = 1, \ldots, M.$$

- The probability samples $A_m = A \cap \mathcal{U}_m$ are of sizes $n_m \leq N_m$.
- If the sizes N_m are known, one can apply the direct estimators

$$\hat{t}_m^{\mathrm{H}} = \frac{N_m}{\widehat{N}_m} \sum_{i \in A_m} d_i y_i \quad \text{with} \quad \widehat{N}_m = \sum_{i \in A_m} d_i, \quad m = 1, \ldots, M,$$

  of the totals t_m, where $d_i = 1/\pi_i$ are design weights and π_i are the inclusion probabilities by the sampling design $\mathrm{p}(\cdot)$.
- The variances $\psi_m^{\mathrm{H}} = \mathrm{var}_{\mathrm{p}}(\hat{t}_m^{\mathrm{H}})$ may be too large for small n_m.

Elements of our methodology (main literature)

1. A stratification of $\mathcal{U}$ into B and $\mathcal{U}\backslash B$ by Kim & Tam (2021) suggests treating B as complete auxiliary information.
2. We extend the calibration estimation of Kim & Tam (2021) to the model-calibration approach of Wu & Sitter (2001).
3. The model-calibrated domain estimates are further modeled using, for example, the Fay & Herriot (1979) model.

Kim, J.-K., Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.* 89:382–401.

Wu, C., Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *JASA* 96:185–193.

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

Model-calibration approach

The model $(\mathcal{M})$ is fitted using the data $(y_i, y_i^*, \mathbf{x}_i)$, $i \in A \cap B$. Let $\hat{y}_i$, $i \in B$, be the predictions of y_i obtained from the fitted model.

The model-calibration approach by Wu & Sitter (2001) means to find the weights w_i, $i \in A$, in

$$\hat{t}_m^{\mathrm{MC}} = \sum_{i \in A_m} w_i y_i, \quad m = 1, \ldots, M,$$

minimizing the distance measure

$$\Phi_m = \sum_{i \in A_m} d_i \left(\frac{w_i}{d_i} - 1 \right)^2,$$

for each $m = 1, \ldots, M$, subject to certain area-specific *calibration constraints* as in Kim & Tam (2021) but where auxiliary data are used through the fitted values $\hat{y}_i$, $i \in B$.

Calibration constraints for incomplete auxiliary data

Let us introduce the indicator variable

$$\delta_i = \begin{cases} 1 & \text{if } i \in B, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that all intersections of the sets A_m and $B_m = B \cap \mathcal{U}_m$ are neither empty nor too small.

For each $m = 1, \ldots, M$, we find the weights $\{w_i,\ i \in A_m\}$ by minimizing the distance Φ_m subject to the calibration constraints

$$\sum_{i \in A_m} w_i \delta_i = N_{B_m}, \quad \sum_{i \in A_m} w_i \delta_i \hat{y}_i = \sum_{i \in B_m} \hat{y}_i,$$

and

$$\sum_{i \in A_m} w_i (1 - \delta_i) = N_m - N_{B_m},$$

where N_{B_m} is the size of the non-probability sample subset B_m.

Note: the auxiliary variable y^* is already integrated.

The data for the Fay–Herriot (FH) model (Fay & Herriot, 1979):

- The model-calibrated estimators $\hat{t}_m^{\mathrm{MC}}$ are treated as the direct estimators because they are approximately design-unbiased under certain conditions (Wu & Sitter, 2001).
- Estimators $\tilde{\psi}_m^{\mathrm{MC}}$ of the variances $\psi_m^{\mathrm{MC}} = \mathrm{var}(\hat{t}_m^{\mathrm{MC}})$.
- Exactly known area-level covariates $\mathbf{z}_m = (z_{m1}, \ldots, z_{mq})'$, $q \leq p$, selected from aggregates of auxiliary data $\mathbf{x}_i$, $i \in \mathcal{U}_m$.

The standard FH model is the linear mixed model

$$\hat{t}_m^{\mathrm{MC}} = \mathbf{z}_m' \boldsymbol{\beta} + v_m + \varepsilon_m, \quad m = 1, \ldots, M,$$

where $\varepsilon_m \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_m^{\mathrm{MC}})$ are sampling errors, $v_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2)$ are random area effects independent of ε_m, and $\boldsymbol{\beta}$ are fixed effects.

