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

Andrius Čiginas[1,2]

(joint work with Donatas Šlevinskas[1,2] & Ieva Burakauskaitė[1,2])

[1]State Data Agency (Statistics Lithuania), [2]Vilnius University

European Conference on Quality in Official Statistics 2024

Auxiliary information in sample surveys

- Auxiliary data is important for efficient survey planning and estimation. Especially if it is related to study variables well.
- Even administrative data does not necessarily fully cover survey populations.
- Data that do not fully cover the populations:
  - a probability samples;
  - b non-probability samples;
  - c big data samples.

In the context of domain-level modeling in small area estimation:

- Estimated covariates may be unavailable for some domains, especially in case a.
- Estimated covariates are biased in cases b and c.

Job vacancy data and the problem

- 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 very partially covers the survey population;
  - as non-probability (or big data) sample is not representative;
  - roughly approximates job vacancies.

We estimate the total job vacancies in municipalities of Lithuania. Applying direct design-based estimators like Horvitz–Thompson or Hájek estimators, the five-number summary for estimates of the coefficients of variation is, for instance,

$$(11.33, 37.22, 48.78, 63.07, 109.1)$$

in percents.

How to exploit incomplete but related auxiliary data

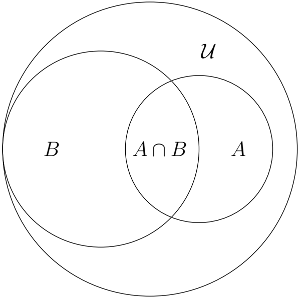
$\mathcal{U} = \{1, \ldots, N\}$ is a finite population, $A \subset \mathcal{U}$ is a probability sample of size n, $B \subset \mathcal{U}$ is a bigger non-probability sample of size N_B, and $A \cap B$ is abundant.

A stratification of $\mathcal{U}$ into B and $\mathcal{U} \backslash B$ by Kim & Tam (2021) suggests treating B as complete auxiliary information.

The values y_i, $i \in A$, of the study variable y and the values y_i^*, $i \in B$, of the contaminated variable y^* are collected. There are known auxiliary vector values $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $p \geqslant 1$, $i \in \mathcal{U}$.

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

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



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

- a linear regression;
- a non-linear parametric model;
- a non-parametric (nearest neighbor) 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 the area $\mathcal{U}_m$ contains N_m elements.
- We aim to estimate the domain totals
$$t_m = \sum_{i \in \mathcal{U}_m} y_i, \quad m = 1, \ldots, M.$$
- The probability sample $A_m = A \cap \mathcal{U}_m$ is of size $n_m \leqslant N_m$ in the mth domain.
- If the sizes N_m are assumed to be known, the direct Hájek estimators of the totals t_m are
$$\hat{t}_m^{\mathrm{H}} = \frac{N_m}{\widehat{N}_m} \sum_{i \in A_m} d_i y_i \quad \text{with} \quad \widehat{N}_m = \sum_{i \in A_m} d_i, \quad m = 1, \ldots, M,$$
where $d_i = 1/\pi_i$ are design weights and π_i are the first-order 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.

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 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 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.

Further small area estimation modeling

Note: the auxiliary variable y^* close to the study variable y is already integrated through the model-calibration estimators $\hat{t}_m^{\mathrm{MC}}$.

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 $\psi_m^{\mathrm{MC}} = \mathrm{var}_{\mathrm{p}}(\hat{t}_m^{\mathrm{MC}})$.
- Exactly known area-level covariates $\mathbf{z}_m = (z_{m1}, \ldots, z_{mq})'$, $q \leqslant p$, selected from aggregates of auxiliary data $\mathbf{x}_i$, $i \in \mathcal{U}$.

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, $m = 1, \dots, M$, 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)

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

Application to job vacancies. Data

- The population $\mathcal{U}$ of companies is of size $N = 34\,087$ in the first quarter of 2023. There are $M = 60$ municipalities in $\mathcal{U}$.
- A 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 only component in the auxiliary vector $\mathbf{x}_i$ that we use is the number of employees in the last month of a quarter.
- The scraped weekly OJA data are transformed:
  1. the number of new OJAs is evaluated and recorded for each identified company;
  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 summed.

  The derived values y_i^* represent the non-probability sample B of size $N_B = 12\,528$.
- There are $3\,468$ observations $(y_i, y_i^*, \mathbf{x}_i)$ in the set $A \cap B$.

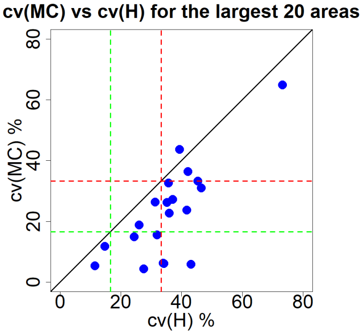
Applying the model-calibration estimators

Candidate models for count variables y and y^*:

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

The choice is the kNN model, where the vector characteristics $(y_i^*, \mathbf{x}_i)$ are used to find the $k = 3$ nearest neighbors in the set $A \cap B$ whose average of values y_i is the prediction $\hat{y}_i$, $i \in B$.

Since in smaller domains, the intersections $A_m \cap B_m$ are sometimes small and there is a dominance of zero values of y and y^*, we apply the model-calibration estimators only to the largest 20 municipalities (by size N_m) and use the Hájek estimators for the rest.



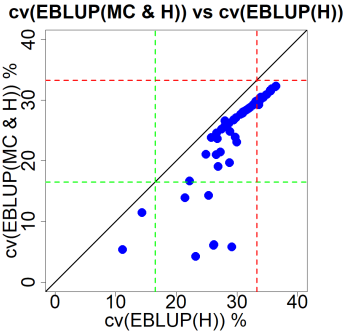
Applying EBLUP

The FH model is applied to the log-transformed combined model-calibration and Hájek estimates, where the covariates are $\mathbf{z}_m = \log(\sum_{i \in \mathcal{U}_m} \mathbf{x}_i)$. The five-number summary of the estimates of the coefficients of variation is

$$(4.28, 23.39, 27.71, 29.88, 32.33)$$

in percents.

The aggregated number of employees is a powerful predictor of job vacancies. Therefore, you might ask: would it be enough to model the Hájek estimates alone? That is, what are the benefits of OJA data?



Conclusions

- 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 estimates 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.* 74:269–277.

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.