Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics)

Inferenza per big data assistita da metodi di stima per piccole aree: un applicazione sulle OBEC

Monica Pratesi, Francesco Schirripa Spagnolo, Gaia Bertarelli, Stefano Marchetti, Monica Scannapieco, Nicola Salvati, Donato Summa

Abstract Nowadays, the availability of a huge amount of data produced by a wide range of new technologies, so-called big data, is increasing. However, data obtainable from big data sources are often the result of a non-probability sampling process and adjusting for the selection bias is an important practical problem. In this paper, we propose a novel method of reducing the selection bias associated with the big data source in the context of Small Area Estimation (SAE). Our approach is based on data integration and the combination of a big data sample and a probability sample. An application on OBEC (on-line based enterprise characteristics) combining Istat sampling survey and web scraping data has been proposed.

Abstract *Attualmente, la disponibilità di grandi quantità di dati che vengono prodotti da nuove tecnologie, c.d. big data, è sempre più in crescita. Tuttavia, tali big data sono spesso il risultato di un processo di campionamento non probabilistico ed è necessario considerare il problema del bias di selezione. In questo lavoro, proponiamo un nuovo metodo per ridurre il bias di selezione associato ai big data nel contesto della stima per piccole aree. Il nostro approccio si basa sullla metodologia integrazione dii dati ed, in particolare, sulla integrazione di un campione di big data e un campione probabilistico. Viene proposta un'applicazione sulle OBEC (caratteristiche dell'impresa on-line based) che combina i dati di indagine campionaria Istat e web scraping.*

Monica Pratesi
Istat and Dipartimento di Economia e Management Università di Pisa e-mail: monica.pratesi@istat.it

Francesco Schirripa Spagnolo; Stefano Marchetti; Nicola Salvati
Dipartimento di Economia e Management Università di Pisa e-mail: francesco.schirripa@unipi.it; stefano.marchetti@unipi.it; nicola.salvati@unipi.it

Gaia Bertarelli
Istituto di Management Scuola Superiore Sant'Anna e-mail: gaia.bertarelli@santanna.it

Monica Scannapieco; Donato Summa
Istat, e-mail: scannapi@istat.it; donato.summa@istat.it

Key words: Data integration; Small Area Estimation; Big data; Official Statistics

1 Introduction

In recent years, there has been a growing demand for more and more detailed official data in order to implement more targeted policies. This has increased the need for appropriate statistical methods to produce reliable statistics for subdomains of a population (such as geographical areas or socio-economic groups). For many decades, probability surveys have been the standard for producing Official Statistics. Due to technological innovations, over the past decade, there has been an unprecedented increase in the volume of "new" data, such as transaction data, social media data, internet of things and scrape data from websites, sensor data and satellite images and so on. Generally, they are called *big data*. Furthermore, the decline in response rates in probability surveys associated with the the increasing cost of data collection have become senior issues for producing official statistics in developed countries.

Big data sources are often the results of non probability sampling processes but, at the same time, they offer very rich data sets: the data can be classified by geographical domains and/or also cross-classified by social and demographic domains (such as gender, educational level for individuals or economic activities for enterprises). 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, and non-probability samples in general, 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 [5].

So inference from big data sources/domain level data needs to be rethought and selection bias adjustments introduced.

In this context Small area estimation (SAE) methods can contribute as a useful tool to integrate data from probability and non-probability sources. Usually, small area techniques provide official statistics at domain of study level using probability surveys and other sources of available information from which the estimators can borrow strength.

In this work, we assume that we have access to a non-probability sample and a probability survey sample from the same finite population and that the target variable is observed only in the big data source. This situation, tend to be very common in practice and very interesting for future use of big data sources.

2 Effect of the selecion bias when the study variable is not observed in the probability sample

We consider a population U of size N divided into m non-overlapping subsets (domains of study or areas) U_i of size N_i, $i = 1,\dots,m$. Let y_{ij} denote the value of the target variable for the unit j belonging to the area i. We assume to have two samples referred to this population of interest: a non-probability sample and a probability sample. Moreover, we assume that the study variable is observed only in the non-probability sample.

