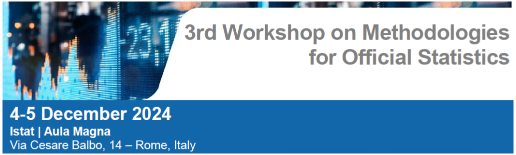
Controlling selection bias in non-probability sample using small area estimation: an application to official statistics

Francesco Schirripa Spagnolo

Department of Economics and Management, University of Pisa
Centro ASESD Camilo Dagum
francesco.schirripa@unipi.it



Joint work with . . . and acknowledgement

- Gaia Bertarelli (Ca' Foscari University of Venice; Centro ASESD Camilo Dagum)
- Nicola Salvati (University of Pisa; Centro ASESD Camilo Dagum)
- Stefano Marchetti (University of Pisa; Centro ASESD Camilo Dagum)
- Donato Summa (ISTAT)
- Monica Scannapieco (Italian National Authority for Cybersecurity; formerly at ISTAT)
- Monica Pratesi (ISTAT; University of Pisa; Centro ASESD Camilo Dagum)

Funding

This work has been supported by

- Ministry of University and Research (MUR) as part of the FSE REACT-EU—PON 2014-2020 'Research and Innovation' resources—Innovation Action—DM MUR 1062/2021—Title of the Research: 'Statistical Machine Learning nelle Indagini Campionarie';
- Project 'Quantification in the Context of Dataset Shift' (QuaDaSh) (Bando 2022 PNRR Prot. P2022TB5JF)
- Project 'Future Artificial Intelligence Research' (FAIR— PE00000013)
- MAPPE project, Programma 'PE GRINS—GRINS—GROWING RESILIENT, INCLUSIVE AND SUSTAINABLE (cod. PE0000018 CUP: J33C22002910001).
- Trusted Smart Statistics—Web Intelligence Network' (2020-PL-SmartStat— 1010358)

The presentation at a glance

- **Methodological key point**: Bias Correction of the estimates from a non-probability sample at survey unplanned domain level.
- **Idea**: We extend the work of Kim and Wang (2019) and Kim et al. (2021) at Small Area level.
  - Target variable comes only from Big Data sources (in this case the number of observations can be large or not).
  - The small areas are domains considered in a probability survey.
  - Proposal: a double robust (DR) estimator that combines
    1. propensity weighting to improve the representativeness of the non-probability sample obtaining inverse probability weighted estimators (Chen and Wu, 2020),
    2. statistical model to predict the units which are not in the big data sample (Valliant et al., 2000)
- **Application**: Estimating the proportion of Italian Enterprises sensitive of SDGs at provincial (NUTS3) level.

Overview

Framework

Data

Methodology

Simulation Scenarios

Application

Pros, Cons and Future Works

Framework

Framework

- At the end of 2022, the European Parliament adopted the Corporate Sustainability Reporting Directive (CSRD) which obliges companies to publish data on the impact of their activities on environment, people and planet (Dinh et al., 2023).
- Primary importance for National Statistical Institutes (NSIs) is to estimate the proportion of Italian Enterprises sensitive of SDGs (**SDG Enterprise Sensitiveness** (SDGES)).
- Information needed at granular level in order to target policy and fundings $\rightarrow$ provincial (NUTS3) level.
- SDGES is not directly measurable by traditional surveys implemented by ISTAT $\rightarrow$ ***Big Data* obtained throw Online Web Scraping**
- Estimates coming from a non-probability sample could be bias.
- To correct the selection bias it is possible to use a probabilistic sample but this introduces the problem of small sample for the desired level of aggregation.

Data

Data

- Target population U is represented by all the Italian enterprises with ≥ 10 employees in one of the following of economic activities (2-digits NACE): (i) Manufacturing, (ii) Industry, (iii) Wholesale and retail trade, (iv) Other services activities.
- *Non-probabilistic* sample of Italian enterprises obtained by a web scraping procedure (B)
  - URLs retrieval from ASIA (Italian Statistical Business Register) register $B \subset U$ (not all enterprise in U have a website);
  - Retrieving the text of the websites from B (SDGs words related);
  - Identify if an enterprise is sensitive to sustainability goals (value 1) or not (value 0) (our **target variable SDGES**) by ML a binary classifier;

Data

- We obtained an organized dataset with 10 variables:
    - number of employees of the enterprise averaged over the years
    - turnover volume indicator in classes (14 classes)
    - NACE code (4 classes)
    - VAT Code
    - name of the enterprise
    - address
    - municipality
    - province
    - Zip Code
    - Target variable
- B sample size $n_B = 51753$

Data

- *A* is a probabilistic sample from the *Istat Special survey on Enterprises perspectives after Covid-19 emergency* (Costa et al., 2022):
  - Sub-sample of the survey that selects enterprises with 10 or more employees in the four considered NACE sectors;
  - Survey sample size $n_A = 19606$;
  - NUTS3 by our target population are considered as small areas;
  - Sample sizes in the provinces ranges from 24 to 1220 (less than 100 enterprises in 35% of the areas);
  - In *A* we have an indicator variable that denotes if a URL is available (value 1) or not (value 0).