EBLUP based on the FH model

The empirical best linear unbiased predictions (EBLUPs) of the domain totals t_m, $m = 1, \ldots, M$, are expressed as the linear combinations

$$\hat{t}_m^{\mathsf{FH}} = \hat{\gamma}_m \hat{t}_m^{\mathsf{MC}} + (1 - \hat{\gamma}_m)\mathbf{z}_m' \hat{\boldsymbol{\beta}} \quad \text{with} \quad \hat{\gamma}_m = \frac{\hat{\sigma}_v^2}{\tilde{\psi}_m^{\mathsf{MC}} + \hat{\sigma}_v^2},$$

and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{m=1}^{M} \frac{\mathbf{z}_m \mathbf{z}_m'}{\tilde{\psi}_m^{\mathsf{MC}} + \hat{\sigma}_v^2} \right)^{-1} \sum_{m=1}^{M} \frac{\mathbf{z}_m \hat{t}_m^{\mathsf{MC}}}{\tilde{\psi}_m^{\mathsf{MC}} + \hat{\sigma}_v^2},$$

where $\hat{\sigma}_v^2$ is an estimator of the variance σ_v^2 of random area effects.

Note: for skewed data, a variance-stabilizing transformation – such as the logarithm – might be applied if needed.

Monthly turnover example

Aim – to improve direct *GREG* estimates of total turnover in NACE activity groups of service enterprises.

Main variables:

- y_i – monthly turnover;
- y_i^* – corresponding turnover derived from VAT data;
- x_{i1} – previous year's annual turnover.

Sample B size relative to A: $N_B/n \approx 6.2$.

Measurement error model $(\mathcal{M})$ – *linear regression* since y and y^* are continuous variables.

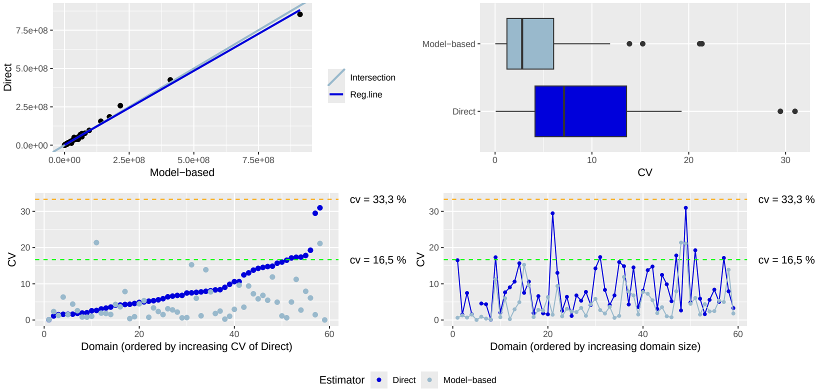


Figure 1(a): Comparison of direct estimates and EBLUPs for 2024 M11.

Good (CV $\leqslant$ 16.5%), sufficient (16.5% $<$ CV $\leqslant$ 33.3%), unreliable (CV $>$ 33.3%).

Monthly turnover example | multiple months

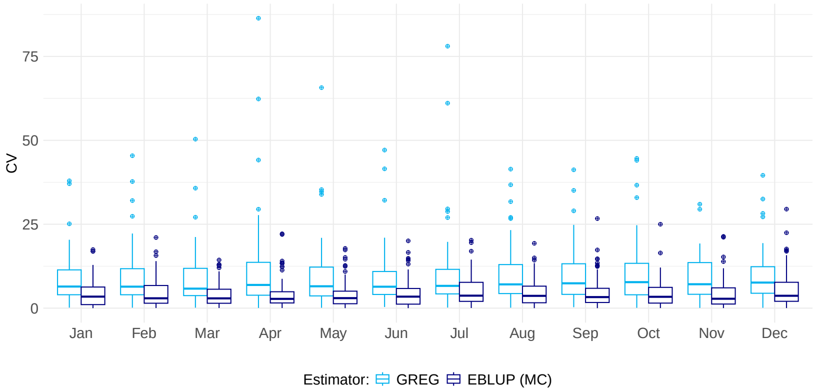


Figure 1(b): Comparison of direct GREG estimates and EBLUPs.

Annual investment example

Aim – to improve direct Horvitz–Thompson estimates of total investment in tangible assets in NACE activity groups.

Main variables:

- y_i – annual investment (sum over quarters);
- y_i^* – corresponding investment from a cut-off sample of the Structural Business Statistics;
- x_{i1} – previous year's annual investment from the Structural Business Statistics.

Sample B size relative to A: $N_B/n \approx 1.1$.

Measurement error model $(\mathcal{M})$ – *two-part approach:*
(a) logistic regression for the probability of zero;
(b) linear regression for the positive part;
since y and y^* are continuous variables with many zeros.

Annual investment example | a single year

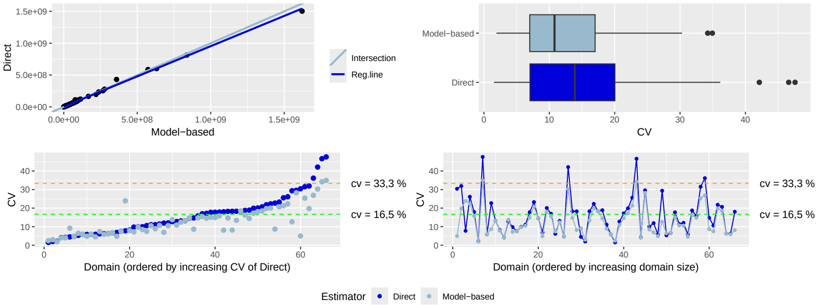


Figure 2(a): Comparison of direct estimates and EBLUPs for 2023.

