Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 27–45
https://doi.org/10.1093/jrsssa/qnae115
Advance access publication 18 November 2024
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Inference for big data assisted by small area methods: an application on sustainable development goals sensitivity of enterprises in Italy

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

[1]Department of Economics and Management, University of Pisa, Via Cosimo Ridolfi 10, Pisa 56124, Italy
[2]Department of Economics, Ca' Foscari University of Venice, Campus S. Giobbe, Cannaregio 873, Venice 30121, Italy
[3]Department for Development of Methods and Technologies for Production and Dissemination of Statistical Information, Italian National Institute of Statistics (ISTAT), Via Cesare Balbo 16, Rome 00184, Italy

Address for correspondence: Gaia Bertarelli, Department of Economics, Ca' Foscari University of Venice, Campus S. Giobbe, Cannaregio 873, Venice 30121, Italy. Email: gaia.bertarelli@unive.it

Abstract

In this study, we proposed a new method for estimating the sensitivity of enterprises in Italy to the United Nation's sustainable development goals at the provincial level using web-scraping data (a nonprobability sample) because this value is not surveyed by the Italian National Institute of Statistics. The proposed method used a probability sample to reduce the selection bias of estimates obtained from the nonprobability sample in the context of small area estimation and integrated nonprobability and probability samples using a double robust estimator that combined (i) propensity weighting to improve the representativeness of the nonprobability sample and (ii) a statistical model to predict the units that were not in the nonprobability sample. A bootstrap procedure for estimating variance was also proposed. To validate the proposed method, a Monte Carlo simulation was performed. Results showed that the proposed method allowed the correction of bias from the nonprobability sample while maintaining a good level of estimate reliability.

Keywords: data integration, nonprobability sample, selection bias, sustainability, unplanned domains

1 Introduction

Several recent events, such as the COVID-19 pandemic, have forced new thinking directions and opened up new opportunities to advance knowledge and provide more useful data to enhance the quality of human life, now and in the future. In particular, information has to be available at very high frequencies (daily, weekly, and monthly) and with sufficient granularity to respond quickly and effectively to the needs of policymakers and society at large. Consequently, National Statistical Institutes (NSIs) have been challenged with providing timely, accessible, and detailed statistical data (UNECE, 2022). To respond to these challenges, NSIs and academia are exploring new data sources, such as big data, as alternatives to traditional data collection approaches. In addition, NSIs and academia are studying innovative statistical approaches to define and maintain high-data quality (Beaumont, 2020; Citro, 2014).

The Italian National Statistical Institute (ISTAT) has been investigating the potential of big data sources for official statistics since 2013. ISTAT followed the strategic indications reported in two

Received: January 15, 2024. **Revised:** October 18, 2024. **Accepted:** October 23, 2024

reference documents besides the European Statistical System: (i) the Scheveningen memorandum (Eurostat, 2013), where European NSIs were called to investigate the possible use of big data sources to support the production of official statistics, and (ii) the Bucharest memorandum (Eurostat, 2018), which indicated investments necessary to produce big data-based statistics as part of official statistics for all effects and purposes.

Moreover, in the last 5 years, ISTAT has developed a strategic program of investments in the use of new data sources to complement and enrich official statistics named *Roadmap for Trusted Smart Statistics*. In this framework, the possibility of using web-scraping techniques associated with the estimation phase of text and data mining algorithms was explored to replace traditional data collection and estimation equipment for enterprise characteristics or to combine them in an integrated approach (Barcaroli & Scannapieco, 2019). In addition, other NSIs around the world are moving in the same direction.

Contextually, ISTAT has started an intense activity for the definition and progressive implementation of a statistical reference framework that can provide initial empirical evidence on the characteristics of sustainable behaviours of companies and produce some initial indicators on diffusion and orientation toward sustainability in Italian productive fabrics (Istat, 2020a, 2020b).

At the end of 2022, the European Parliament adopted the corporate sustainability reporting directive, which, starting in 2024, will oblige many companies to publish data on the impact of their activities on the environment, people, and the planet (Dinh et al., 2023). Consequently, it has become essential for NSIs to estimate the commitment of national companies to sustainability, at least at the provincial (NUTS 3) level. This approach is of paramount importance because it allows for an accurate and detailed assessment of the impact of businesses on local communities and the surrounding environment. As highlighted in a study conducted by the Policy Department for Economic, Scientific, and Quality of Life Policies (Policy Department for Economic, 2020), measuring sustainability at the provincial level provides a more comprehensive picture of sustainability challenges and opportunities, enabling local authorities to adopt targeted policies and businesses to adjust their corporate practices more effectively. Thus, provincial-level analysis provides a clearer picture of areas where efforts need to be concentrated to enhance sustainability and promote well-being at the local level, thereby contributing to the achievement of global sustainable development goals (SDGs).

Therefore, one of the main goals of ISTAT is to estimate the proportion of enterprises sensitive to the SDGs of the 2030 Agenda of the United Nations (UN) at the provincial level in Italy. Unfortunately, from the business surveys currently conducted by ISTAT, the sustainable behaviours of firms are not yet directly measurable.

Therefore, a solution is to use data from the websites of enterprises elaborated by machine learning (ML) and text mining techniques. This is a case of the usage of online-based enterprise characteristics (OBEC). However, the biggest problem with this solution is that data from websites, as coming from a nonprobability sample, can be affected by possible selection bias. Consequently, making reliable inferences from these data alone is very challenging, and their naive usage could result in severely biased estimates because of selection bias and measurement error. Indeed, the *bigness* of data is no substitute for representativeness. Moreover, as Meng (2018) highlighted, without adjustments to reduce bias associated with even bigger data, we can run into the so-called big data paradox: *the bigger the data, the surer we fool ourselves* (Meng, 2018). Thus, an increasing number of NSIs are experimenting with the use of new data sources to produce the same or new statistical information in a multisource environment, where traditional surveys are integrated with big data sources. In this study, we use website data as big data and a nonprobability survey sample, although the big data generation process can be much more complex than the survey data process.

To consider selection bias, we assume that some auxiliary variables are common between big data and a probability sample. An imputation model can then be trained (fit) on big data and used to predict the target variable on the probability sample to obtain target parameter (e.g. means and totals) estimates. This approach implies that the reliability of estimates is also related to the probability sample quality.

As aforementioned, estimates about sustainability must be obtained at least at the provincial level in Italy (NUTS 3), and such a level of aggregation is not planned in the design of business surveys usually conducted by ISTAT. Therefore, another issue arises concerning the use of probability samples to correct the selection bias of big data. The sample size at the provincial level is

usually small, and its use at this level of aggregation can invalidate the self-selection correction method. Usually, the problem of small sample sizes at the desired level of aggregation is known as a small area problem; that is, the domain/area sample size is not sufficiently large to obtain reliable direct estimates. There is considerable research on the small area estimation (SAE) problem. For a review of various SAE methods, see Rao and Molina (2015) and Pratesi (2016).

In summary, we face the following two issues: (i) we must resort to big data for our target variables, but big data are affected by self-selection bias (and other nonsampling problems); therefore, we must use a probability sample to correct the selection bias, thereby introducing (ii) the small sample problem for the desired level of aggregation (province level [NUTS 3]). Thus, in this study, we develop a bias correction model to obtain reliable estimates at the survey unplanned domain level, which is to be applied when values of the target variable are available only in big data and comparable auxiliary information is available in a probability sample.

We based our work on the study of Kim and Wang (2019), which proposed a double robust (DR) estimator that combined (i) propensity weighting to improve nonprobability sample representativeness obtaining inverse probability weighted estimators (Chen et al., 2020) with (ii) a statistical model to predict units that are not in the big data (Valliant et al., 2000). Other approaches to account for selection bias apply to big data standard weighting and calibration methods known from classical probability sampling (Baker et al., 2013; Marella, 2023). In addition, multilevel regression and poststratification is a popular method for addressing selection bias in subgroup estimation, and Si (2023) proposed a new framework for data integration within this approach.

For the use of big data in SAE, we list some interesting papers herein. For instance, Porter et al. (2014) used Google Trends searches as covariates in a spatial Fay–Herriot (FH) model (Fay & Herriot, 1979) to analyse the percentage of Spanish-speaking households in the eastern half of the U.S. Meanwhile, Schmid et al. (2017) used mobile phone data as covariates in a classical area-level FH model to estimate the literacy rate in Senegal by gender at the local (*commune*) level. Marchetti et al. (2015) used big data sources to study the estimation of poverty indicators for local areas in Tuscany (Italian region). Similarly, Marchetti et al. (2016) used text mined from Twitter to estimate the share of food expenditure at the local level in Italy. These papers demonstrate the predictive power of big data sources in improving SAE efficiency. However, these techniques cannot be applied in our case, where the target variable is unavailable in a probability sample survey.

The remainder of this article is structured as follows. Section 2 describes the business survey data used in the application and the ML approach adopted to obtain big data on the sustainable behaviours of firms in Italy. In Section 3, we introduce notations, recall the effect of selection bias when the target variable is unobserved in a probability sample, and describe our proposed method for obtaining reliable estimates at the provincial level when the target variable is available only from big data that share some common variables with a probability sample. In addition, a bootstrap technique for variance estimation is proposed. In Section 4, we discuss the application of sustainability, which concerns the prevalence of enterprises sensitive to the SDGs of the UN 2030 Agenda at the provincial level (NUTS 3) in Italy. Section 5 is devoted to validating the proposed method using Monte Carlo (MC) simulations. Finally, the conclusions and future work are presented in Section 6. Moreover, we provide in Section S1 of the online supplementary material with an application to real data to validate the proposed method, where the target is available both in big data and in a probability sample, and additional MC simulations.

2 Data

As aforementioned, data on sustainable behaviours are not currently collected by business surveys conducted by ISTAT. Consequently, they must be extracted from the websites of these enterprises, which entails a self-selection process as not all enterprises decide to maintain a website and publicly express their commitment to SDGs. To address this self-selection bias, a probabilistic survey sample was employed for correction.

ISTAT conducted two probabilistic sampling on enterprises. The first was a sampling survey on *Information and Communication Technologies in enterprises*, which aimed to produce information on the use of the Internet and other networks for various purposes (e.g. e-commerce, e-skills, e-business, social media, and e-government). The second, which is used in our application, was the *ISTAT Special Survey on Enterprises' Perspectives After the COVID-19 Emergency* (started in

November 2020), which provided information about the effects of the COVID-19 pandemic on firms' performances and strategies (e.g. demand dynamics, turnover, employment, investments, and technologies) and the type of reaction, if any, opposed to the shock (e.g. reorganization, downsizing, and digital transformation) (Costa et al., 2022).