Data

- *A* and *B* share variables obtained through a direct (exact) linkage through ASIA:
    1. number of employees (average over the year - continuous);
    2. turnover volume indicator (14 classes);
    3. NACE code;
    4. general and geographical details.
- SDGES is available in *B* and not in *A*.

Methodology

Notation

A finite target population U can be partitioned into m non-overlapping areas U_i of size N_i, $i = 1, \ldots, m$ $i = 1, \ldots, m$

- **Non-probability sample**
  - non-probability sample B of size n_B is available with $B \subset U$,
  - B_i is the non-probability sample in the area i, $B_i \subset U_i$, $i = 1, \ldots, m$, n_{B_i} sample size in area i,
  - Indicator of inclusion: $\delta_{ij} = 1$ if $j \in B_i$, $\delta_{ij} = 0$ otherwise; therefore $n_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$
  - Contains the variable of interest and a series of covariates: $(\mathbf{x}_{ij}, y_{ij})$
- **Probability sample**
  - A survey data A of size n_A is available, A_i is a subset of U_i drawn randomly such that the inclusion probability of the unit j within area i is π_{ij} ($w_{ij} = 1/\pi_{ij}$).
  - Sample size in each area A_i could be small.
  - The survey data do not contain the variable of interest but contain only the covariates $\mathbf{x}_{ij}$ and δ_{ij}

Selection bias

- Target parameter: area means $\theta_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1, \dots, m$
- $y_{ij} = 1$ if SGDES is *YES* for enterprise j in area i, 0 otherwise.
- By using the non-probability sample B we can estimate θ_i by *naive direct estimator*:

$$\tilde{\theta}_{B_i} = n_{B_i}^{-1} \sum_{j \in U_i} \delta_{ij} y_{ij}$$

  y_{ij} is the jth observation in area i
- Although the nonprobability data can have a large sample size, because of the unknown sample selection/inclusion mechanism, they do not represent the target population (Yang and Kim, 2020) $\rightarrow \tilde{\theta}_{B_i}$ is biased.

The proposed approach

Data integration can be used to reduce the bias by combining a probability and a non-probability sample through a vector of common auxiliary variables (Kim and Wang, 2019).

Assumptions:

- We can observe δ_{ij}, the big data sample inclusion indicator, from the sample A.
- The selection mechanism of the big data sample is ignorable:

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

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

Propensity score

- We assume the following model for the propensity scores based on the missing at random (MAR):

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

  where λ is the vector of the regression coefficients.
- The hierarchical structure of the data should be considered in the estimation model of the propensity scores.
- We can use the data $\{(\delta_{ij}, w_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the propensity scores in B:

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

  where $g(\cdot)$ is a (logit) link function; $\hat{\boldsymbol{\lambda}}$ and $\hat{u}_i$ are the ML estimates of $\boldsymbol{\lambda}$ and u_i.
  Even if the area-specific sample size is small, we borrow strength from the whole sample using the above model to obtain stable values of $\hat{p}_{ij}$s.

DR estimator mixed model approach

- We assume that the following working population model holds for sample B:

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

  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 $\mathbf{x}_{ij}$.

- We can use data $\{(y_{ij}, x_{ij})\} \in B$ to fit the working model:

$$\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T\hat{\beta} + \hat{\gamma}_i\right)$$

  where $\hat{\beta}$ and $\hat{\gamma}_i$ are the ML estimates of β and γ_i.

DR estimator

- The DR estimator based on the mixed model approach is given by:

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

where
  - $\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}\hat{\beta} + \hat{\gamma}_i\right)$;
  - $\hat{\beta}$ and $\hat{\gamma}_i$ are the estimated regression coefficients and the random effects based on the B;
  - w_{ij} is the sampling weight of the unit j in area i.
- DR: consistent if the model for propensity scores or the model for the study variable are correctly specified (Kim and Wang, 2019; Rao, 2021).
- Bootstrap procedure to approximate the variance of the estimator.

