Estimation of job vacancies in small population domains using web-scraped data

Ieva Burakauskaitė,
Andrius Čiginas,
Donatas Šlevinskas

Vilnius University (Vilnius, Lithuania)

Joint Statistical Meetings | August 7, 2024

Motivation

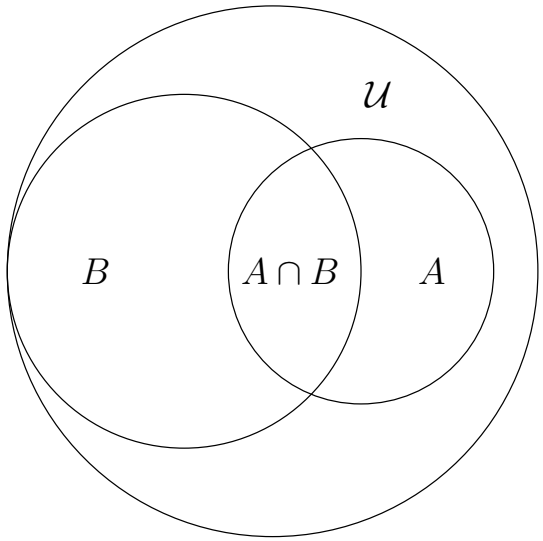
- **Quarterly job vacancy estimates** are obtained from a **probability sample** of enterprises of the Statistical survey on earnings carried out by Statistics Lithuania.

  However, the latter sample is **not designed for the estimation of population parameters in small domains** such as municipalities. The sample sizes are insufficient in some domains since a direct estimator produces estimates with large variances.

- A few **additional sources of information** might be used to increase the estimation accuracy in the small domains, that is,
  - ***complete administrative information*** on the monthly number of employees, economic activity, etc.,
  - ***weekly web-scraped online job advertisements (OJA) data*** (non-probability sample). Even though it covers the survey population only partially and is not representative, it still roughly approximates job vacancies.

Basic setup

- $\mathcal{U} = \{1, \ldots, N\}$ is a finite population,
- $A \subset \mathcal{U}$ is a probability sample of size n from which the values y_i, $i \in A$, of the study variable y are collected,
- $B \subset \mathcal{U}$ is a bigger non-probability sample of size N_B from which the values y_i^*, $i \in B$, of the contaminated variable y^* are collected.



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

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

Direct estimation in population domains

- Suppose the probability sample $A_m = A \cap \mathcal{U}_m$ is of size $n_m \leqslant N_m$ in the mth domain.
- If the domain sizes N_m are assumed to be known, the direct Hájek estimators of the totals t_m are

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

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

Auxiliary data and measurement error model

- A stratification of $\mathcal{U}$ into B and $\mathcal{U}\backslash B$ by Kim & Tam (2021) suggests treating B as complete auxiliary information.
- We assume that for each $i \in \mathcal{U}$ there are known auxiliary vector values $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $p \geqslant 1$.
- A model utilizing a similarity of the contaminated variable y^* to the study variable y,

$$y_i \sim (y_i^*, \mathbf{x}_i), \quad i \in B, \tag{$\mathcal{M}$}$$

  is referred to as the measurement error model and might be:
  - a linear regression,
  - a non-linear parametric model, or
  - a non-parametric (e.g., k-nearest neighbors) model.

Model-calibration approach

- 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.
- Another estimator of the totals t_m is obtained through the model-calibration approach by Wu & Sitter (2001) as

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

where the weights w_i, $i \in A_m$, minimize 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 domain-specific calibration constraints built as in Kim & Tam (2021), where auxiliary data is used through the fitted values $\hat{y}_i$, $i \in B$.

Calibration constraints for incomplete auxiliary data

- Let $\delta_i = \mathbb{I}(i \in B)$ be the indicator variable denoting whether an element $i \in \mathcal{U}$ is in the non-probability sample B or not.
- 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.

FH model for small area estimation

- In the current setting, the components of the standard Fay–Herriot (FH) model (Fay & Herriot, 1979) are
  - the model-calibrated estimators $\hat{t}_m^{\mathrm{MC}}$, treated as the direct estimators since they are approximately design-unbiased under certain conditions (Wu & Sitter, 2001),
  - estimators $\widetilde{\psi}_m^{\mathrm{MC}}$ of the variances $\psi_m^{\mathrm{MC}} = \mathrm{var}_{\mathrm{p}}(\hat{t}_m^{\mathrm{MC}})$,
  - area-level covariates $\mathbf{z}_m = (z_{m1}, \ldots, z_{mq})'$, $q \leqslant p$, selected from aggregates of the known auxiliary data $\mathbf{x}_i$, $i \in \mathcal{U}$.

- The 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 effects independent of ε_m, and $\boldsymbol{\beta}$ is a vector of unknown fixed-effect parameters.

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)

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

and

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

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

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

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

Application to job vacancy data of 2023 Q1

- The population $\mathcal{U}$ consists of $N = 34\,087$ enterprises, and is partitioned into $M = 60$ municipalities.
- The stratified simple random sample A is of size $n = 7\,051$, where y_i, $i \in A$, are job vacancies at the end of the quarter.
- The weekly OJA data are transformed in such a way that
  1. The number of new OJAs is evaluated and recorded for each identified enterprise;
  2. Zeros are assigned to a number of previous and subsequent weeks with no records;
  3. The data of several last weeks of a quarter are aggregated.