Our target population comprises all enterprises with 10 or more employees that belong to four economic activities of interest for ISTAT: (i) manufacturing, (ii) industry, (iii) wholesale and retail trade, and (iv) other services. Economic activities are classified according to NACE (from the French *Nomenclature statistique des activités économiques dans la Communauté européenne*).

2.1 ML procedure for obtaining OBEC

We gathered data from the websites of enterprises to assess what we referred to as OBEC, that is, specific attributes of businesses that could be found on their websites and were pertinent to official statistics. Websites offer regularly updated information that can prove valuable in generating more cost-effective and up-to-date official statistics (UNECE, 2022). In addition, the Internet is a promising new data source. Indeed, the experiences of several NSIs and the recent Eurostat-sponsored Web Intelligence Network (https://cros-legacy.ec.europa.eu/WIN_en) have shown the appropriateness of Internet data in producing new statistics and augmenting existing statistics.

Starting from a set of URLs (i.e. addresses identifying the enterprises' websites), we extracted text from websites and stored them for subsequent analyses. In collecting URLs, we started with the European statistical business registers (ASIA–Statistical Register of Active Enterprises, ISTAT, 2021b), which was established in 1996. It covers all enterprises that perform economic activities contributing to the gross domestic product at market prices in the fields of industry, commerce, and services. It provides identification (name and address) and stratification (main economic activity, size, legal form, dates of creation and cessation, and turnover) variables. In terms of economic activities, although the classifications of ASIA and NACE could not always be the same, they were consistent for the economic activities used in the application. The register is updated annually by integrating administrative and statistical sources. Its regular maintenance ensures the updating of active units, providing an official data source, harmonized at the European level, for the statistical analysis of the business population and its demography. It plays a central role in the field of business statistics. It is used for estimating national economic accounts, supplying sample frames and population data necessary for conducting ISTAT surveys on enterprises, and identifying target populations for the preparation and coordination of surveys, as well as grossing up the survey results.

ASIA provided a perfect framework for our target population (enterprises with 10 or more employees that belong to four economic activities, as specified above), which we denoted by U, and it consisted of 189,074 enterprises. The first step in building a structured dataset from big data on the Internet was to retrieve URLs for all the enterprises in U, if any. To perform this task, we used a group of specific open-source scripts developed ad hoc by ISTAT (Barcaroli et al., 2015) (more details are available upon request from the author).

From U, we extracted information on different characteristics of enterprises with valid URLs, obtaining a subset of U that we denoted by K, which consisted of 51,754 enterprises. The information was about the following variables: (i) the number of employees of the enterprise averaged over the years, (ii) turnover volume indicator in classes (14 classes), (iii) NACE code (four classes), (iv) VAT code, (v) name of the enterprise, (vi) address, (vii) municipality, (viii) province, and (ix) zip code. The probability sample explained below also possessed these variables.

To obtain the target variable *SDG enterprise sensitivity*, which indicated whether the enterprise was sensitive to the SDGs, we built a term-document matrix for each enterprise and applied a binary classifier. The binary classifier was selected by comparing commonly used ML methods and looking for the presence of a set of predefined sustainability-related words[1] on each website. In particular, we tested the following ML methods: naive Bayes, gradient boosting classifier, random forest classifier, neural network classifier, logistic regression, k-nearest neighbours classifier, decision tree classifier, and support vector machine classifier (Russell & Norvig, 2020). The dataset

[1] 'sdg', 'sviluppo', 'sostenibil', 'sustainab', 'develop', 'agenda 2030', 'povert', 'fame', 'nutrizione', 'educazione', 'inclusiv', 'uguaglianza', 'emancipa', 'dignit', 'diseguaglianz', 'insediament', 'climatic', 'pace', 'giustizia', 'responsabil', 'parit', 'istruzione', 'benessere', 'rinnov', and 'energi'.

was split into training and test sets at 70% and 30%, respectively, to train and test the ML methods. The random forest classifier (Breiman, 2001) achieved superior performance, with accuracy = 0.87, precision = 0.85, recall = 0.89, specificity = 0.84, and f1 = 0.87 on the test set.[2]

Finally, after the entire process, from the subset K, we obtained an organized dataset with 10 variables—the target and the nine variables obtained for K—on 51,753 enterprises. We referred to this dataset as B and considered it a nonprobability sample. Note that in this case, for all the units in K, we were able to retrieve information from websites. Therefore, B had the same size as K. Moreover, regarding the variables in B, only three could be used as auxiliary variables in a statistical model: (i) the number of employees of an enterprise averaged over the years, (ii) the turnover volume indicator, and (iii) the NACE code. The province variable was used to identify the area, and the other variables used to identify an enterprise included VAT, name, and address.

2.2 ISTAT Special Survey on Enterprises' Perspectives After the COVID-19 Emergency

The probabilistic survey 'ISTAT Special Survey on Enterprises' Perspectives After the Covid-19 Emergency' (Costa et al., 2022; ISTAT, 2022), provided by ISTAT, is available on the same target population U and is denoted as A. We considered the third edition of the survey, conducted between 16 November and 17 December 2021, which updated the information collected in previous editions by measuring the behaviour and strategies of companies almost 2 years after the pandemic began.

The design of the 'ISTAT Special Survey on Enterprises' Perspectives after the COVID-19 Emergency' is two-phase sampling. The dimensions that defined surveyed companies were as follows: (i) economic activity; (ii) company size in terms of average number of employees (in classes); and (iii) territory of residence (i.e. where the company is located). ASIA was used to define the list of units of the target population. The unit of analysis was the enterprise (considered as a legal unit). In the first phase of the design, a stratified simple random sample was implemented. The strata were defined according to the economic activity, the company size, and the place where the enterprise was located. The variables used to calculate the probabilities of selection of the final units in the second phase were as follows: introduction by the company of at least one innovation, recruitment of new resources by the company, and self-financing of the company. Furthermore, to account for the nonresponse of units in the first phase, the weights were modified using adjustment coefficients obtained as the inverse of the estimated response propensity. The models for estimating the propensity to respond were of the random forest type. For further information on this survey design, please refer to ISTAT (2022).

In particular, in this paper, we considered a specific subsample of the survey that selected enterprises with 10 or more employees in the four considered NACE sectors. The final sample size comprised 19,606 enterprises among approximately 90,000 surveyed enterprises. The survey was not planned to provide reliable estimates at the provincial level (NUTS 3) in Italy and for just a selection of NACE sectors. The sample size in the provinces ranged from 24 to 1,220 and less than 100 enterprises in 35% of the areas. Consequently, we could consider the provinces to be small areas because estimates based on the usual design-based approach had too-large sampling variability.

Datasets B and A shared three variables obtained from a direct (exact) linkage of the ASIA National Statistical Business Register: (i) the number of employees of the enterprise averaged over the years, (ii) the turnover volume indicator in classes, and (iii) the NACE code. Moreover, in the probability sample A, we had a specific variable that denoted whether an enterprise had a known website (i.e. a URL was available). In other words, this variable indicated whether an enterprise included in sample A was also present in sample B or not.

Figure 1 shows a visual representation of the data collection setting underlying the relationship between the ASIA register, probability sample A, and dataset B.

Notably, although we considered a specific application to the Italian context (Figure 1), the proposed method could be used in a very wide range of applications inside and outside NSIs globally.

[2] Let TP, TN, FN, and FP represent the true positive, true negative, false negative, and false positive, respectively. Then, accuracy $=(TP+TN)/(TP+TN+FN+FP)$, precision $=TP/(TP+FP)$, recall $=TP/(TP+FN)$, specificity $=TN/(TN+FP)$, and f1 is the harmonic mean between precision and recall.

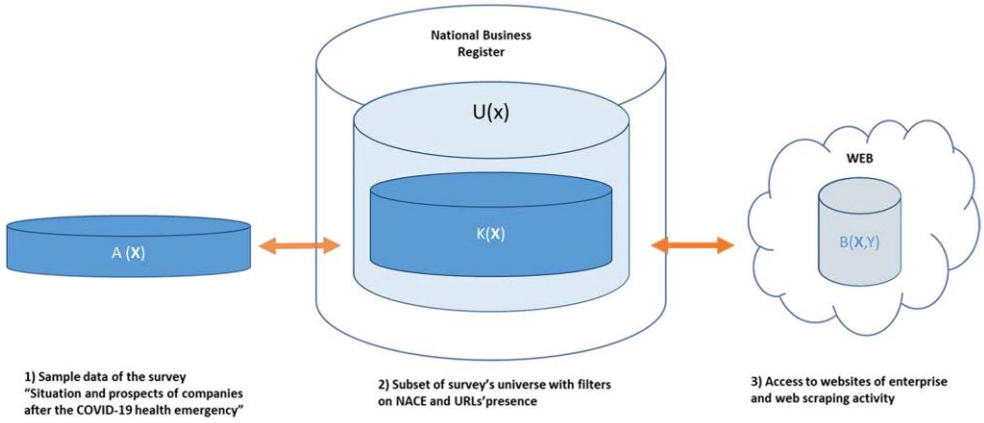

Figure 1. Data collection setting.

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

3.1 Effect of selection bias when the target variable is observed only in a nonprobability sample

Consider a finite population U of size N divided into m nonoverlapping subsets (domains of study or areas) U_i of size N_i, $i = 1, \ldots, m$. Let y_{ij} denote the value of the target variable Y for unit j in area i. Suppose that the parameters of interest are the area means $\theta_i = N_i^{-1} \sum_{j \in U_i} y_{ij}$, $i = 1, \ldots, m$.

First, we assume that we have access to a nonprobability sample (obtained from big data), which we denoted by B, of size n_B for the target population, with $B \subset U$. We also assume that the nonprobability sample is available in each area of interest, and denote by $B_i \subset U_i$ the nonprobability sample in area i, with n_{B_i} the sample size. We further assume that for the nonprobability sample B, we know a vector of variables, denoted by $\mathbf{z}_{ij}$ for unit j in area i, which are used as auxiliary variables in the proposed method.

Second, we assume that a probability sample—denoted by A—of size n_A is available. Sample A is drawn from U, the same target population of B, according to a known sampling design, such that the inclusion probability of unit j in area i is π_{ij} and its sampling weight is $w_{ij} = \pi_{ij}^{-1}$. We also assume that for each area A_i, a sample of size n_{A_i} is available and the areas of interest are 'small areas'. Generally, an area (or domain) is regarded as 'small' if the domain-specific sample size is too small to obtain direct estimates with acceptable statistical significance (Rao & Molina, 2015). These areas can be geographic areas, such as provinces or municipalities, sociodemographic groups (e.g. sex, age, and race), and other subpopulations (e.g. firms belonging to an industry subdivision). In these cases, model-assisted and model-based SAE techniques are usually employed (for a review of SAE methods, see Pratesi, 2016; Rao & Molina, 2015).

