REDUCING SELECTION BIAS IN NON-PROBABILITY SAMPLE BY SMALL AREA ESTIMATION

Francesco Schirripa Spagnolo [1], Gaia Bertarelli[2], Nicola Salvati [1], Donato Summa [3], Monica Scannapieco [3], Stefano Marchetti [1] and Monica Pratesi [13]

[1] Department of Economics and Management, University of Pisa, (e-mail: francesco.schirripa@unipi.it, nicola.salvati2@unipi.it, stefano.marchetti@unipi.it, monica.pratesi@unipi.it)

[2] Department of Economics, Ca' Foscari University of Venice (e-mail: gaia.bertarelli@unive.it)

[3] ISTAT (e-mail: summa@istat.it, scannapieco@istat.it, pratesi@istat.it)

ABSTRACT: Nowadays, the availability of a huge amount of data produced by a wide range of new technologies is increasing. However, data obtainable from these sources are often the result of a non-probability sampling process. We propose a method to reduce the selection bias associated with the big data in the context of Small Area Estimation. Our approach is based on data integration and it combines a big data sample and a probability sample. Real data examples are considered in the context of Italian enterprises sensitiveness towards Sustainable Development Goals and e-commerce.

KEYWORDS: official statistics, big data, data integration, SDGs, e-commerce.

1 Introduction

For many decades probability surveys have been the standard for producing official statistics. However, the decline in response rates in probability surveys associated with the increasing cost of data collection have become big issues for producing official statistics. Due to technological innovations, over the past decade, there has been an unprecedented increase in the volume of "new" data, called *big data*, which are often the results of non-probability sampling processes but, at the same time, they offer very rich data sets. Anyway the "nature" itself of the data, as collected without a probability scheme, opens the door to possible selection bias, even at domain level.
Although, there is a trend to modernize official statistics through a more extensive use of big data, making reliable inferences from a non-probability sample alone is very challenging and a naive use of these data can lead to

biased estimates as affected by selection bias and measurement error. The Italian National Statistical Institute has a strategic program of investments on the use of these new data sources to complement and enrich official statistics. In this context a roadmap document, named "Roadmap for Trusted Smart Statistics"(RTSS), has been released. This work must be laid in the methodological action of the RTSS related to quality improvement by reducing non-representativeness of Big Data sources at survey unplanned domain level.

2 Notation

We consider a population U of size N divided into m non-overlapping subsets U_i of size N_i, $i = 1,\ldots,m$. Let y_{ij} denote the value of the target variable for the unit j belonging to the area i. A non-probability sample B is available for the target population, with $B \subset U$. We assume that the non-probability sample is available in each area of interest: B_i is the non-probability sample in the area i, $B_i \subset U_i$. We denote the inclusion indicator in B_i as δ_{ij}; in other words, $\delta_{ij} = 1$ if $j \in B_i$, $\delta_{ij} = 0$ otherwise; therefore $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$. The study variable y_{ij} is observed only when $\delta_{ij} = 1$. The non-probability sample contains other auxiliary variables, denoted by $\mathbf{x}$.

A survey data of size n, denoted by A, is also available; $A_i \in U_i$ drawn randomly. The survey data do not contain the variable of interest but contain only the auxiliary variables $\mathbf{x}$. The area-specific samples A_i are available in each area, but the number of sample units in each area, $n_i > 0$, is limited. Therefore, the areas of interest can be denoted as "small areas". In general, an area is regarded as "small" if the domain-specific sample size is not large enough to obtain direct estimates with acceptable statistical significance. In these cases, SAE techniques need to be employed.
In summary, the available data can be denoted by $\{(y_{ij}, x_{ij}), i \in B\}$ and $\{(x_{ij}), i \in A\}$, and the quantities of interest are the area means $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\ldots,m$. By using B we can estimate $\bar{Y}_i$ by:

$$\bar{Y}_{B_i} = N_{B_i}^{-1} \sum_{j \in B_i} y_{ij},$$

where $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$ and y_{ij} is the jth observation in the area i. Because of the selection bias and the measurement error, the sample mean $\bar{Y}_{B_i}$ from the non-probability sample is biased, and it does not represent the target population (Kim & Wang, 2019). Therefore, we propose a techniques in order to make valid inference from big data sources when the aim is to provide reliable estimates at small area level.