  The derived values y_i^* represent the job vacancies from the non-probability sample B of size $N_B = 13\,372$.
- There are $3\,669$ observations (y_i, y_i^*, x_i) in the set $A \cap B$, where x_i are the number of employees in the last month of the quarter.

Applying the model-calibration approach

We have considered two models for count variables y and y^*:

- a parametric zero-inflated negative binomial regression,
- a non-parametric k-nearest neighbors (kNN) imputation,

and the latter is chosen. The auxiliary variables (y^*, x) are used to find $k = 3$ nearest neighbors in the set $A \cap B$, and the average of their values y_i is the prediction $\hat{y}_i$, $i \in B$.

In smaller areas, the intersections $A_m \cap B_m$ are sometimes small and zero values comprise a significant part of y and y^*, hence we apply the model-calibration estimators only to *the largest 20 municipalities*, and use the Hájek estimators for the rest.

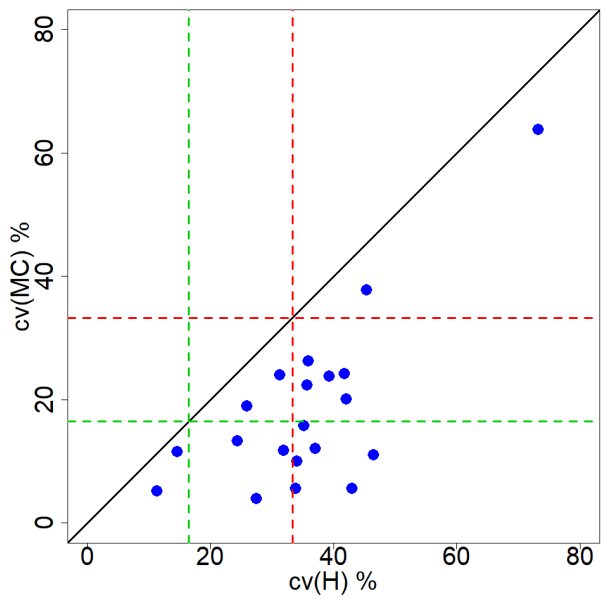


Figure 1: Comparing estimates of the coefficients of variation: *MC* vs. *H*.

Applying the EBLUP approach (I)

- The FH model is applied to the log-transformed combined model-calibration and Hájek estimates, with $z_m = \log(\sum_{i \in \mathcal{U}_m} x_i)$.
- The aggregated number of employees is a good predictor of job vacancies at the area level, and the obtained EBLUPs drastically improve the results. It would seem to be enough to model the Hájek estimates alone.
- However, the combined inclusion of OJA data with the aggregated number of employees into the modeling is observed to better increase the accuracy of estimation.

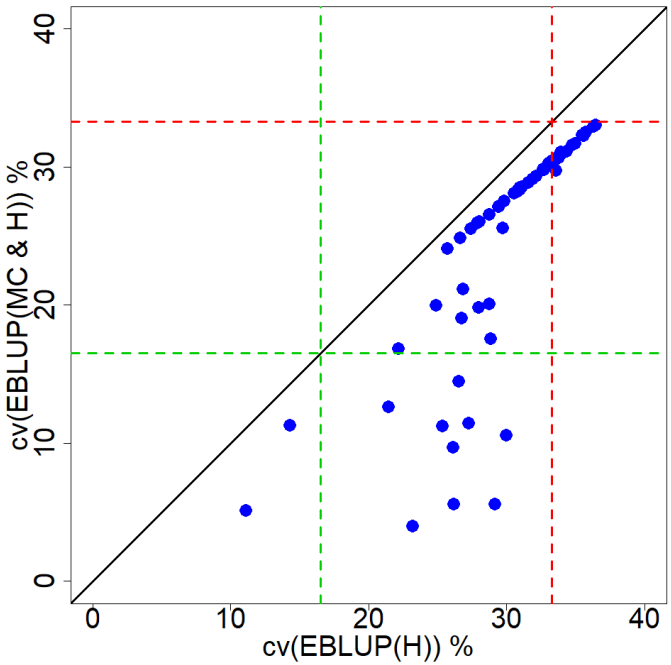


Figure 2: Comparing estimates of the coefficients of variation: *EBLUP(MC & H)* vs. *EBLUP(H)*.

Applying the EBLUP approach (II)

Table 1: Comparing the five-number summary of estimates of the coefficients of variation (in percent).

Estimator	Min.	1st Quartile	Median	3rd Quartile	Max.
Hájek	11.33	37.22	48.78	63.07	109.10
FH (H)	11.10	27.33	30.66	33.43	36.45
FH (MC & H)	3.98	19.88	28.12	30.38	33.05

Final remarks

- Given an additional variable observed in a non-probability sample and close to the study variable, we present a general methodology for how it can be used to refine the estimation of totals (or means) in small population domains. The methodology circumvents the problems of auxiliary data incompleteness and bias.
- In the application, we integrate the OJA data with the probability sample data to estimate the job vacancy totals in municipalities. The overall improvement in accuracy over other standard estimators depends on how many areas are sufficiently covered by the non-probability sample.
- The application also shows how important administrative information commonly used in official statistics can be when utilized in small area estimation models.

References

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(366):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(2):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(453):185–193.