Bootstrap variance estimation

1. Extract a sample with replacement of size n_A from A using a sampling design with inclusion probabilities π_{ij} to obtain a bootstrap replicate $\{(\delta_{ij}^\star, w_{ij}^\star, \mathbf{x}_{ij}^\star)\} \in A^\star$.
2. Extract a srswr of size n_B from B to obtain a bootstrap replicate $\{(y_{ij}^\star, \mathbf{x}_{ij}^\star)\} \in B^\star$
3. Obtain the bootstrap propensity score $\hat{p}_{ij}^\star(\mathbf{x}_{ij}, \hat{\boldsymbol{\lambda}}^\star, \hat{u}_i^\star)$ by using scaled bootstrap weights, $\tilde{w}_{ij}^\star = w_{ij}^\star N_i / \sum_{j\in i} w_{ij}^\star$.
4. Fit the model on the bootstrap sample $B^\star$ to estimate the regression coefficients $\hat{\beta}^\star$ and area-specific random effects $\hat{\gamma}_i^\star$.
5. Obtain the DR estimator $\hat{\theta}_{i;DR}^\star$
6. Repeat steps 1–5 independently for L times. The resulting bootstrap variance estimator of $\hat{\theta}_{i;DR}$ is computed as follows (Kim et al., 2021):

$$\hat{V}(\hat{\theta}_{i;DR}) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_{i;DR}^{\star(l)} - \hat{\theta}_{i;DR}\right)^2$$

Simulation Scenarios

Simulation Scenarios

Limited simulations were performed to

1. compare the SAE DR estimator based on the mixed model approach with the naive direct estimator (from a nonprobability sample)
2. check the validity of the proposed variance for the SAE DR estimator.

The setup for the simulation is based on Chambers et al. (2016); Kim and Wang (2019).

Simulation setting

i) Linear model:

$$y_{ij}|u_i \sim \text{Bernoulli}(\pi_{ij}),\ i = 1, \ldots, m;\ j = 1, \ldots, N_i$$
$$\pi_{ij} = \exp(\eta_{ij})\{1 + \exp(\eta_{ij})\}^{-1}$$
$$\eta_{ij} = x_{1,ij} + x_{2,ij} + u_i$$

ii) Nonlinear model:

$$y_{ij}|u_i \sim \text{Bernoulli}(\pi_{ij}),\ i = 1, \ldots, m;\ j = 1, \ldots, N_i$$
$$\pi_{ij} = \exp(\eta_{ij})\{1 + \exp(\eta_{ij})\}^{-1}$$
$$\eta_{ij} = 0.5(x_{1,ij} - 1.5)^2 + x_{2,ij} + u_i$$

Simulation setting

- $x_{1,ij} \sim N(1, 0.5)$ and $x_{2,ij} \sim \mathrm{Unif}(a_i, b_i)$, for $a_i = -1$ and $b_i = m/16$, $i = 1, \ldots, m$; $j = 1, \ldots, N_i$.
- $u_i \sim N(0, \sigma_u^2 = 0.25)$.
- $m = 100$ small areas, $N_i = 1,000$.
- SSRWR within each area to obtain an independent sample A of size $n = 1,000$ with $n_i = 10$.
- $\delta_{ij} \sim \mathrm{Ber}(p_{ij})$ independently for $j = 1 \ldots N$ and $i = 1 \ldots m$,

Simulation setting

Two propensity score models:

i) Linear propensity model:

$$p_{ij} = \frac{\exp(x_{2,ij} + \gamma_i)}{1 + \exp(x_{2,ij} + \gamma_i)} \tag{1}$$

ii) Nonlinear propensity score model:

$$p_{ij} = \frac{\exp(-0.5 + 0.5 \cdot (x_{2,ij} - 2)^2 + \gamma_i)}{1 + \exp(-0.5 + 0.5 \cdot (x_{2,ij} - 2)^2 + \gamma_i)} \tag{2}$$

$\gamma_i \sim N(0, 0.1)$

Simulation scenarios

Four scenarios obtained by combining the outcome and propensity score models

1) Both the outcome regression model and the big data propensity score model are linear.
2) The outcome regression model is linear, and the big data propensity score model is nonlinear.
3) The outcome regression model is nonlinear, whereas the big data propensity score model is linear.
4) Both the outcome regression model and the big data propensity score model are nonlinear.

DR estimator

- The parameter of interest was the population proportion in each small area, θ_i.
- To obtain the SAE DR estimator, $\hat{\theta}_{i;DR}$, we used a random-intercept logistic model as the working propensity score model:

$$\text{logit}(p_{ij}(\mathbf{x}, \boldsymbol{\lambda}, u_i)) = \lambda_0 + \lambda_1 x_{2,ij} + u_i,$$

and we used the following random-intercept logistic model for the outcome:

$$\text{logit}(y_{ij}) = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \gamma_i.$$

Indicator of performance

For each scenario, we conducted $R = 500$ MC simulations.