3 Reducing selection bias in big data: a data integration approach using SAE methods

We consider a data integration method for combining probability and non-probability samples in order to reduce the bias which is assisted by unit level small area model, following the approach of Kim and Wan (2019). We consider the case in which the survey data and the big data are available in each small area of interest. We also assume that the selection mechanism for the big data is non-informative :

$$P(\delta_{ij} = 1|\mathbf{x}_{ij}, y_{ij}; u_i) = P(\delta_{ij} = 1|\mathbf{x}_{ij}; u_i)$$

where u_i is an area-specific random effect characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$.

Moreover, we can observe δ_{ij}, the big data sample inclusion indicator, from the sample A. We can use the data $\{(\delta_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the for the propensity scores $P(\delta_{ij} = 1|\mathbf{x}_{ij}) = p(\mathbf{x}, \lambda)$ in sample B based on the missing at random. Usually, a logistic regression model for the binary variable δ_{ij} can be used in order to obtain estimators $\hat{p}_{ij}$ in sample B.

In order to take into account the hierarchical structure of the data, we consider the following generalized linear random intercept model for the propensity scores:

$$\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i) = g^{-1}(\mathbf{x}_{ij}^T\hat{\lambda} + \hat{u}_i),$$

where $g(\cdot)$ is a logit link function; $\hat{\lambda}$ and $\hat{u}_i$ are the ML estimates of λ and u_i.

To develop our estimator we suppose that the following working population model holds for sample B:

$$E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T\beta + \gamma_i\right), \tag{1}$$

where $h(\cdot)$ is the link function, assumed to be known and invertible, γ_i is the area-specific random effect for area i characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$. It should be noted that the covariates used here could be different from those used to fit the propensity model. Model in equation (1) includes three important special cases: the linear model obtained with $h(\cdot)$ equal to the identity function and y_{ij} is a continuous variable; logistic generalized linear random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized linear random intercept model where $h(\cdot)$ is the log link function

and the individual y_{ij} values are taken to be independent Poisson random variable. Using data from the big data sample B, assuming the model is correctly specified, we obtain an estimator of $\hat{\beta}$ which is consistent for β (Rao, 2021). Then a doubly robust (DR) estimator of the mean is given by:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j\in B_i}\frac{1}{\hat{p}_{ij}(\hat{\lambda},\hat{u}_i)}(y_{ij}-\hat{\mu}_{ij})+\frac{N_i}{n_i}\sum_{j\in A_i}\hat{\mu}_{ij}\right\}, \tag{2}$$

where $\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}\hat{\beta}+\hat{\gamma}_i\right)$ and $\hat{\beta}$ and $\hat{\gamma}_i$ are respectively the estimated regression coefficients and the random effects based on the big data sample.

The estimator in Eq. (2) is DR in the sense that it is consistent if both the model for propensity scores and the model for the study variable are correctly specified (Kim & Wang, 2019, Rao, 2021).

4 Real data examples

The proposed methodology has been applied to estimate the proportion of enterprises sensitive to Sustainable Development Goals (SDGs) of the 2030 Agenda at the provincial level in Italy. The Big Data sample is represented by the enterprises' websites accessed due to a web scraping procedure. The probabilistic sample dataset is a sub-sample of the survey "*Situazione e prospettive delle imprese nell'emergenza sanitaria Covid-19"* (2020). The target variable is a binary indicator computed for each enterprise and represents if the enterprise is sensitive or not to SDGs. This indicator has been computed through machine learning methods by analyzing the big data sample and looking for a set of pre-defined SDGs-related words on each website. Furthermore, an application related to the diffusion of e-commerce in Italian companies, using the same data of the application on sustainability, will be considered.

References

KIM, JAE KWANG, & WANG, ZHONGLEI. 2019. Sampling techniques for big data analysis. *International Statistical Review*, **87**, S177–S191.

RAO, JNK. 2021. On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, **83**, 242–272.