State data agency
Statistics Lithuania

Combining online job advertisements with probability sample data for enhanced small area estimation of job vacancies

Donatas Šlevinskas, Andrius Čiginas, Ieva Burakauskaitė

Outline

- Introduction
  - Job vacancy estimation in domains
  - Integration of NP samples
- Methodology
  - The case of NP based on OJA
  - Further small area estimation modeling
  - EBLUP based on the FH model
- Empirical results
  - Effectiveness for a single quarter
  - Effectiveness for multiple quarters

Auxiliary information in sample surveys

- What is the main product of National Statistical Institutes (NSIs)? *Official statistics*.
- NSIs aim for improvement:
  - → **by timeliness** – more frequent estimates,
  - → **by granularity** – more detailed level estimates.
- Typically sample designs are optimized for population-level estimates. Small domains often have:

limited or unplanned sample coverage

↓

small sample sizes

↓

high variability or unreliable estimates

- A possible solution: incorporate administrative data or other non-traditional data sources (mobile network, social media, etc.) to supplement existing probability sample data.

Job vacancy data and the problem

Data source	Target variable, y	Auxiliary data, $\mathbf{x}$
NP sample	×	✓
P sample	✓	✓

- Probability sample data on job vacancies in companies are collected in the quarterly Statistical survey on earnings.
- There is complete administrative information on the monthly number of employees, economic activity, etc.
- Transformed online job advertisement (OJA) data:
  - only partially covers the survey population;
  - as non-probability (or big data) sample is not representative;
  - roughly approximates job vacancies by nonlinear relationship.

Direct job vacancy estimation in domains

- Let $\mathcal{U}$ be the finite population and $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M$ be its partition into M non-overlapping domains, $|\mathcal{U}_m| = N_m$.
- The aim is to estimate domain totals

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

- The probability sample A_m is of size $n_m \leq N_m$ in the m-th domain.
- The inaccuracy of the estimator can also be expressed using the Coefficient of Variation (CV):

$$\mathrm{CV}(\hat{t}_m) = \sqrt{\mathrm{var}(\hat{t}_m)}/\hat{t}_m.$$

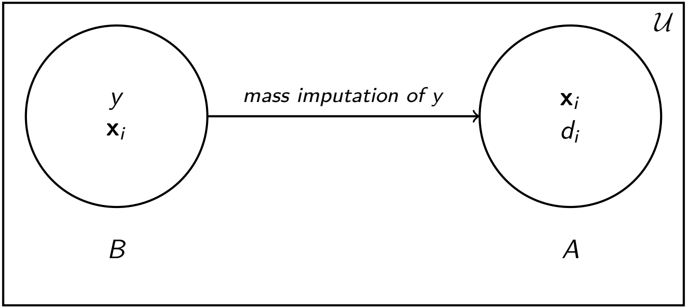
- If the sizes N_m are assumed to be known, the direct Hájek estimators of the totals t_m are

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

  where $d_i = 1/\pi_i$ are design weights and π_i are the first-order inclusion probabilities.
- The variances $\psi_m^{\mathrm{H}} = \mathrm{var}(\hat{t}_m^{\mathrm{H}})$ may be too large for small n_m.

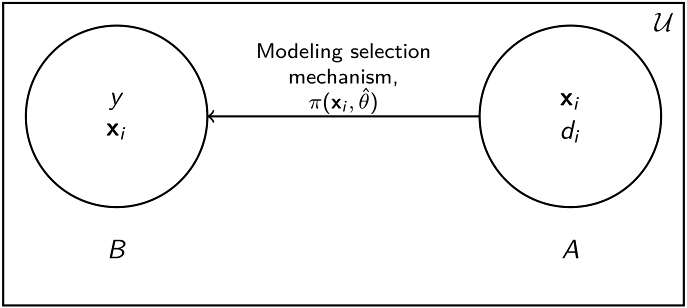
Possible cases of NP integration

- A and B – probability and non-probability samples respectively,
- y_i – the target variable for which a parameter (such as total, mean, or quantile) needs to be estimated,.
- $\mathbf{x}_i$ – auxiliary covariate vector,
- d_i – design weight of ith unit.



