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

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

[1]*State Data Agency (Statistics Lithuania), Lithuania*

[2]*Vilnius University, Lithuania*

Abstract

We combine the probability sample data on job vacancies with online job advertisements (OJA) information and administrative data to improve the estimates of job vacancy totals in small population domains like municipalities. Since OJA data is a non-probability sample covering only a limited part of the survey population and its selection mechanism is unknown, we apply non-probability sample integration techniques to incorporate this information properly into the small area estimation models. The methodology proposed based on this application can be used in other estimation problems where incomplete additional information is available from administrative or alternative data sources.

Keywords: non-probability sample, nearest neighbor imputation, model-calibration, small area estimation, area-level model

1. Introduction

Auxiliary information available in probability sample surveys is important in obtaining as accurate parameter estimates in the finite population and its domains as possible. Having auxiliary data related to the study variables at the unit or domain (area) level provides a range of models to choose from that can improve the direct design-based estimates. Estimation approaches supported by models in different ways are well developed for estimating parameters in the population or its larger domains (Särndal et al., 1992; Wu & Thompson, 2020) and in small population areas (Rao & Molina, 2015).

The classical literature on survey statistics usually deals with idealized additional information: the values of auxiliary variables are assumed to fully cover the survey population at a detailed or at least some aggregated level. However, even the administrative data sources commonly used in official statistics often cover only a part of the population and, therefore, standard methods like calibration estimators (Deville & Särndal, 1992) cannot be applied immediately. In addition to administrative data, many other potentially useful sources do not meet the ideal (completeness or coverage) conditions, such as sample data from other probability surveys, non-probability samples, and big data samples. If earlier there were only separate attempts to properly utilize (integrate) all these types of additional data for estimation of population parameters, in recent years such research has been rapidly developed, as reviewed in Yang and Kim (2020).

In the context of estimation in small population domains, when using aggregated auxiliary data in area-level models like in the famous Fay–Herriot (FH) model (Fay & Herriot, 1979), there are also specific challenges to employing incomplete auxiliary information. Ybarra and Lohr (2008) use unbiased estimators of auxiliary domain means obtained from the same or other preferably larger probability samples as covariates in area-level models. They show that a naive application of the FH model can lead to worse results than the direct estimation. They propose to use a modified version of the FH model allowing for measurement error in the auxiliary aggregates. A drawback of this approach is that the estimated covariates may be unavailable for some domains if the auxiliary variables are taken from samples that are not large enough. Although non-probability samples can be much larger than probability samples, it can be difficult to ensure that the estimators of auxiliary domain characteristics based on them would be approximately unbiased or to assess the biases of such estimators. This difficulty is due to the typically unknown selection mechanism of the non-probability samples (Rao, 2021; Wu, 2022). These biases cannot be ignored (Meng, 2018) and make the method of Ybarra and Lohr (2008) not immediately applicable because it requires the estimated mean squared errors (MSEs) for the estimated covariates. Among exceptions is the application of Marchetti et al. (2015), where a big data covariate is based on a non-probability sample treated as a simple random sample. Some more recent applications using area-level big data as additional predictors in various models are reviewed in Rao (2021). It should be noted that some applications ignore population coverage errors in certain big data variables, such as those compiled from Google Trends, mobile operators, or social network data. Administrative data covering almost the entire survey population can be treated in the same way, but this is more reasonable.

In the application that motivates our study, the coverage of the population in big data is only very partial and it is a non-probability (voluntary) sample with an unknown selection mechanism. We have a quarterly probability sample of Lithuanian companies, from which the sums (totals) of job vacancies in municipalities are estimated. Apart from administrative information such as the monthly number of employees known for all companies in the population, we are most interested in the online job advertisement (OJA) data. The latter information, scraped weekly from the major job posting portals, is more related to the study variable. No matter how transformed, the OJA data cannot replace the values of the study variable but can be utilized as auxiliary information in modeling.