To summarize the results, we used the following performance indicators:

- $\text{RB}(\tau_i) = R^{-1} \sum_{r=1}^{R} \frac{\left(\tau_i^{(r)} - \theta_i^{(r)}\right)}{\theta_i^{(r)}} \times 100$
- $\text{MSE}(\tau_i) = R^{-1} \sum_{r=1}^{R} \left(\tau_i^{(r)} - \theta_i^{(r)}\right)^2$

where τ_i is an estimator in area i (the compared estimators are SAE DR ($\hat{\theta}_{i,DR}$) and naive direct ($\tilde{\theta}_{B_i}$)), τ_i^r is its estimate obtained in the r-th MC replication, and θ_i is the population mean (the *true* value).

Simulation Results: Median over the areas over 500 MC simulations

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Naive direct	SAE DR	Naive direct	SAE DR	Naive direct	SAE DR	Naive direct	SAE DR
RB (%)	7.632	0.022	-4.746	-0.057	4.257	-0.016	-2.643	0.037
MSE	0.003	0.004	0.001	0.004	0.001	0.003	0.001	0.002

Median of RB of bootstrap and CR over the 500 simulations

Scenario	*RB (%)*	*CR (%)*
Scenario 1	0.218	91.1
Scenario 2	-3.067	90.9
Scenario 3	2.320	91.8
Scenario 2	0.902	90.6

IC:$[\hat{\theta}_{i;DR} - z_{\alpha/2}\widehat{SE}(\hat{\theta}_{i;DR}), \hat{\theta}_{i;DR} + z_{\alpha/2}\widehat{SE}(\hat{\theta}_{i;DR})]$.

Application

SDGES Results

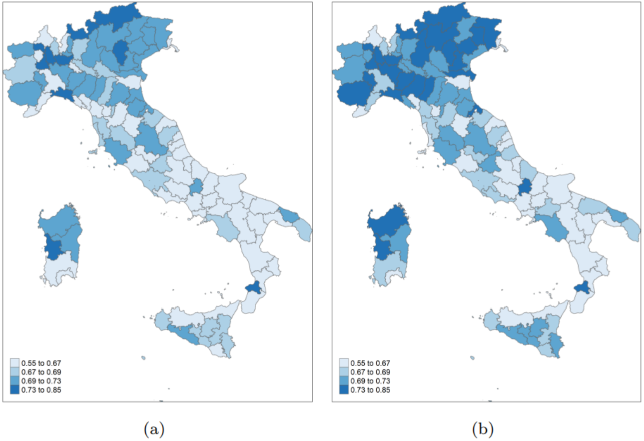


Figure 1: SDGES for the Italian provinces using the DR estimator (a) and the naive direct estimator (b)

SDGES Results

- North–south dualism, with greater attention to sustainability in the north.
- Bolzano (84.1%), Vercelli (77.8%), and Vibo Valentia (75.2%)
- Massa Carrara (59.0%), Crotone (59.9%), and Campobasso (60.8%)
- SAE DR estimator seems to smoothen the estimates more, as expected, according to the use of a model to correct bias.
- Similar geographical distribution, but the bias of the naive direct estimator could mislead policymakers.
- SAE DR estimates for 106 out of 107 areas had coefficients of variation (CV) below 16.6%

Pros, Cons and Future Works

Pros & Cons

- **Pros** of the proposed approach:
  - Represents one of the first attempt to obtain reliable estimates from a non-probability sample at Small Area Level.
  - Results highlight that the proposals tend to reduce the selection bias of the big data sample.
- **Cons** of the proposed approach:
  - A probabilistic survey is still needed.
  - The indicator of inclusion is not always available: reduction in the number of possible applications.
  - Only approximate bootstrap variance can be estimated (at the moment).
  - (not strictly connected to the proposed bias correction) Heavy influenced by the Machine Learning model (words selection, focus on the definition of the target variable).

References

Chambers, R., N. Salvati, N. Tzavidis, N. (2016) Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK *Journal of the Royal Statistical Society: Series A 179*(2), pp. 453–479

Chen, Y., Li, P. and C. Wu (2020) Doubly robust inference with nonprobability survey samples *Journal of the American Statistical Association 115*(532), pp. 2011–2021

Costa, S., S. De Santis, and R. Monducci (2022) Reacting to the covid-19 crisis: state, strategies and perspectives of italian firms. *Rivista di Statistica Ufficiale/Review of official statistics 1*, pp. 73–107

Dinh, T., A. Husmann, and G. Melloni (2016) Corporate sustainability reporting in europe: A scoping review. *Accounting in Europe 20*(1), pp. 1–29.

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021) Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), pp. 941–963.

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

Rao, J. (2021) On making valid inferences by integrating data from surveys and other sources *Sankhya B 83*(1), pp. 242–272.

Valliant, R., A. H. Dorfman, and R. M. Royall (2000) *Finite population sampling and inference: a prediction approach* John Wiley .

Yang, S. and J. K. Kim (2020) Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, pp. 1–26.

Francesco Schirripa Spagnolo

Dipartimento Economia e Management
Università di Pisa

francesco.schirripa@unipi.it