Third, let us assume that for the probability sample A, we know an inclusion indicator variable δ_{ij}, with $\delta_{ij} = 1$ if $j \in B_i$ (the unit j is both in A_i and in B_i) and $\delta_{ij} = 0$; otherwise (the unit j is only in A_i), $i = 1, \ldots, m$, $j = 1, \ldots, N_i$. The inclusion indicator variable δ_{ij} can be attached also in the nonprobability sample B and, in this case, is obviously always equal to one ($\delta_{ij} = 1 \; \forall \; j \in B_i$).

We further assume that the study variable y_{ij} is only observed for units in B. Therefore, the probability sample A does not contain the variable of interest; however, we assume that it contains some of the auxiliary variables in B. Without loss of generality, let these auxiliary variables be denoted by $\mathbf{x}_{ij}$ for unit j in area i. Moreover, let us recall that for units in sample A, we know the inclusion indicator variable δ_{ij}—that is, if for unit j in A_i, $\delta_{ij} = 1$, then the unit j is also in B_i; if for unit j in A_i, $\delta_{ij} = 0$, then the unit j is not in B_i (and y_{ij} is consequently unknown). This is possible because δ_{ij} is known from sample A, as stated earlier. The opposite case is not true; that is, for units in B, $\delta_{ij} = 1 \; \forall \; i = 1, \ldots, m$ and $j = 1, \ldots, n_{B_i}$ (obviously every unit in B is in B), and we do not know if unit j is or is not in A_i. However, this is not important for the application of the proposed method. In our application, δ_{ij} indicates the presence/absence of a URL, B is a nonprobability sample built

through web scraping from the universe U, whereas A is a probability sample selected from U, for which the variable about the presence of a URL is surveyed. Therefore, all the units in B have $\delta_{ij} = 1$ given that all have a URL (otherwise they cannot be in B), the units in A have $\delta_{ij} = 1$ if they have a URL (and, in this case, the unit is also in B), and $\delta_{ij} = 0$ if they do not have a URL (and, in this case, the unit is not in B).

The nonprobability sample B and the probability sample A have some common auxiliary variables. Consequently, the available data can be denoted by $\{(\delta_{ij} = 1, y_{ij}, \mathbf{z}_{ij}),\ j \in B_i,\ i = 1, \ldots, m\}$, and $\{(\delta_{ij}, \mathbf{x}_{ij}, w_{ij}),\ j \in A_i,\ i = 1, \ldots, m\}$. Moreover, among $\mathbf{z}_{ij}$ and $\mathbf{x}_{ij}$, there are some common variables.

To define a naive estimator and the self-selection error, we attach the variable δ_{ij} to the population U so that $\delta_{ij} = 1$ if the unit j is in B_i and $\delta_{ij} = 0$ if the unit j is not in B_i (note that knowing δ_{ij} at the population level is not required by the proposed method). Using the nonprobability sample B, we can estimate our target parameters—the population area means θ_i—using the following equation:

$$\tilde{\theta}_{B_i} = \frac{1}{n_{B_i}} \sum_{j \in U_i} \delta_{ij} y_{ij} = \frac{1}{n_{B_i}} \sum_{j \in B_i} y_{ij}. \tag{1}$$

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 & Kim, 2020). Therefore, the sample mean $\tilde{\theta}_{B_i}$ obtained from the nonprobability sample data is biased.

According to Kim and Wang (2019), the error of $\tilde{\theta}_{B_i}$ is given by

$$\tilde{\theta}_{B_i} - \theta_i = \frac{1}{f_{B_i}} \operatorname{cov}(\delta_{ij}, y_{ij}),$$

where $f_{B_i} = n_{B_i}/N_i$ and

$$\operatorname{cov}(\delta_{ij}, y_{ij}) = \frac{1}{N_i} \sum_{j \in U_i} (\delta_{ij} - \bar{\delta}_{N_i})(y_{ij} - \theta_i),$$

with $\bar{\delta}_{N_i} = N_i^{-1} \sum_{j \in U_i} \delta_{ij}$. For a general random mechanism, we have

$$\begin{aligned} E_{\delta_{ij}}[(\tilde{\theta}_{B_i} - \theta_i)^2] &= \frac{1}{f_{B_i}^2} E_{\delta_{ij}}[\rho(\delta_{ij}, y_{ij})^2 \operatorname{var}(\delta_{ij}) \operatorname{var}(y_{ij})] \\ &= E_{\delta_{ij}}[\rho(\delta_{ij}, y_{ij})^2] \cdot \left(\frac{1}{f_{B_i}} - 1\right) \cdot \sigma_i^2, \end{aligned} \tag{2}$$

where $E_{\delta_{ij}}[\,\cdot\,]$ denotes the expectation with respect to the random mechanism δ_{ij} and σ_i^2 denotes the population variance of y in area i.

The identity in equation (2) highlights that the selection bias of $\tilde{\theta}_{B_i}$ is determined by three factors. According to the terminology of Meng (2018), $\rho(\delta_{ij}, y_{ij})$ is called *data quality*, $\sqrt{(1/f_{B_i} - 1)}$ is *data quantity*, and σ_i is *problem difficulty*. Data quality represents the most critical term in determining bias and is approximately zero on average under simple random sampling. The first term in equation (2), $E_{\delta_{ij}}[\rho(\delta_{ij}, y_{ij})^2]$, called *data defect index* (DDI), determines the level of departure from simple random sampling. Indeed, under an equal probability sampling design, $E_{\delta_{ij}}[\rho(\delta_{ij}, y_{ij})] = 0$, and the DDI is of the order $O(1/N_i)$, implying that $E_{\delta_{ij}}[(\tilde{\theta}_{B_i} - \theta_i)^2] = O(n_{B_i}^{-1})$. Meanwhile, for other sampling designs with $E_{\delta_{ij}}[\rho(\delta_{ij}, y_{ij})] \neq 0$, the DDI becomes significant with the order $O(1)$, implying that $E_{\delta_{ij}}[(\tilde{\theta}_{B_i} - \theta_i)^2] = O(n_{B_i}^{-1}(N_i - 1))$. Thus, a nonprobability sampling design subjects the analysis results to selection bias.

3.2 Inference from a nonprobability sample combined with a probability sample in SAE

The above setting is suitable for the applied problem presented in this study. We now propose a technique to make a valid inference from a nonprobability sample (e.g. from big data sources)

that shares auxiliary variables with a probability sample to obtain reliable estimates at a small area level.

Combining information from a nonprobability sample and auxiliary information from a probability sample can be considered a data integration problem. Data integration, which is designed to combine information from two independent surveys of the same target population, is an adequate statistical approach for dealing with nonprobability sample selection bias by including a probability sample (Kim & Tam, 2021; Lohr & Raghunathan, 2017).

In the SAE framework, the use of multiple data sources is very common. Indeed, SAE methods combine survey data with auxiliary data sources, such as census or administrative data. This auxiliary/census information is then used as explanatory variables in a regression equation to predict the target variable. *Standard industry* SAE models employ a hierarchical model that includes random area effects to account for between-area variation. SAE models are classified into two categories according to the available data on the target variable: (i) area-level models and (ii) unit-level models. If information is available at the unit level, the standard unit-level SAE models proposed by Battese et al. (1988), Jiang and Lahiri (2001), and Jiang (2003) may be used.

When the quantities of interest are the area means, they can be defined in terms of a linear combination between the observed and unobserved units as follows:

$$\theta_i = N_i^{-1}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} y_{ij}\right\}, \tag{3}$$

where s_i and r_i are, respectively, the sets of sampled and unsampled units in area i. Substituting the values for the unsampled units with their predictions, we obtain the following:

$$\hat{\theta}_i = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}\right], \tag{4}$$

where $\hat{y}_{ij}$ is predicted according to the nature of the outcome (continuous or discrete). We consider the case where a probability sample A and a nonprobability sample (obtained from big data) B are available, with each small area containing at least two units, and we assume that the selection mechanism for the nonprobability sample B is noninformative; that is,

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

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 $\mathbf{x}_{ij}$. Notably, we can attach δ_{ij} (the nonprobability sample B inclusion indicator) to sample A. In other words, among the elements of sample A, membership information can be obtained in the nonprobabilistic sample B.

We can use the data $\{(\delta_{ij}, w_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for participation probabilities or propensity scores ($P(\delta_{ij} = 1|\mathbf{x}_{ij}) = p(\mathbf{x}_{ij}, \boldsymbol{\lambda})$) in sample B based on missing-at-random, where $\boldsymbol{\lambda}$ is the vector of the binary regression coefficients. Usually, a logistic regression model for the binary variable δ_{ij} is employed to obtain the estimators of $p(\mathbf{x}_{ij}, \boldsymbol{\lambda})$, $\hat{p}_{ij}(\hat{\boldsymbol{\lambda}})$, in sample B. However, the hierarchical structure of the data should be considered in the estimation model of the propensity scores. Therefore, the participation probabilities $p(\mathbf{x}_{ij}, \boldsymbol{\lambda})$ are also conditioned on the random effects u_i:

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

Consequently, we consider the following generalized linear random intercept model for the propensity scores:

$$\hat{p}_{ij}(\mathbf{x}_{ij}, \hat{\boldsymbol{\lambda}}, \hat{u}_i) = g^{-1}(\mathbf{x}_{ij}^T\hat{\boldsymbol{\lambda}} + \hat{u}_i), \tag{5}$$

where $g(\cdot)$ denotes a logit link function and $\hat{\boldsymbol{\lambda}}$ 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. More details can be found in Rao and Molina (2015).

To estimate the parameters of the generalized linear random intercept model for the propensity scores in equation (5), we use a pseudo maximum-likelihood (PML) approach. In the framework of a multilevel model with random effects, the PML approach has been proposed by several authors (Asparouhov, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006; Schirripa Spagnolo et al., 2020). All these approaches include sampling weights in the estimation procedure in multilevel models. In particular, census likelihood is estimated by weighting the sample likelihood by including sampling weights in the log-likelihood function. In our probability sample data, we have first-level sampling weights that account for unequal selection probabilities of the enterprises. Meanwhile, as the provinces are not planned domains in the survey, they do not need weights. Thus, we left the second level (provinces) unweighted in the estimation procedure. Accordingly, the log-likelihood function for the sample units can be expressed as

$$\log L(\boldsymbol{\lambda}, \sigma_u^2) = \sum_{i=1}^{m} \log \int \left[\exp\left\{ \sum_{j \in A_i} w_{ij} \log L_{ij}(\boldsymbol{\lambda}, \sigma_u^2 | u_i) \right\} \right] f(u_i)\, \mathrm{d}u_i, \tag{6}$$

where σ_u^2 is the variance of the random effects u_i, w_{ij} denotes the sampling weight of unit j in area i, $f(u_i)$ is the density of the random effects, and $\log L_{ij}(\boldsymbol{\lambda}, \sigma_u^2 | u_i)$ is the log-likelihood contribution of the first-level units (enterprises) conditioned on random effects. The maximization of the weighted log-likelihood in equation (6) involves the computation of several integrals that do not have a closed-form solution; thus, a numerical approximation technique is required. In particular, we use the `R` function `glmer` of the package `lme4` (Bates et al., 2015), which implements adaptive Gauss–Hermite quadrature.