Kim and Tam (2021) assume the linear regression relationship between the variable from a big non-probability sample and the study variable. This is a unit-level measurement error model. To estimate the population total, their idea is to stratify the population into a big data stratum and a missing data stratum. Then they apply the calibration estimation method with different conditions imposed on these artificial strata to exploit the big data sample as complete auxiliary information. Such Deville and Särndal (1992) type calibration could be applied to each population domain or small area separately, but it is not suitable in general if the underlying unit-level dependence between the variables is not linear as a non-linear relation between the count type variables in our application.

We use the idea of Kim and Tam (2021) but through the model-calibration (MC) approach of Wu and Sitter (2001), which allows more general underlying unit-level models. The MC approach to improving the direct probability sample-based estimates in small areas is based on the predictions of the study variable in the big data stratum, which are further used in calibration constraints. Under certain conditions, the MC estimators of the population totals are asymptotically design-unbiased (Wu & Sitter, 2001). Due to this property, our estimation approach outlined in Section 2 has the second step modeling the MC estimates using exactly known area-level covariates.

The application exploiting the incomplete OJA data and complete auxiliary information on the number of employees is presented in Section 3. Since it is complicated to make strong parametric model assumptions for unit-level measurement errors, we apply a non-parametric nearest neighbor imputation model to predict the job vacancies in the big data stratum defined by the available OJA data. Then the model-calibration of the probability sample weights is applied separately in each municipality. The MC estimates are further modeled using the FH model to obtain the empirical best linear unbiased predictions (EBLUPs), where the aggregated number of employees appears to be a good explanatory variable. We summarize our research findings in Section 4.

2. Methodology outline

Consider a finite population $U = \{1, \ldots, N\}$ of size N. Let A be a probability sample of size n drawn from U according to a probability sampling design with the first-order inclusion probabilities π_i, $i \in A$, and the values y_i, $i \in A$, of the study variable y are collected. We suppose that the vector values $x_i = \left(x_{i1}, \ldots, x_{ip}\right)'$, $p \geq 1$, of the auxiliary variables x are known for $i \in U$.

Let $U = U_1 \cup \cdots \cup U_M$ be the partition of the population into the non-overlapping domains, where the area U_m contains N_m elements. Then the domain sample $A_m = A \cap U_m$ is of size $n_m \leq N_m$. We aim to estimate the domain totals

$$t_m = \sum_{i \in U_m} y_i \,, \qquad m = 1, \dots, M. \tag{1}$$

If the probability sampling design does not ensure fixed domain sample sizes n_m, they can be too small to get sufficiently accurate direct estimates. It means that the Hájek estimators

$$\hat{t}_m^{\mathrm{H}} = \frac{N_m}{\widehat{N}_m} \sum_{i \in A_m} d_i y_i \quad \text{with} \quad \widehat{N}_m = \sum_{i \in A_m} d_i \,, \qquad m = 1, \dots, M, \tag{2}$$

of totals (1), where $d_i = 1/\pi_i$ are design weights and the numbers N_m are assumed to be known, or other design-based estimators utilizing auxiliary data x may have too high variances.

In addition, a larger sample $B \subset U$ of size N_B is available, but its selection mechanism is unknown. In the latter non-probability sample, the values of y are measured with an error. We assume that the samples A and B are linked at the unit level, and the values y_i of the study variable y can be related to the values y_i^* of the contaminated variable y^* observed in the sample B through a parametric or non-parametric measurement error model, which possibly uses some of the variables x as covariates as well. It can be a simple linear regression model as in Kim and Tam (2021), more general non-linear parametric models considered in Wu and Sitter (2001), or non-parametric nearest neighbor models (Yang et al., 2021).