Possible cases of NP integration (2)

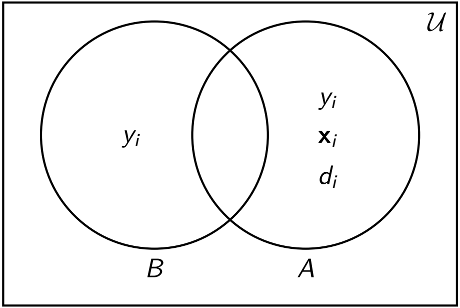
- A and B – probability and non-probability samples respectively,
- y_i – the target variable for which a parameter (such as total, mean, or quantile) needs to be estimated,.
- $\mathbf{x}_i$ – auxiliary covariate vector,
- d_i – design weight of ith unit,
- $\pi(\mathbf{x}_i, \hat{\theta}), i \in B$ – estimated propensity scores.



Possible cases of NP integration (3)

Kim & Tam (2021) regression data integration estimator:

- δ_i – inclusion into B indicator,
- w_i – calibrated weight of unit ith,
- $D(\cdot,\cdot)$ – distance function.



$$\hat{t}_m^{RegDI} = \sum_{i \in A_m} w_i \, y_i,$$

$$\mathbf{w} = \arg\min_{\mathbf{w}} D\big(\mathbf{w}, \mathbf{d}\big),$$

subject to:

$$\sum_{i \in A_m} w_i \, \delta_i = N_{B_m},$$

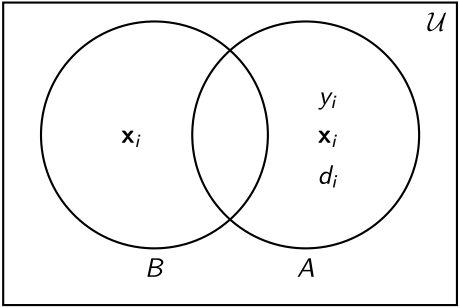
$$\sum_{i \in A_m} w_i \, \big(1 - \delta_i\big) = N_m - N_{B_m},$$

$$\sum_{i \in A_m} w_i \, \delta_i \, y_i = \sum_{i \in B_m} y_i.$$

The case of NP based on OJA

Modified regression data integration estimator based on model-calibration: *(Wu & Sitter, 2001)*

- δ_i – inclusion into B indicator,
- $D(\cdot, \cdot)$ – distance function,
- $\hat{\mu}_i = \hat{\mu}_i(\mathbf{x}_i, \hat{\theta})$ – predictions of y_i based on model that was fitted on $A \cap B$ data.



$$\hat{t}_m^{\text{MC}} = \sum_{i \in A_m} w_i \, y_i,$$

$$\mathbf{w} = \arg\min_{\mathbf{w}} D\Big(\mathbf{w}, \mathbf{d}\Big),$$

subject to:

$$\sum_{i \in A_m} w_i \, \delta_i = N_{B_m},$$

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

$$\sum_{i \in A_m} w_i \, \delta_i \, \hat{\mu}_i = \sum_{i \in B_m} \hat{\mu}_i.$$

Further small area estimation modeling

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

- The model-calibrated estimators $\hat{t}_m^{\mathrm{MC}}$ 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 $\mathbf{z}_m = (z_{m1}, \ldots, z_{mq})'$, $q \le p$, selected from aggregates of auxiliary data $\mathbf{x}_i$, $i \in \mathcal{U}_m$.

The standard FH model is the linear mixed model

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

where $\varepsilon_m \overset{\mathrm{ind}}{\sim} \mathcal{N}(0, \psi_m^{\mathrm{MC}})$ are sampling errors, $v_m \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma_v^2)$ are random area effects independent of ε_m, and β 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 *(Fay & Herriot, 1979)*

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

and

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

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

For data like job vacancies, the standard FH model should be applied to the log-transformed estimators *(Rao & Molina, 2015)*

$$\log(\hat{t}_m^{\mathrm{MC}}) \quad \text{with} \quad \mathrm{var}\left(\log(\hat{t}_m^{\mathrm{MC}})\right) \approx (\hat{t}_m^{\mathrm{MC}})^{-2}\, \mathrm{var}(\hat{t}_m^{\mathrm{MC}}).$$