In developing our estimator, we assume that the following working population model holds for sample B:

$$E[y_{ij} | \mathbf{z}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{z}_{ij}^T \boldsymbol{\beta} + \gamma_i\right), \tag{7}$$

where $h(\cdot)$ denotes the link function, assumed to be known and invertible, $\mathbf{z}_{ij}$ is a vector of auxiliary variables for unit j in area i, and γ_i denotes the area-specific random effect for area i characterizing the between-area differences in the distribution of y_{ij} given $\mathbf{z}_{ij}$. Notably, the auxiliary variables used here—$\mathbf{z}_{ij}$s—could be the same as those used to fit the propensity model—$\mathbf{x}_{ij}$s. In our application, we have the availability of only three auxiliary variables, specified in Section 2.2. In this particular case, there is no need to resort to variable selection methods. However, in the case of many auxiliary variables, some selection methods for SAE are available, such as those in Jiang et al. (2008); Müller et al. (2013); Tibshirani (1996); van den Brakel and Buelens (2014). The model in equation (7) includes three key special cases: the linear model obtained with $h(\cdot)$ is equal to the identity function, and y_{ij} is a continuous variable; the logistic generalized linear random intercept model, where $h(\cdot)$ is the logit link function, and the outcome variable is binomial; and the Poisson-log generalized linear random intercept model, where $h(\cdot)$ is the log link function, and the y_{ij} values are considered independent Poisson random realizations. In addition, in this case, the model in equation (7) is estimated using the `R` function `glmer` of the package `lme4` (Bates et al., 2015). Notably, in this case, we cannot include sampling weights in the estimation procedure because we use data from the nonprobability sample B.

Using data from the nonprobability sample B and assuming that the model is correctly specified, we obtain an estimator $\hat{\boldsymbol{\beta}}$, which is consistent for $\boldsymbol{\beta}$ and a predictor $\hat{\gamma}_i$ (Rao, 2021).

Finally, a SAE DR estimator of the population average in area i is given by

$$\hat{\theta}_{i;\mathrm{DR}} = \frac{1}{N_i}\left\{ \sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\boldsymbol{\lambda}}, \hat{u}_i)} (y_{ij} - \hat{\mu}_{ij}) + \sum_{j \in A_i} w_{ij} \hat{\mu}_{ij} \right\}, \tag{8}$$

where $\hat{\mu}_{ij} = h^{-1}(\mathbf{z}_{ij}\hat{\boldsymbol{\beta}} + \hat{\gamma}_i)$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}_i$, respectively, denote the estimated regression coefficients and random effects based on the nonprobability sample B and w_{ij} denotes the sampling weight of unit j in area i.

The estimator in equation (8) is DR in the sense that it is consistent if only one between the model for propensity scores and the model for the study variable is correctly specified (Kim & Wang, 2019; Rao, 2021). To show the double robustness of $\hat{\theta}_{i;\mathrm{DR}}$, suppose that the target variable is observed in the probability sample A; then, the Horvitz–Thompson estimator of θ_i would be equal to

$$\hat{\theta}_{i;HT} = \frac{1}{N_i} \sum_{j \in A_i} w_{ij} y_{ij}.$$

Therefore, we can express

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

where $\hat{e}_{ij} = y_{ij} - \hat{\mu}_{ij}$. If the propensity score model in equation (5) is correctly specified,

$$\mathrm{E}_{\delta_{ij}}\left(\hat{\theta}_{i;\mathrm{DR}} - \hat{\theta}_{i;HT}\right) \approx \frac{1}{N_i} \left\{ \sum_{j \in U_i} \hat{e}_{ij} - \sum_{j \in A_i} w_{ij} \hat{e}_{ij} \right\}$$

is design-unbiased of zero. Therefore, $\hat{\theta}_{i;\mathrm{DR}}$ is asymptotically unbiased under the model in Equation (5). In addition, if $E[y_{ij}|\mathbf{x}_{ij}, \gamma_i]$ is correctly specified, then

$$\frac{1}{N_i} \mathrm{E}\left(\sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\boldsymbol{\lambda}}, \hat{u}_i)} \hat{e}_{ij} | B \right) = \frac{1}{N_i} \sum_{j \in B_i} \frac{1}{\hat{p}_{ij}(\hat{\boldsymbol{\lambda}}, \hat{u}_i)} \mathrm{E}(\hat{e}_{ij}|B),$$

$$\frac{1}{N_i} \mathrm{E}\left(\sum_{j \in A_i} w_{ij} \hat{e}_{ij} \right) = \frac{1}{N_i} \sum_{j \in U_i} \mathrm{E}(\hat{e}_{ij}|B),$$

and $\mathrm{E}(\hat{e}_{ij}|B) = 0$ (under $E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij}$ and the MAR assumption). It follows that $\mathrm{E}(\hat{\theta}_{i;\mathrm{DR}} - \hat{\theta}_{i;HT}) \approx 0$ if the outcome regression model is correctly specified. Thus, we establish the double robustness of $\hat{\theta}_{i;\mathrm{DR}}$. See also Kim and Wang (2019) and Yang and Kim (2020) for further reference.

3.3 Bootstrap variance estimation

Deriving an analytical estimator of the variance of the estimator in equation (8) is not straightforward. Alternatively, we propose a bootstrap procedure to approximate it. The steps of the bootstrap procedure are as follows:

1. Extract a sample of size n_A from sample A using the inclusion probabilities π_{ij} to obtain a bootstrap replicate denoted by $\{(\delta^{\star}_{ij}, w^{\star}_{ij}, \mathbf{x}^{\star}_{ij})\} \in A^{\star}$.
2. Extract a simple random sample with replacement of size n_B from sample B to obtain the bootstrap $\{(y^{\star}_{ij}, \mathbf{x}^{\star}_{ij})\} \in B^{\star}$.
3. Use the model in equation (5) applied to the bootstrap sample to obtain the bootstrap propensity score $\hat{p}^{\star}_{ij}(\mathbf{x}_{ij}, \hat{\boldsymbol{\lambda}}^{\star}, \hat{u}^{\star}_i)$ by using scaled bootstrap weights, $\tilde{w}^{\star}_{ij} = w^{\star}_{ij} N_i / \sum_{j \in i} w^{\star}_{ij}$.
4. Fit the model in equation (7) on the bootstrap sample $B^{\star}$ to estimate the regression coefficients $\hat{\boldsymbol{\beta}}^{\star}$ and area-specific random effects $\hat{\gamma}^{\star}_i$.
5. Use equation (8) to obtain the DR estimator $\hat{\theta}^{\star}_{i;\mathrm{DR}}$.
6. Repeat Steps 1–5 independently for L times. The resulting bootstrap variance estimator of $\hat{\theta}_{i;\mathrm{DR}}$ is computed as follows (Kim et al., 2021):

$$\hat{V}(\hat{\theta}_{i;\mathrm{DR}}) = \frac{1}{L} \sum_{l=1}^{L} \left(\hat{\theta}^{\star(l)}_{i;\mathrm{DR}} - \hat{\theta}_{i;\mathrm{DR}} \right)^2, \tag{9}$$

where $\hat{\theta}^{\star(l)}_{i;\mathrm{DR}}$ denotes the replicate version of the estimator $\hat{\theta}_{i;\mathrm{DR}}$ in area i for replication l.

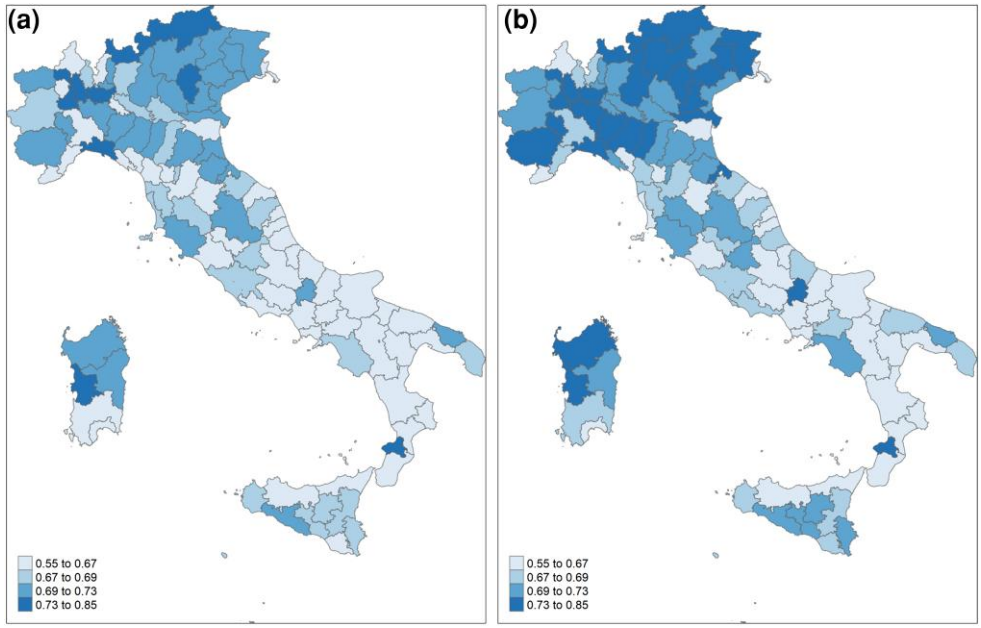

Figure 2. Estimated proportion of *sustainable development goal-sensitive enterprises* in the Italian provinces using the small area estimation double robust estimator (a) and naive direct estimator (b).

4 Application results

From OBEC, the SDG sensitivity of enterprises could be derived. Our target parameter was the proportion of SDG-sensitive enterprises at the provincial level (NUTS 3). To this end, we used the data described in Section 2. In summary, we used the nonprobability sample B obtained by scraping the websites of the target population of enterprises, from which we obtained the binary target variable *SDG sensitivity* and other auxiliary variables and a probability sample A, 'Situation and Prospects of Companies after the COVID-19 Health Emergency', which shared auxiliary variables with B and where the province sample sizes (from A) were small. Therefore, given the available data, we applied the proposed SAE DR method to obtain the desired estimates.