If one of these models is suitable to describe the relationship between the variables y and y^* and the intersection $A \cap B$ is abundant enough for model fitting, we can apply the MC methodology according to Wu and Sitter (2001), where auxiliary information is used through the fitted values of y. Let $\hat{y}_i$, $i \in B$, be the predictions of y_i obtained from the fitted model. The next step in our calibration version is to find the weights w_i, $i \in A$, in

$$\hat{t}_m^{\mathrm{MC}} = \sum_{i \in A_m} w_i y_i \,, \qquad m = 1, \dots, M, \tag{3}$$

minimizing an average distance between the sets $\{w_i,\, i \in A_m\}$ and $\{d_i,\, i \in A_m\}$ for each $m = 1, \dots, M$, subject to certain area-specific calibration constraints built as in Kim and Tam (2021). To construct the weights, let us introduce the indicator variable

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

which is also observed in the probability sample A. Suppose that at least the non-probability sample B is large enough that all intersections of the sets A_m and $B_m = B \cap U_m$ are not empty. Otherwise, or if some intersections appear too small, we can apply simpler Hájek estimators (2) or other design-based estimators in the respective domains. For each $m = 1, \ldots, M$, we find the weights $\{w_i,\ i \in A_m\}$ by minimizing the distance measure

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

subject to the constraints

$$\sum_{i \in A_m} w_i\, \delta_i = N_{B_m}, \quad \sum_{i \in A_m} w_i\, \delta_i \hat{y}_i = \sum_{i \in B_m} \hat{y}_i\,, \quad \text{and} \quad \sum_{i \in A_m} w_i (1 - \delta_i) = N_m - N_{B_m},$$

where N_{B_m} is the size of the non-probability sample subset B_m. The variances of estimators (3) are estimated by applying standard linearization or replication methods. In practice, the variance estimates can be calculated using the function `calibrate` from `R` package `survey` (Lumley, 2010).

We treat the MC estimators (3) as approximately design-unbiased estimators of the domain totals (1). Then one can use them as the direct estimators in the standard FH model of Fay and Herriot (1979) or its extensions (Rao & Molina, 2015) to build EBLUPs of (1). The area-level covariates used in these models are selected from the aggregated auxiliary variables x. Recent work by Harmening et al. (2023) provides convenient tools for applying the most common variants of the FH model using `R` package `emdi`. Let us denote by $\hat{t}_m^{\mathrm{FH}}$, $m = 1, \ldots, M$, the EBLUPs of the domain totals (1).

3. Application

Data on job vacancies are collected in one of the statistical surveys of the State Data Agency (Statistics Lithuania). We demonstrate the application of the presented methodology to a probability sample of companies for the first quarter of 2023. This is a stratified simple random sample A of size $n = 7\ 051$ drawn from the population U of size $N = 34\ 087$. The observed values y_i of the study variable y are job vacancies at the end of the quarter in the sampled companies. A dataset of completely known auxiliary data x contains supplementary variables such as the number of employees and some variables indicating economic activity available from administrative data sources and statistical registers.

We aim to estimate the totals of the variable y in $M = 60$ municipalities. As the total sample size is not uniformly distributed across municipalities and the variance of the study variable y

is large, the five-number summary $(11.33, 37.22, 48.78, 63.07, 109.1)$ for estimates of the coefficients of variation (in percents) of the Hájek estimates (2) shows that the latter estimates cannot be published for most municipalities.

The scraped weekly OJA data are transformed to better approximate the variable y. Since only the number of new OJAs is evaluated and recorded for each identified company, we first choose to assign zeros to a number of previous and subsequent weeks (for example, for 13 and 26 weeks, respectively) with no records. Then, we derive the variable y^* by summing the data of several last weeks of a quarter (for example, 6 weeks). The determined values of y^* define the non-probability sample B of size $N_B = 12\,528$.

Both variables y and y^* are count variables with many zero values, and there are $3\ 468$ observations (y_i, y_i^*, x_i) in the intersection $A \cap B$. For the prediction part of the MC method, one option is to fit a parametric model like a zero-inflated negative binomial regression. However, the latter regression is sensitive to outliers, and its efficiency varies greatly depending on the quarter. Our chosen alternative – the nearest neighbor imputation model – works much more accurately and is robust to outliers. Together with the variable y^*, the number of employees in the last month of a quarter (the auxiliary variable from the dataset x) is used to find the 3 nearest neighbors in the set $A \cap B$ whose average of values y_i is the prediction $\hat{y}_i$.

