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

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

Abstract

The growing demand for high-frequency, granular, and timely statistical information is prompting National Statistical Institutes (NSIs) to incorporate new data sources, such as web scraping and big data, alongside traditional surveys. This study tackles the challenge of assessing the sensitivity of Italian enterprises to the United Nations Sustainable Development Goals (SDGs) at a detailed provincial level. The Italian National Institute of Statistics (ISTAT) currently lacks direct survey data on enterprises' SDG-related behaviour, motivating the use of web-scraped data from enterprise websites as a non-probability data source. However, the inherent selection bias in non-probability samples poses a significant challenge to reliable inference. To address this, we propose a data integration approach by employing small area estimation (SAE) techniques. Our method integrates non-probability big data with a probability sample to correct for selection bias and improve the precision of the estimates. Our proposed doubly robust (DR) estimator integrates propensity score weighting, to adjust for the representativeness of the non-probability sample, with model-based predictions for units not captured by the big data sample. Additionally, a bootstrap procedure for variance estimation enhances the robustness of our inference framework.

We validated our approach using Monte Carlo simulations to assess the estimator's robustness across various scenarios, including cases of model misspecification. The simulations confirm that the DR estimator effectively mitigates bias and yields reliable estimates even under incorrect specifications of the outcome regression or propensity score models. Applying our methodology to real web-scraped data reveals significant spatial heterogeneity in SDG-sensitive enterprises across Italian provinces, with higher sensitivity observed in the north compared to the south, and some provinces showing extreme values. These findings highlight the critical need for detailed data to support targeted policy actions. This work demonstrates the potential of integrating big data with SAE methods to derive insights into corporate sustainability practices, facilitating informed decision-making and policy formulation. By addressing selection bias in non-probability samples and overcoming the limitations of small sample sizes at detailed geographical levels, our approach offers a scalable solution for enhancing the quality of official statistics in the big data era, applicable beyond the Italian context to other domains requiring small area inference from combined data sources.

Keywords: Survey estimator, big data, web scraping.

[1] Francesco Schirripa (francesco.schirripa@unipi.it), Nicola Salvati (nicola.salvati@unipi.it), Università di Pisa; Gaia Bertarelli(gaia.bertarelli@unive.it), Università di Venezia; Donato Summa (donato.summa@istat.it), Monica Pratesi (monica.pratesi@istat.it), Stefania Marchesi (stmarche@istat.it), Italian National Institute of Statistics –Istat; Monica Scannapieco (scannapi@istat.it), formerly at Italian National Institute of Statistics – Istat.