Good (CV $\leqslant$ 16.5%), sufficient (16.5% $<$ CV $\leqslant$ 33.3%), unreliable (CV $>$ 33.3%).

Annual investment example | a single year

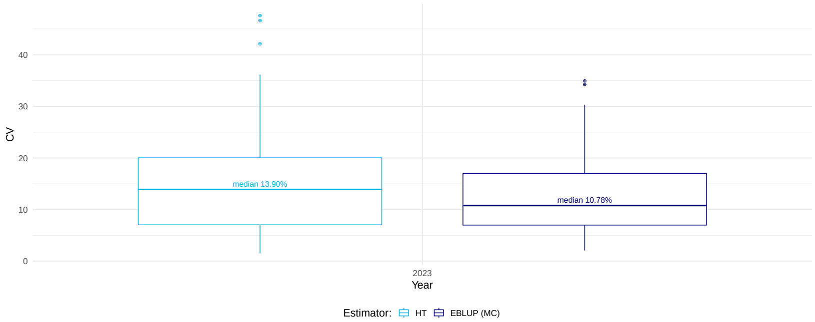


Figure 2(b): Comparison of Horvitz–Thompson estimates and EBLUPs.

Quarterly job vacancies example

Aim – to improve direct Hájek estimates of municipal job vacancies.

Main variables:

- y_i – job vacancies at the end of the quarter;
- y_i^* – corresponding online job advertisement information;
- x_{i1} – last month's number of employees.

Sample B size relative to A: $N_B/n \approx 1.8$.

Measurement error model $(\mathcal{M})$ – *non-parametric k-nearest-neighbors imputation* since y and y^* are count variables with many zeros.

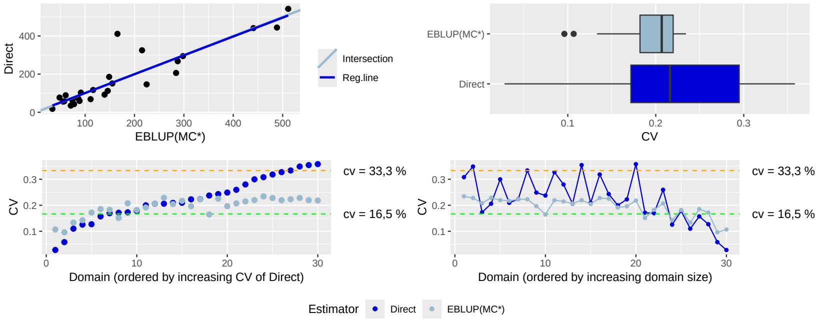


Figure 3(a): Comparison of direct estimates and EBLUPs for 2024 Q2.

Good (CV $\leqslant$ 16.5%), sufficient (16.5% $<$ CV $\leqslant$ 33.3%), unreliable (CV $>$ 33.3%).

Quarterly job vacancies example | multiple quarters

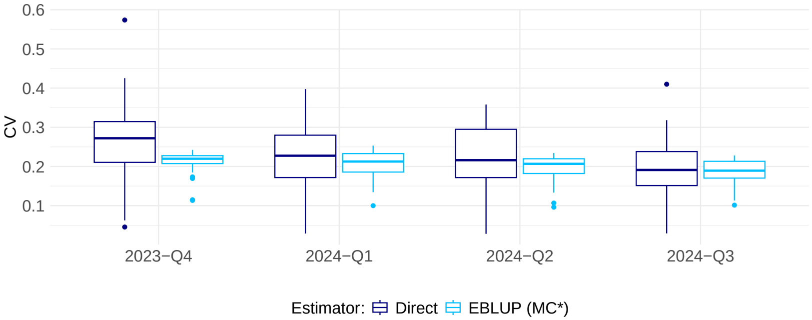


Figure 3(b): Comparison of direct Hájek estimates and EBLUPs.

Conclusions and suggestions

- Important variables for policy decisions often have analogs in other data sources, but these external data cover only a part of the target population. Our methodology provides a way to address this issue.
- The examples show that the improvement depends on how large the non-probability sample is relative to the probability sample.
- As a second step, for key variables, estimation in domains might be considered through small area estimation techniques.

State data agency
Statistics Lithuania

Thank you!

Small area estimation using incomplete auxiliary information

Donatas Šlevinskas[1,2], Andrius Čiginas[1,2], Ieva Burakauskaitė[1,2]

[1]State Data Agency (Statistics Lithuania)

[2]Vilnius University