In Figure 2, we present two maps: the estimated proportion of *SDG sensitivity* using the proposed SAE DR estimator (left panel) and the naive direct estimator ($\tilde{\theta}_{B_i}$, right panel).

The three provinces with the highest proportion of enterprises sensitive to SDGs (estimated using the SAE DR method) were Bolzano (84.1%), Vercelli (77.8%), and Vibo Valentia (75.2%), whereas those with the lowest values were Massa Carrara (59.0%), Crotone (59.9%), and Campobasso (60.8%). Despite this, a clear north–south dualism, with greater attention to sustainability in the north, was evident from both maps. The SAE DR estimator seemed to smoothen the estimates more, as expected, according to the use of a model to correct bias. Although the two geographical distributions of the estimates were similar, the probable bias of the naive direct estimator could mislead policymakers.

By assessing the variability of the estimates, it was found that the SAE DR estimates for 106 out of 107 areas had coefficients of variation (CV) below 16.6%, rendering them reliable according to the classification standards set by Statistics Canada (Statistics Canada, 2010). For the remaining area, the CV was acceptable because it fell between 16.6% and 33.3%.

To evaluate how the estimated propensity scores impacted the final estimates, we also employed a more standard SAE predictor. In particular, we used the following predictor: $\hat{\theta}_{i;S} = N_i^{-1}\{\sum_{j\in B_i} y_{ij} + \sum_{j\in A_i} w_{ij}\hat{\mu}_{ij}(1-\delta_{ij})\}$, where $\hat{\mu}_{ij} = h^{-1}(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{\gamma}_i)$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}_i$, respectively, denote the estimated regression coefficients and random effects based on the nonprobability sample B and w_{ij} denotes the sampling weight of unit j in area i. In this case, we tended to underestimate the proportion of SDG-sensitive enterprises, whereas the SAE DR method provided estimates

between 0.59 and 0.84, with a median value of 0.67; by avoiding the inclusion of the propensity scores, the range of the estimates was between 0.38 and 0.73, with a median value of 0.56.

The analysis presented in this article serves, to the best of our knowledge, as a first attempt to estimate the interest in sustainability issues in companies at the Italian provincial level.

On this topic, the Italian newspaper 'Il Sole 24 Ore' (an influential Italian newspaper on economic topics) has been producing a ranking of the most sustainable Italian enterprises since 2021 called Sustainability Leader (Il Sole 24 Ore, 2023). Sustainability Leader is a list of 240 Italian firms considered the most sustainable. For the 2023 edition, 1,500 Italian firms were specifically analysed on the basis of their published sustainability reports and financial statements. Enterprises active in Italy and belonging to the companies with the highest turnover or listed on the Italian stock exchange were chosen and then categorized into two based on their 2021 turnover size: large companies with 200 winners, whose 2021 turnover was more than 100 million euros, and medium to small companies with 40 winners, whose turnover was less than 100 million euros. For the identified companies, we verified whether the company published a sustainability report or an equivalent report (e.g. nonfinancial statement, integrated report, and social balance sheet). In addition, a registration procedure was activated for the Sustainability Leader competition on the journal website to collect voluntary businesses. The registration phase was active from 25 October to 31 December 2021. Companies could submit their sustainability and financial reports online, if available. Alternatively, a questionnaire was made available to the target companies, where the data required for the analysis could be inserted. In this case, the released data were verified again by the participating companies through a document signed by their chief executive officer or manager. The analysis was based on the responsibility of companies (or corporate social responsibility). This ranking considered three dimensions: environmental, social, and economic.

Although the ranking is essential in monitoring SDG sensitivities for enterprises in Italy, we believe that it is inadequate for understanding the actual situation of the country in terms of sensitivity to sustainability. First, the ranking produced by 'Il Sole 24 Ore' considers only quite large companies that agree to participate in the analysis without considering that small- and medium-sized enterprises account for the majority of Italian firms, providing approximately 80% of the industrial and service labour force and generating approximately two-thirds of turnover and added value (OECD, 2022). Although not all small companies have a website, even if they are widespread, especially following the COVID-19 pandemic, we believe that the proposed method can better capture the degree of attention to sustainability in our country by considering all territories currently excluded because of the absence of large companies in their area. Second, the main objective of public funding is to highlight the most difficult situations to organize investments and raise awareness on the issue of sustainability. Therefore, our analysis shifted attention from companies to territories, allowing us to consider not only big companies that volunteer, but also the entire Italian territory, thereby highlighting the provinces that require greater public intervention. Finally, the Sustainability Leader ranking represents a picture of the Italian context that is not updated. For example, the list compiled in 2023 provides information on firms' conditions in 2021. Meanwhile, our analysis, which can be updated without time constraints, can provide a current snapshot of the country, considering the simplicity of updating company websites.

Our analysis offers the possibility of scaling down to a very granular territorial level, such as provinces (NUTS 3), to produce results that would otherwise only be available at the level of the four macroregions (northern–western, northern–eastern, central, and southern Italy). Indeed, ISTAT released a report on the sustainable practices of companies in 2022 and the prospects for 2023–2025 (ISTAT , 2023). The data made available in this report came from the monthly questionnaire usually used for trust surveys, which was specifically implemented to include an 'ad hoc section' regarding the sustainability and circularity of production processes. The obtained results were reliable only at the level of the above four macroregions. Furthermore, the implementation of a specific one-off questionnaire does not allow information systematization, which is possible with the proposed method.

Moreover, our approach can be viewed as a complementary analysis of the Permanent Census of the Enterprises, launched by ISTAT in 2019 to collect data on several aspects of the Italian business system, including actions performed for environmental sustainability and social responsibility (Istat, 2021a). Census data reveal that, in 2018, more than 60% of firms with at least three employees declared that they had implemented sustainability actions, and an increasing trend was expected in the 3 years 2019–2021. However, these data are released every 3 years, and the

procedure used to obtain them is resource-intensive. Meanwhile, our approach offers a technique for obtaining data more frequently and with fewer resources required.

5 Validation of the proposed method through MC simulations

To evaluate the reliability of the proposed method, we ran a small MC simulation to evaluate the statistical properties of the proposed SAE DR estimator and the bootstrap estimator of its variance.

Additionally, to validate the proposed method using real data, we present in Section S1 of the online supplementary material an application with real data where SAE DR and direct naive estimates are compared with an unbiased direct estimator obtained from a sample survey.

The simulation was performed to (i) compare the SAE DR estimator based on the mixed model approach with the naive direct estimator (from a nonprobability sample) and (ii) check the validity of the proposed variance for the SAE DR estimator. The setup for the simulation was based on Kim and Wang (2019) and Chambers et al. (2016).

We considered the following two outcome models for generating a finite population for $m = 100$ small areas

(i) Linear model

$$\begin{aligned} y_{ij}|\gamma_i &\sim \text{Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, m; \; j = 1, \ldots, N_i \\ \pi_{ij} &= \exp(\eta_{ij})\{1 + \exp(\eta_{ij})\}^{-1} \\ \eta_{ij} &= x_{1,ij} + x_{2,ij} + \gamma_i \end{aligned} \tag{10}$$

(ii) Nonlinear model

$$\begin{aligned} y_{ij}|\gamma_i &\sim \text{Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, m; \; j = 1, \ldots, N_i \\ \pi_{ij} &= \exp(\eta_{ij})\{1 + \exp(\eta_{ij})\}^{-1} \\ \eta_{ij} &= 0.5(x_{1,ij} - 1.5)^2 + x_{2,ij} + \gamma_i. \end{aligned} \tag{11}$$

For both models, two auxiliary variables were generated as follows: $x_{1,ij} \sim N(1, 0.5)$ and $x_{2,ij} \sim \text{Unif}(a_i, b_i)$, for $a_i = -1$ and $b_i = m/16$, $i = 1, \ldots, m$; $j = 1, \ldots, N_i$. The small area effects γ_i were independently drawn from a normal distribution with a mean of 0 and variance $\sigma^2_\gamma = 0.25$. We fixed the population size in each small area to be $N_i = 1,000$. We used simple random sampling without replacement within each area to obtain an independent sample A of size $n = 1,000$, with $n_i = 10$. The sampling indicator of the nonprobability sample was generated by $\delta_{ij} \sim \text{Ber}(p_{ij})$ independently for $j = 1, \ldots, N$ and $i = 1, \ldots, m$, and we considered the following two propensity score models:

(i) Linear propensity score model

$$p_{ij} = \frac{\exp(x_{2,ij} + u_i)}{1 + \exp(x_{2,ij} + u_i)}. \tag{12}$$

(ii) Nonlinear propensity score model

$$p_{ij} = \frac{\exp(-0.5 + 0.5 \cdot (x_{2,ij} - 2)^2 + u_i)}{1 + \exp(-0.5 + 0.5 \cdot (x_{2,ij} - 2)^2 + u_i)}, \tag{13}$$

where $u_i \sim N(0, 0.1)$.

We considered 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. A finite population is generated using equation (10), and the sampling indicator of the big data sample is generated using equation (12).

2. The outcome regression model is linear, and the big data propensity score model is nonlinear. Equations (10) and (13) are used to generate the finite population and the sampling indicator of the big data sample, respectively.
3. The outcome regression model is nonlinear, whereas the big data propensity score model is linear. The finite population and the sampling indicator of the big data sample are generated using equations (11) and (12), respectively.
4. Both the outcome regression model and the big data propensity score model are nonlinear. The finite population is generated using equation (11), whereas the sampling indicator of the big data sample is generated using equation (13).

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

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

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

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

These choices allowed us to evaluate the performance of our estimator in the presence of misspecification of the propensity model (Scenario 2), outcome regression model (Scenario 3), and both models (Scenario 4).