A non-probability sample, denoted by 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 contains other auxiliary variables, denoted by $\mathbf{x}$.

A survey data of size n, denoted by A, is available; A_i is a subset of U_i drawn randomly. The survey data do not contain the variable of interest but contain only 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, a domain (or area) is regarded as “small” if the domain-specific sample size is not large enough to obtain direct estimates with acceptable statistical significance [7]. These areas can be geographic areas, such as provinces or municipalities and other sub-populations, such as the firms belonging to a industry subdivision. 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 $\{(x_{ij}), i \in A\}$.

The quantities of interest are the area means $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\dots,m$.

By using the non-probability sample we can estimate $\bar{Y}_i$ by:

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

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. Indeed, non-probability samples have unknown selection/inclusion mechanisms and are typically biased, and they do not represent the target population [3, 8]. Thus, a non-probability sampling design, makes the analysis results subject to selection bias.

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 sources: a data integration approach using Small Area Estimation methods

Data integration represents a quite new research area aimed at combining information from two independent surveys on the same target population [4].

Using multiple data sources is common in SAE; indeed, small area methods combine the data from a survey with predictions from a regression model using covariates from the administrative or census data. The SAE models are classified into two categories according to the available data on the target variable: (i) area level models and (ii) unit levels model. The *standard* SAE models use hierarchical model in which the deviation of an area mean from the overall mean is represented by a random effect.

If information at unit level is available, the standard unit-level small area model proposed by [1] may be used. In this case, the hierarchical model used for the individual response of the survey individual j in area i is:

$$y_{ij} = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \tag{1}$$

where the area-specific random effects u_i and individual level errors e_{ij} are assumed to be normally distributed with mean 0 and variance σ_u^2 and σ_e^2, respectively.

We suppose that the quantities of interest are the area means, it possible to express the mean in terms of linear combination between observed and unobserved units as follows

$$\theta_i = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i}(\mathbf{x}_{ij}^T\hat{\beta} + \hat{u}_i)\right], \tag{2}$$

where $\hat{\beta}$ and $\hat{u}_i$ are the BLUE for β and u_d and s_i is the set of the sampled units in area i and r_i is the set of the non-sampled units in area i.. Replacing the variance components by their estimators we obtain the Empirical Best Linear Unbiased Predictor (EBLUP).

Being assisted by unit level approach, we propose a new method to producing statistics at local level when the variable of interest has been recorded only in the non-probability sample. In particular, we consider a data integration method for combining probability and non-probability samples (i.e. big data sample) assisted by unit level small area model, following the approach of [3], in order to reduce the bias.

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 no-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. In other words, among the elements in sample A, it is possible to obtain the membership information from the big data sample B.

We can use the data $\{(\delta_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the for the participation probabilities or propensity scores ($P(\delta_{ij} = 1|\mathbf{x}_{ij} = p(\mathbf{x}, \lambda))$ in sample B based on the missing at random (MAR). 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 in to account the hierarchical structure of the data, we consider the following generalized liner 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.

In order to develop our estimator we suppose a working population model holds for sample B. We assume that the following working population model holds for sample B:

$$E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T \beta + \gamma_i\right), \tag{3}$$

where 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 $\mathbf{x}_{ij}$. Model in equation (3) 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 liner random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized liner random intercept model where $h(\cdot)$ is the log link function and the individual $y_{ij}i$ 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 [6].

Then a doubly robust (DR) estimator of the mean is given by:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i)}(y_{ij} - \hat{\mu}_{ij}) + \sum_{j \in A_i} \hat{\mu}_{ij}\right\}, \tag{4}$$

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. (4) 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 [3, 6].