Effectiveness for a single quarter

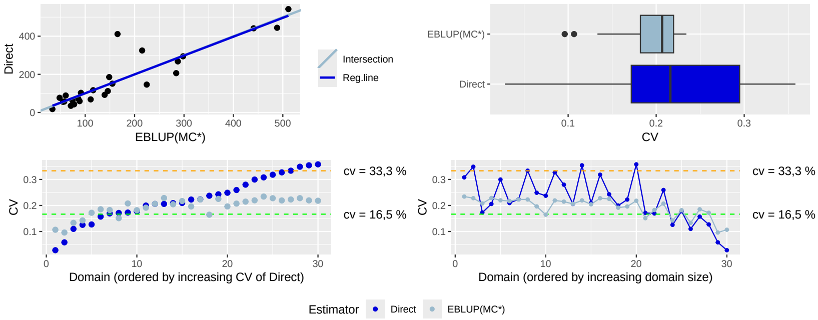


Figure 1: Comparison of direct estimates and EBLUPs for a period of 2024 Q2.
Note: good (CV $\leqslant$ 16.5%), sufficient (16.5% $<$ CV $\leqslant$ 33.3%), unreliable (CV $>$ 33.3%)

Effectiveness for multiple quarters

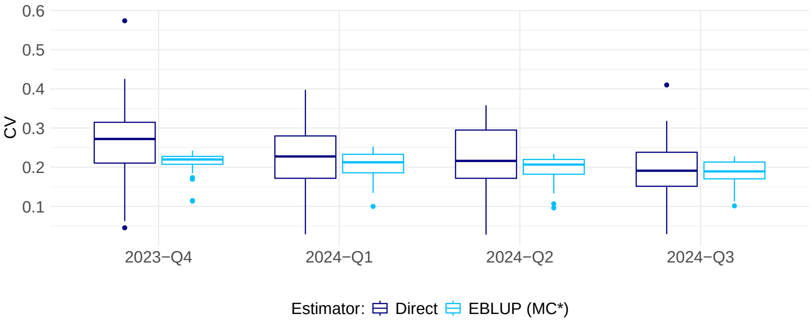


Figure 2: Comparison of direct estimates and EBLUPs.

Effectiveness for multiple quarters (2)

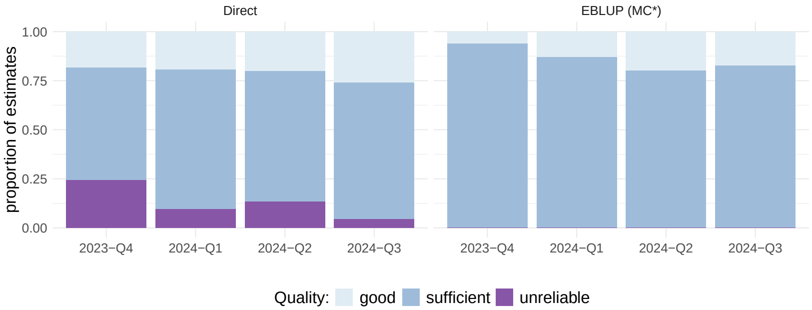


Figure 3: Trends in Direct and EBLUP estimates by quality groups.

Note: good (CV $\leqslant$ 16.5%), sufficient (16.5% $<$ CV $\leqslant$ 33.3%), unreliable (CV $>$ 33.3%).

Tools overview

- Initial preprocessing and record linkage:
  - Performed in Python using Spark for efficient data processing.
- Model building and model calibration estimates:
  - Conducted in R using the `StatMatch` and `survey` packages.
- Final EBLUP estimates and diagnostics:
  - Generated using the `emdi` package in R for small area estimation.

Literature sources

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

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.

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

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

State data agency
Statistics Lithuania

Thank you for attention

State data agency
Statistics Lithuania

Combining online job advertisements with probability sample data for enhanced small area estimation of job vacancies

Donatas Šlevinskas, Andrius Čiginas, Ieva Burakauskaitė