For each scenario, we conducted $R = 500$ MC simulations to measure the bias and variability of the SAE DR estimator. To summarize the results, we used the following performance indicators:

- $\mathrm{Bias}(\tau_i) = R^{-1} \sum_{r=1}^{R} (\tau_i^{(r)} - \theta_i^{(r)})$
- $\mathrm{AbsBias}(\tau_i) = |R^{-1} \sum_{r=1}^{R} (\tau_i^{(r)} - \theta_i^{(r)})$
- $\mathrm{RB}(\tau_i) = R^{-1} \sum_{r=1}^{R} \frac{(\tau_i^{(r)} - \theta_i^{(r)})}{\theta_i^{(r)}} \times 100$
- $\mathrm{MSE}(\tau_i) = R^{-1} \sum_{r=1}^{R} (\tau_i^{(r)} - \theta_i^{(r)})^2$

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

Table 1 summarizes the simulation results. For all four scenarios, the SAE DR estimator outperformed the naive direct estimator in terms of bias. Indeed, the naive direct estimator of the sample mean was biased in all scenarios, indicating a positive relative bias (RB) in the first and third scenarios of approximately 7.6% and 4.3%, respectively, and a negative RB in the second and fourth scenarios when the propensity score model was nonlinear. The RB of the SAE DR estimator was virtually null, indicating that it tended to be robust to the misspecification of both the outcome regression and propensity score models. Regarding MSE, the naive direct estimator tended to have slightly lower values than those of the SAE DR estimator, but the difference was negligible. In Section S2 of the online supplementary material, we replicate the four simulation scenarios where the sample A is selected according to Poisson sampling. The results consistently showed that the SAE DR estimator had a significantly lower bias than the naive direct estimator in all scenarios.

As highlighted above, the second aim of this simulation study was to evaluate the performance of the proposed bootstrap procedure for estimating variance (equation (9)). In Table 2, we report the median of the area-specific bias expressed in relative terms (%) of the bootstrap SE estimator, $\widehat{\mathrm{SE}}(\hat{\theta}_{i;\mathrm{DR}}) = \sqrt{\hat{V}(\hat{\theta}_{i;\mathrm{DR}})}$. We generated 500 bootstrap replicates.

The results in Table 2 demonstrate that the proposed bootstrap procedure performed well across all scenarios. Specifically, the bootstrap variance estimator showed a slight positive bias in the first scenario, where both the outcome regression model and the big data propensity score model were linear, as well as in the final scenario, where both models were nonlinear. In Scenario

Table 1. Median over the areas of Bias, AbsBias, relative bias, and MSE over 500 Monte Carlo simulations for each scenario

Estimator	Bias	AbsBias	RB (%)	MSE
Scenario 1				
Naive direct	0.053	0.053	7.632	0.003
SAE DR	0.000	0.002	0.022	0.004
Scenario 2				
Naive direct	−0.035	0.035	−4.746	0.001
SAE DR	0.000	0.003	−0.057	0.004
Scenario 3				
Naive direct	0.034	0.034	4.257	0.001
SAE DR	0.000	0.002	−0.016	0.003
Scenario 4				
Naive direct	−0.023	0.023	−2.643	0.001
SAE DR	0.000	0.002	0.037	0.002

4, where only the outcome regression model was nonlinear, RB increased to approximately 2.3%. In these cases, the bootstrap procedure appeared to be conservative. Conversely, a negative bias was observed in Scenario 2, where the outcome regression model was linear but the big data propensity score model was nonlinear. However, even in this instance, the bias remained within an acceptable range. Notably, this procedure represented an initial attempt to estimate the variability of the proposed SAE DR estimator. Further research is needed to explore variance estimation through additional and more complex resampling techniques or analytical methods.

We also show in Table 3 the empirical coverage rates (CRs) and the semilengths for nominal 95% confidence intervals (CIs) based on normal-like CIs

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

CR was defined as the percentage of times the CI included the true population mean of the 500 simulations.

The actual CRs tended to be lower than the nominal level, especially in the last two scenarios. Notably, the practice of using estimated standard errors to build CIs, although common, has faced criticism. Hall and Maiti (2006), along with Chatterjee et al. (2008), explored the application of bootstrap methods for constructing CIs for small area parameters. They argued that the asymptotic assumptions underlying normal theory CIs may not hold in the context of small sample sizes typical of SAE. Although an attempt was made in Section S2 of the online supplementary material, further investigation into the use of bootstrap techniques for constructing CIs for the proposed DR SAE estimator remains a topic for future research.

6 Final remarks

In an era characterized by the spread of new data sources and unprecedented data availability, UN's 2030 Agenda for SDGs, with its overall goal of leaving no one behind, requires disaggregated and up-to-date data and statistics. Particularly, considering the recent European directives, it is essential to monitor the behaviour of Italian enterprises in terms of SDGs to evaluate policies and distribute European funding. Big data, considered in this study as nonprobability samples obtained from websites, are crucial sources of information—frequently updated, widely available, and inexpensive. Furthermore, owing to their wealth of information, they are often able to provide information not available in traditional probabilistic surveys. However, nonprobability samples tend to be affected by selection bias and other nonsampling errors. Consequently, to obtain high-quality indicators, these errors must be considered in estimation processes. An effective strategy to

Table 2. Median over the areas of relative bias (RB) of the bootstrap variance estimators over the 500 Monte Carlo simulations for each scenario

Scenario	RB (%)
Scenario 1	0.218
Scenario 2	−3.067
Scenario 3	2.320
Scenario 4	0.902

Table 3. Quartiles over the areas of empirical coverage rates and the median of semilengths of confidence intervals

Scenario	Coverage			Semilength
	Q1	Median	Q3	
Scenario 1	89.2	91.1	92.4	0.12
Scenario 2	89.2	90.9	92.2	0.12
Scenario 3	87.4	89.4	91.8	0.09
Scenario 4	86.4	88.5	90.6	0.09

adjust for self-selection is the use of a probability sample that shares common variables with the nonprobability sample. When the domains or areas of interest are not planned in the probability sample, the self-selection adjustment method can result in unreliable estimates because of the small sample sizes of the areas/domains. Therefore, in this study, we propose a new SAE DR estimator, which allowed us to obtain reliable estimates when the target variable was available only in the nonprobability sample and a probability sample with variables common to the nonprobability sample was available. This was realized in a framework where the areas of interest were unplanned in the design of a probabilistic survey; thus, they were usually small areas. The SAE DR estimator combined propensity weighting to improve the representativeness of the nonprobability sample and then used a statistical model to predict the study variable values for units of the probability sample. The SAE DR estimator allowed reliable and robust estimates at the provincial level (NUTS 3) regarding the sustainability of Italian enterprises, thereby reducing the bias inherent in nonprobability samples while maintaining acceptable estimate reliability. We also proposed a bootstrap procedure for the empirical estimation of the variance of the estimates. The focus of future work will be the development of an analytical variance estimator for the SAE DR predictor.

The results were validated using MC simulations, which showed that our estimator performed well in all tested scenarios. Moreover, the results were validated using a real-data application to an e-commerce case in Italy, as shown in Section S1 of the online supplementary material. In this case, unbiased estimates from a probability survey were available, indicating that our method provided satisfactory predictive performance and bias correction.

The findings of this study suggest that considerable work still needs to be done regarding sensitivity to the issues of the UN 2030 Agenda in Italy. Indeed, although the percentage of companies sensitive to SDGs in some provinces is approximately 80%, they are very few and are mainly located in the northern part of Italy. Meanwhile, in most of the provinces of the southern regions, the proportion is approximately 60%. However, there are exceptions, and spatial heterogeneity of the phenomena inside the regions is observed. This shows the relevance of disaggregating estimates to allow for analysis at the provincial level (NUTS 3).

Acknowledgments

The authors thank the anonymous reviewers and the associate editors for their valuable suggestions.

Conflicts of interest: None declared.

Funding

The work of Francesco Schirripa Spagnolo was carried out with the support of the 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'. The work of Nicola Salvati was carried out with the support of the project 'Quantification in the Context of Dataset Shift' (QuaDaSh) (Bando 2022 PNRR Prot. P2022TB5JF) and 'Future Artificial Intelligence Research' (FAIR—PE00000013). The work of Nicola Salvati and Francesco Schirripa Spagnolo was carried out with the support of the 'MAPPE' project, Programma 'PE GRINS—GRINS—GROWING RESILIENT, INCLUSIVE AND SUSTAINABLE' (cod. PE0000018 CUP: J33C22002910001). The work of Monica Pratesi, Monica Scannapieco, and Donato Summa was carried out with the support of the project 'Trusted Smart Statistics—Web Intelligence Network' (2020-PL-SmartStat—1010358).

Data availability

The data that support the findings of this study are available from the Italian National Statistics Institute (Istat) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Istat. The code for generating and analysing the simulated data presented in Section 5 is available upon request.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Asparouhov T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, *35*(3), 439–460. https://doi.org/10.1080/03610920500476598

Baker R., Brick J. M., Bates N. A., Battaglia M., Couper M. P., Dever J. A., Gile K. J., & Tourangeau R. (2013). Summary report of the AAPOR Task Force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Barcaroli G., Nurra A., Salamone S., Scannapieco M., Scarnò M., & Summa D. (2015). Internet as data source in the Istat survey on ICT in enterprises. *Austrian Journal of Statistics*, *44*(2), 31–43. https://doi.org/10.17713/ajs.v44i2.53

Barcaroli G., & Scannapieco M. (2019). Integration of ICT survey data and internet data from enterprises websites at the Italian National Institute of Statistics. *Statistical Journal of the IAOS*, *35*(4), 643–656. https://doi.org/10.3233/SJI-190553

Bates D., Mächler M., Bolker B., & Walker S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Battese G. E., Harter R. M., & Fuller W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36. https://doi.org/10.1080/01621459.1988.10478561

Beaumont J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, *46*(1), 1–28. https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-eng.pdf

Breiman L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chambers R., 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: Statistics in Society*, *179*(2), 453–479. https://doi.org/10.1111/rssa.12123

Chatterjee S., Lahiri P., & Li H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Annals of Statistics*, *36*(3), 1221–1245. https://doi.org/10.1214/07-AOS512

Chen Y., Li P., & Wu C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Citro C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*(2), 137–162. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X

Costa S., De Santis S., & Monducci R. (2022). Reacting to the COVID-19 crisis: State, strategies and perspectives of Italian firms. *Rivista di Statistica Ufficiale/Review of Official Statistics*, *1*, 73–107. https://www.istat.it/it/files/2022/05/RSU_1-2022_Article-3.pdf

Dinh T., Husmann A., & Melloni G. (2023). Corporate sustainability reporting in Europe: a scoping review. *Accounting in Europe*, *20*(1), 1–29. https://doi.org/10.1080/17449480.2022.2149345

Eurostat (2013). Scheveningen memorandum on big data and official statistics. In *DGINS2013 conference*. https://ec.europa.eu/eurostat/documents/13019146/13237859/Scheveningen-memorandum-27-09-13.pdf/2e730cdc-862f-4f27-bb43-2486c30298b6?t=1401195050000 [Accessed 06 November 2023].

Eurostat (2018). Bucharest memorandum on official statistics in a datafied society (trusted smart statistics). In *DGINS2018 conference*. https://ec.europa.eu/eurostat/documents/13019146/13239158/The+Bucharest+Memorandum+on+Trusted+Smart+Statistics+FINAL.pdf/59a1a348-a97c-4803-be45-6140af08e4d7?t=1539760880000 [Accessed 06 November 2023].