4 Application Setting: Estimating Online-Based Enterprise Characteristics

Let us consider a setting in which the Big Data source is represented by the websites of enterprises that are accessed as a result of a web scraping procedure. Starting from a set of URLs (i.e. addresses identifying the enterprise websites), the procedure accesses URLs, extracts texts from the sites and stores such texts for subsequent analyses. In particular, text analyses can be performed to estimate the so-called Online-based Enterprise Characteristics (OBEC), i.e. some characteristics of businesses that are available on their own websites. In this specific setting, we assume to start from the Italian Statistical Business Register and being tU the universe of enterprises with equal or more than 10 employees, we select the subset S having a (valid) URL available. The Big Data sample B is accessed starting from S and will consist of all the texts of scraped websites. Notice that, assuming that URLs are all valid, the cardinality of S is equal to cardinality of B, i.e. B is the online representation of enterprises in S. By using B, we would like to compute a Yes/No indicator Y, considering if the enterprise is sensitive or not to Sustainable Development Goals of the 2030 Agenda. The indicator, named SDG enterprise sensitiveness, can be computed by analyzing B and looking for the presence of a set of pre-defined SDG related words on each website. B and S share a set of X variables that include Vat Code, Name of the Enterprise, Address, Municipality, Province, Zip Code, NACE code and Number of employees. In addition, X variables are also common to specific survey data A; in this application, we will use data of the “ ICT usage in enterprises” survey. Considering A and B, let us observe that we can consider a specific variable that denote enterprises present in A but not in B; the variable reports if an enterprise has a known website, i.e. a URL is available, or not. Figure 1 reports a visual representation of the application setting.

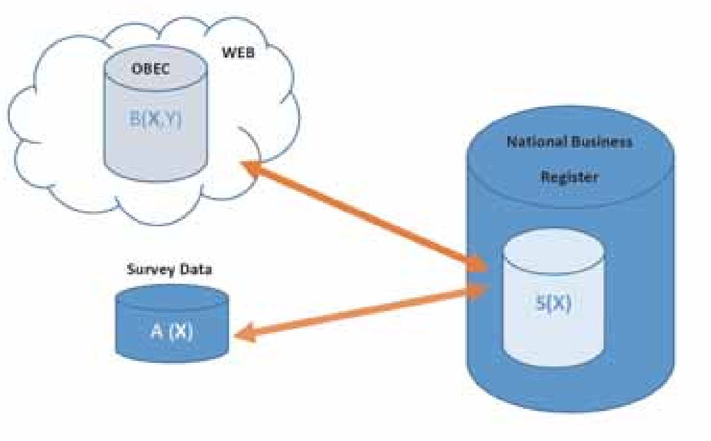

Fig. 1 Application Setting

In summary, as illustrated in this example, big data sources are a treasure of information that runs the risk of being underestimated as not connected with existing official data. They offer data affected by selection bias, as already stated in many scientific papers (see, among others, 2, 6) and adjusting for this selection bias in big data is an important and urgent problem. The effect of selection bias is likely to be even more serious at domain level when the domains are defined by socio-demographic groups. Age groups, gender, educational level, zone of residence, geography in general are often highly correlated with digital divide. This last is often the factor explaining self-selection bias and the presence/absence in big data sources of individuals, households and firms.

In this work we dealt with the problem of making reliable inference for small domains when the target variable is stored in a non-probability sample (big data sample) which is assumed to be available in each area and the number of units in each area is quite large. In particular, we propose a method based on the integration of a probability and a non-probability sample in order to reduce the selection bias associated with big data when the aim is to predict statistics at the local level.

References

[1] Battese, G.E., Harter, R.M., Fuller, W.A.: An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association **83**, 28–36 (1988)

[2] Beaumont, J.-F.: Are probability surveys bound to disappear for the production of official statistics?. Survey Methodology **46**, 1–28 (2020)

[3] Kim, J.K., Wang, Z.: Sampling techniques for big data analysis. International Statistical Review **87**, S177–S191 (2019)

[4] Lohr, S.L., Raghunathan, T.E: Combining survey data with other data sources. Statistical Science **32**, 293–312 (2017)

[5] Meng, X.-L.: Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics **12**, 685–726 (2018)

[6] Rao, J.N.K.: On making valid inferences by integrating data from surveys and other sources. Sankhya B **83**, 242–272 (2021)

[7] Rao, J.N.K., Molina, I.: Small area estimation. John Wiley & Sons, New York (2015)

[8] Yang, S., Kim, J.K.: Statistical data integration in survey sampling: A review. Japanese Journal of Statistics and Data Science (2020) doi: 10.1007/s42081-020-00093-w

Statistical methods and models for Sports Analytics