However, the percentage coverage of municipalities in the non-probability sample varies, as seen from the five-number summary $(7.87, 29.69, 34.89, 39.43, 43.78)$. Therefore, the intersections $A_m \cap B_m$ are sometimes small in smaller areas. Moreover, such areas may be dominated by zero values of the variables y and y^*. For these reasons, we safely apply the MC estimators (3) only to the largest 20 municipalities (by size N_m) and use the Hájek estimators (2) for the rest. Therefore, compared to the estimates of the coefficients of variation for the Hájek estimates, the improvement is only for the estimates of already acceptable accuracy, according to the summary $(4.30, 31.84, 48.78, 63.07, 109.1)$ of the combined coefficients of variation.

The log-transformed combined MC and Hájek estimates are modeled using the FH model. The model covariates are log-transformed domain totals of the number of employees in the last month of a quarter. The number of employees is a good predictor of job vacancies at least at the area level, so the resulting EBLUPs $\hat{t}_m^{\mathrm{FH}}$ drastically improve the results. The five-number summary $(4.28, 23.39, 27.71, 29.88, 32.33)$ of the estimates of the coefficients of variation shows that, with a precision warning, all EBLUPs can be published. If only the Hájek estimates are modeled analogously, the respective accuracy summary $(11.1, 27.33, 30.66, 33.43, 36.45)$ shows worse estimation results.

4. Conclusions

It happens in sample surveys that an additional variable close in nature to the study variable is not completely known in the population, or at least in a large probability sample drawn from it. We present a fairly general methodology for how a variable observed in a non-probability sample can be used to refine the estimation of totals (or means) in small population domains. This is the way to solve the problems of data incompleteness and bias.

In the considered 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 direct design-based 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

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, *87*(418), 376-38. https://doi.org/10.2307/2290268

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. https://doi.org/10.2307/2286322

Harmening, S., & Kreutzmann, A.-K., & Schmidt, S., & Salvati, N., & Schmid, T. (2023). A framework for producing small area estimates based on area-level models in R. *R J.*, *15*(1), 316-341. https://doi.org/10.32614/RJ-2023-039

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. https://doi.org/10.1111/insr.12434

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: John Wiley & Sons.

Marchetti, S., & Giusti, C., & Pratesi, M., & Salvati, N., & Giannotti, F., & Pedreschi, D., & Rinzivillo, S., & Pappalardo, L., & Gabrielli, L. (2015). Small area model-based estimators using big data sources. *J. Off. Stat.*, *31*(2), 263-281. http://dx.doi.org/10.1515/JOS-2015-0017

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.*, *12*(2), 685-726. https://doi.org/10.1214/18-aoas1161sf

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*(1), 242-272. https://doi.org/10.1007/s13571-020-00227-w

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* (2 ed.). New Jersey: John Wiley & Sons. https://doi.org/10.1002/9781118735855

Särndal, C.-E., & Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-4378-6

Wu, C. (2022). Statistical inference with non-probability survey samples. *Surv. Methodol.*, *48*(2), 283-311.

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. https://doi.org/10.1198/016214501750333054

Wu, C., & Thompson, M. E. (2020). *Sampling Theory and Practice*. New York: Springer. https://doi.org/10.1007/978-3-030-44246-0

Yang, S., & Kim, J.-K. (2020). Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.*, *3*, 625-650. https://doi.org/10.1007/s42081-020-00093-w

Yang, S., & Kim, J.-K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Surv. Methodol.*, *47*(1), 29-58.

Ybarra, L. M. R., & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*(4), 919-931. https://doi.org/10.1093/biomet/asn048