Fay R. E., & Herriot R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Grilli L., & Pratesi M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, *30*(1), 93–104. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20040016997

Hall P., & Maiti T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Annals of Statistics*, *34*, 1733–1750. https://doi.org/10.1214/009053606000000579

Il Sole 24 Ore (2023). Leader della sostenibilità 2023. https://lab24.ilsole24ore.com/leader-sostenibilita/ [Accessed 25 October 2023].

ISTAT (2020a). Comportamenti d'impresa e sviluppo sostenibile, Statistiche sperimentali. Istat. https://www.istat.it/it/files/2020/03/Imprese-e-sostenibilita-statistiche-sperimentali.pdf [Accessed 06 November 2023].

ISTAT (2020b). Sostenibilità nelle imprese: Aspetti ambientali e sociali., Censimenti Permanenti Imprese Istat. https://www.istat.it/it/files//2020/06/Sostenibilità-nelle-imprese.pdf [Accessed 06 November 2023].

ISTAT (2021a). Rapporto sulle imprese 2021. Struttura, comportamenti e performance dal censimento permanente. https://www.istat.it/storage/rapporti-tematici/imprese2021/Rapportoimprese2021.pdf [Accessed 06 November 2023].

ISTAT (2021b). Struttura e dimensione delle imprese secondo la nuova definizone—REgistro ASIA 2019. Nota metodologica, Technical report, Istat. https://www.istat.it/it/files//2021/11/Nota-metodologica-Registro-2019.pdf.

ISTAT (2022). Sistuazione e prospettive delle imprese nell'emergenza sanitaria COVID-19. Nota Metodologica. https://www.istat.it/it/files//2022/02/REPORT-COVID-IMPRESE_2022.pdf [Accessed 29 August 2023].

ISTAT (2023). Pratiche sostenibili delle imprese nel 2022 e le prospettive 2023–2025. https://www.istat.it/it/files//2023/04/Pratiche-sostenibili-delle-imprese.pdf [Accessed 25 October 2023].

Jiang J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, *111*(1-2), 117–127. https://doi.org/10.1016/S0378-3758(02)00293-8

Jiang J., & Lahiri P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, *53*(2), 217–243. https://doi.org/10.1023/A:1012410420337

Jiang J., Rao J. S., Gu Z., & Nguyen T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, *36*(4), 1669–1692. https://doi.org/10.1214/07-AOS517

Kim J. K., Park S., Chen Y., & Wu C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(3), 941–963. https://doi.org/10.1111/rssa.12696

Kim J.-K., & Tam S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, *89*(2), 382–401. https://doi.org/10.1111/insr.v89.2

Kim J. K., & Wang Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, *87*(S1), S177–S191. https://doi.org/10.1111/insr.12290

Lohr S. L., & Raghunathan T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293–312. https://doi.org/10.1214/16-STS584

Marchetti S., Giusti C., & Pratesi M. (2016). The use of twitter data to improve small area estimates of households'share of food consumption expenditure in Italy. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *10*(2-3), 79–93. https://doi.org/10.1007/s11943-016-0190-4

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. *Journal of Official Statistics*, *31*(2), 263–281. https://doi.org/10.1515/jos-2015-0017

Marella D. (2023). Adjusting for selection bias in nonprobability samples by empirical likelihood approach. *Journal of Official Statistics*, *39*(2), 151–172. https://doi.org/10.2478/jos-2023-0008

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. *The Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Müller S., Scealy J. L., & Welsh A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*(2), 135–167. https://doi.org/10.1214/12-STS410

OECD (2022). *Financing SMEs and entrepreneurs 2022*. OECD.

Policy Department for Economic, Scientific and Quality of Life Policies (2020). Social sustainability concepts and benchmarks (Technical report). European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648782/IPOL_STU(2020)648782_EN.pdf.

Porter A. T., Holan S. H., Wikle C. K., & Cressie N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics*, *10*, 27–42. https://doi.org/10.1016/j.spasta.2014.07.001

Pratesi M. (2016). *Analysis of poverty data by small area estimation*. John Wiley & Sons.

Rabe-Hesketh S., & Skrondal A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *169*(4), 805–827. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Rao J. (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., & Molina I. (2015). *Small area estimation*. John Wiley & Sons.

Russell S. J., & Norvig P. (2020). *Artificial intelligence a modern approach*. Pearson Education Limited.

Schirripa Spagnolo F., Salvati N., D'Agostino A., & Nicaise I. (2020). The use of sampling weights in *M*-quantile random-effects regression: An application to programme for international student assessment mathematics scores. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, *69*(4), 991–1012. https://doi.org/10.1111/rssc.12418

Schmid T., Bruckschen F., Salvati N., & Zbiranski T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in senegal. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *180*(4), 1163–1190. https://doi.org/10.1111/rssa.12305

Si Y. (2023). 'On the use of auxiliary variables in multilevel regression and poststratification', arXiv, arXiv:2011.00360v4, https://doi.org/10.48550/arXiv.2011.00360, preprint: not peer reviewed.

Statistics Canada (2010). Survey of household spending 2006: Data quality indicators. https://www150.statcan.gc.ca/n1/en/pub/62f0026m/62f0026m2010003-eng.pdf?st=HSMcu0Tt [Accessed 26 November 2023].

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

UNECE (2022). *Machine learning for official statistics*. United Nations.

Valliant R., Dorfman A. H., & Royall R. M. (2000). *Finite population sampling and inference: A prediction approach*. John Wiley. Number 04; QA276. 6, V3

van den Brakel J. A., & Buelens B. (2014). Covariate selection for small area estimation in repeated sample surveys. *Statistics in Transition New Series*, *16*(4), 523–540. https://doi.org/10.59170/stattrans

Yang S., & Kim J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, *3*(2), 625–650. https://doi.org/10.1007/s42081-020-00093-w

Supplementary Material on Inference for big data assisted by small area methods: an application on SDGs sensitivity of enterprises in Italy

S1 Validation of the proposed method by real data application

In this section, we validate the reliability of the Small Area Double Robust Estimator (SAE DR) through an application on real data, where we have access to a nonprobability sample, B, and a probability sample, A. In this context, we assumed that the same set of auxiliary variables, $\mathbf{x}_{ij}$, and the target variable, y_{ij}, were collected for both A and B, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. It is worth noting that, in the application in the main document, the target variable was not surveyed in sample A.

Our goal was to estimate the prevalence of enterprises engaged in e-commerce activities at the provincial level (NUTS 3) in Italy for a reference year. For the probability sample A, we used the "ISTAT Special Survey on Enterprises' Perspectives After the Covid-19 Emergency", as in the application in the paper. In this case the target variable "presence of e-commerce" was surveyed, along with a set of auxiliary variables. We also constructed a nonprobability sample B, where the target variable "presence of e-commerce" was obtained from web-scraping the enterprises' websites with valid URLs available from the ASIA register, which also provided the same auxiliary variables as those in A. Further details can be found in the paper.

In this setting, we applied the SAE DR estimator, masking the target variable from sample A (i.e., assuming that y_{ij} is only available in sample B), and compared the naive direct estimates (obtained directly from B) and the SAE DR estimates with the direct estimates from sample A, which are theoretically unbiased and are considered the *gold standard*.

Using the data scraped from sample B, we designed a binary classifier to identify whether an enterprise engages in e-commerce (i.e., sells goods or services through its website). This process is similar to the sustainability application presented in the paper. We compared different binary classifiers using training and test data (split 70% and 30%, respectively) from the enterprises in

B for which web scraping was successful. The optimal predictive model was the random forest with the following performance on the test set: accuracy=0.87, precision=0.80, recall=0.90, specificity=0.80, f1=0.88[1]. We then compared the the e-commerce prevalence estimates obtained from this binary classification in the big data with the direct estimates from the survey, which serve as the *gold standard* from an unbiased estimator.

The first step in estimating population proportions was to fit the propensity scores using general linear mixed model, as described in the paper. We fitted this model by including, as auxiliary variables, the turnover volume indicator (re-coded in two classes). The second step was to fit the outcome regression model by including the following auxiliary variables: (i) the number of employees of the enterprise averaged over the years, (ii) the turnover volume indicator in classes, and (iii) the NACE code.

Figure S1 shows the map of the estimated proportion of enterprises performing e-commerce across Italian provinces, using the SAE DR estimator.

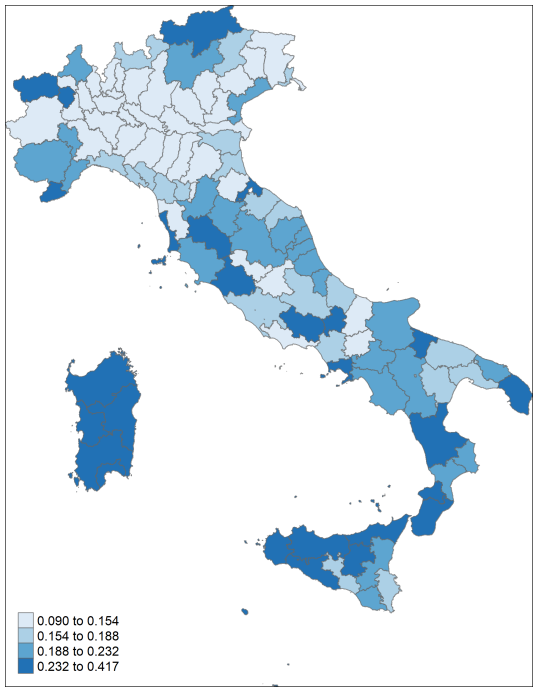

Figure S1: Map of the estimated proportion of enterprises performing e-commerce for the Italian provinces using SAE DR estimator.

[1]Consider a confusion matrix and let TP, TN, FN, and FP be the true positive, true negative, false negative, and false positive, respectively. Then, accuracy = $(TP+TN)/(TP+TN+FN+FP)$, precision = $TP/(TP+FP)$, recall i= $TP/(TP+FN)$, specificity = $TN/(TN+FP)$, and f1 is the harmonic mean between precision and recall.

Since the target variable was also available in the probability sample A, we used the direct estimator applied to that sample as the gold standard to evaluate the SAE DR estimator's performance.

We did not use small area estimates for comparison, as SAE methods generally yield point estimates similar to direct estimates but with reduced variability. Therefore, even if a shrinkage effect on point estimates exists, we decided to compare our results only with the unbiased direct estimator (from the probability sample A), so we did not need to obtain population area means for the auxiliary variables, which are needed to apply SAE industry unit-level methods, thereby prolonging the real data validation. Importantly, the unbiasedness of direct estimates, even with small sample sizes, is essential in this validation. We also estimated the provincial-level proportion of e-commerce enterprises from sample B using a naive direct estimator ($\tilde{\theta}_{B_i}$).

To evaluate the performance of the naive direct estimator from the nonprobability sample B (naive direct) and the SAE DR estimator, we computed their relative bias (RB) by using the direct estimates from A ($\hat{\theta}_{A_i}$) as the reference estimator:

$$\mathrm{RB}(\tau_i) = \frac{\tau_i - \hat{\theta}_{A_i}}{\hat{\theta}_{A_i}} \times 100,$$

where τ_i denotes the considered estimator, SAE DR ($\hat{\theta}_{i,DR}$) or naive direct ($\tilde{\theta}_{B_i}$).

The median RB of the SAE DR estimator was 9.57%, whereas that of the naive direct estimator was 13.67%. Therefore, our method provided less biased estimates.

To further compare the naive direct estimates from sample B, SAE DR estimates, and direct (gold standard) estimates from sample A, we calculated how often the direct estimates from sample A fell within the 95% confidence intervals (CIs) of the SAE DR and naive direct estimators. The results are presented in Figure S2.

The proportions of the unbiased direct estimates that fell within 95% CIs of the naive direct estimator and the 95% CIs of the SAE DR estimator were 39% and 66%, respectively. As shown in Figure S2, the SAE DR estimator outperformed the naive direct estimator when the bias (the absolute difference between the direct estimator from A, gold standard, and the naive direct estimator from B) was large.

When the bias between the gold standard and naive direct estimator from sample B was small (as seen on the left side of Figure S2), there was no noticeable difference between the SAE DR and naive direct CIs. Indeed, as the bias increased, the SAE DR CIs differed from the naive direct CIs and were closer to the direct CIs. We also evaluated whether this pattern held for areas the CVs of the SAE DR and naive direct estimates were below 16.6%, i.e., the areas whose estimates are considered reliable for both estimators. For these areas (74 out of 107), the proportions of the unbiased direct estimates inside the 95% CIs of the naive direct and SAE DR estimator were, respectively, 25% and 50%, indicating a significant difference.

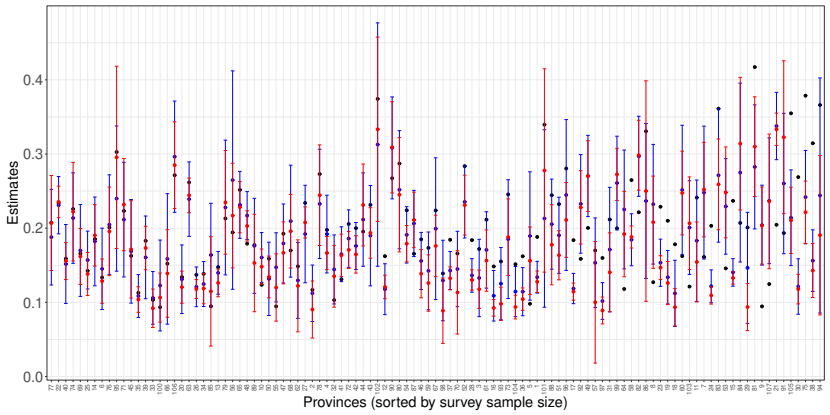

Figure S2: Plot of the direct (gold standard) estimates (black), naive direct estimates with their 95% CIs (red), and SAE DR estimates with their 95% CIs (blue). Areas sorted by absolute difference between the direct estimator from A (gold standard) and the naive direct estimator from B in ascending order.

Based on the results obtained with real-case data compared with the *golden standard* direct estimates obtained on A, we conclude that the proposed SAE DR estimator can, at least partially, correct for the bias caused by the self-selection of the nonprobability sample B.

Table S1 lists the CVs of the SAE DR and naive direct estimators. The standard error (SE) of the SAE DR estimator was estimated using the proposed bootstrap approach based on $L = 500$ bootstrap samples. The SE of the naive direct estimator was obtained via analytical approximation using the function `direct` of the `sae` package in `R` (Molina and Marhuenda, 2015). Notably, the classification of the reliability of the estimates was performed according to Statistics Canada rules (estimates with CV $< 16.6\%$ are reliable for general use, with $16.6\% <$ CV 33.3% should be accompanied by warnings to users, with CV $> 33.3\%$ are considered unreliable (Statistics Canada, 2010)). While the SAE DR estimator lost precision compared to the naive direct estimator, this was expected in a bias correction context. Nonetheless, the loss in efficiency was minimal, and no provinces had CVs larger than 33.3

Table S1: Number of provinces by CV for naive direct and SAE DR estimators on B.

Est.	$< 16.6\%$	between 16.6% and 33.3%	$> 33.3\%$
Naive direct	88	18	1
SAE DR	57	50	0

S2 Additional simulation scenarios

To evaluate the performance of the proposed estimator under a sampling design with unequal inclusion probabilities, we replicated the scenarios 1-4 presented in Section 5 of the paper. In these scenarios, samples inside the areas were selected according to Poisson sampling with inclusion probabilities proportional to $x_{2;ij}$, $\pi_{ij} = 0.02x_{2;ij} + 0.05$ (Fabrizi et al., 2014).

The results, presented in Table S2, show that across all scenarios, the naive direct estimator exhibited larger bias, absolute bias, and relative bias compared to the SAE DR estimator. Bias reduction was particularly noticeable in Scenarios 1 and 2 compared to Scenarios 3 and 4. Nonetheless, in all cases, even with a different sampling scheme, our proposed SAE DR estimator achieved a substantial bias reduction. These findings are consistent with the simulations conducted under simple random sampling within areas.

Table S2: Median over the areas of Bias, AbsBias, RB, and MSE over 500 MC simulations for each scenario - Poisson sampling.

Estimator	*Bias*	*AbsBias*	*RB (%)*	*MSE*
Scenario 1				
Naive direct	0.051	0.051	7.737	0.003
SAE DR	0.008	0.008	1.370	0.006
Scenario 2				
Naive direct	-0.036	0.036	-5.193	0.004
SAE DR	0.009	0.009	1.291	0.006
Scenario 3				
Naive direct	0.032	0.032	3.837	0.001
SAE DR	0.005	0.005	0.621	0.004
Scenario 4				
Naive direct	-0.021	0.021	-2.507	0.001
SAE DR	0.006	0.006	0.702	0.004

To estimate the variance of the DR SAE estimator in this context, we tested an alternative bootstrap procedure inspired by the work of Antal and Tillé (2011). This approach can be employed to estimate the variance in cases of unequal inclusion probabilities. Specifically, we implemented a two-step bootstrap approach that accounts explicitly for these unequal probabilities. In the first step, to incorporate the survey sampling design in resampling from A, we used the general bootstrap procedure proposed by Antal and Tillé (2011). In this framework, part of the subsampled units was selected without replacement,

while another portion was selected with replacement, adjusting for the finite population setting. This bootstrap method is versatile and can be adapted to various survey sampling designswhether with or without replacement, and whether with equal or unequal inclusion probabilities. The proposed approach eliminated the need for scaling or weighting the sample and ensured that the bootstrap and original samples had the same expected size.

In the second step, we sampled from the non-probability sample B using simple random sampling, assuming an unknown response propensity in the population. The steps of this bootstrap procedure are as follows:

1. xtract a sample of size n_A from the probabilistic survey A using the Antal and Tillé (2011) approach, following the sampling design of A to obtain a bootstrap replicate denoted by $\{(\delta_j^\star, w_j^\star, \mathbf{x}_j^\star)\} \in A^\star$. Here $\mathbf{w}^\star = \{w_j^\star\}$ is the set of replication weights based on the sampling design for the probability sample A. Then $\{(\delta_{ij}^\star, w_{ij}^\star, \mathbf{x}_{ij}^\star)\} \in A_i^\star$ with $i = 1, \ldots, m$ is the subsample of $A^\star$ in the small area i.

2. Extract a simple random sample with replacement of size n_B from sample B, obtaining the bootstrap sample $(y_j^\star, \mathbf{x}_j^\star) \in B^\star$. In this case, $(y_{ij}^\star, \mathbf{x}_{ij}^\star) \in B_i^\star$, where $i = 1, \ldots, m$ denotes the subsample of $B^\star$ in the small area i.

3. Use the model in Equation (5) applied to the bootstrap sample to obtain the bootstrap propensity score $\hat{p}_{ij}^\star(\mathbf{x}_{ij}, \hat{\boldsymbol{\lambda}}^\star, \hat{u}_i^\star)$.

4. Fit the model in Equation (7) on the bootstrap sample $B^\star$ to estimate the regression coefficients $\hat{\boldsymbol{\beta}}^\star$ and area-specific random effects $\hat{\gamma}_i^\star$.

5. Use Equation (8) to obtain the DR estimator $\hat{\theta}_{i;DR}^\star$ using the weights $w_{ij}^\star$ obtained in 1.

6. Repeat steps 1–5, independently for L times. The resulting bootstrap variance estimator of $\hat{\theta}_{i;DR}$ is computed as in Equation (9).

It should be noted that the first step of this procedure must be adapted to the specific problem under consideration, depending on the sampling design of survey A.

In Table S3, we reported the median of the area-specific bias expressed in relative terms (%) of the bootstrap SE estimator, obtained with 200 bootstrap replicates.

The proposed bootstrap variance estimator tended to to slightly underestimate the variance. Specifically, in all scenarios we observed a negative RB, with the underestimation being more pronounced in Scenario 4, where the RB was about -13%.

In Table S4, we also present the empirical coverage rates (CRs) and the semilengths for nominal 95% CIs based on normal-like CIs. As expected, due to the negative bias of the boostrap variance estimator, actual CRs tended to be lower than the nominal level. Therefore, further research on the use of bootstrap techniques for constructing confidence intervals is needed and is left for future exploration.

Table S3: Median over the areas of RB of the bootstrap variance estimators over the 500 MC simulations for each scenario - Poisson sampling.

Scenario	*RB (%)*
Scenario 1	-1.471
Scenario 2	-5.898
Scenario 3	-1.456
Scenario 4	-13.196

Table S4: Quartiles over the areas of empirical CRs and the median of semilengths of CIs - Poisson sampling.

Scenario	Coverage			Semi-Length
	Q1	Median	Q3	
Scenario 1	79.6	86.8	91.3	0.14
Scenario 2	79.1	85.3	90.1	0.14
Scenario 3	75.3	82.0	88.5	0.10
Scenario 4	70.2	79.6	85.8	0.09

References

Antal, E. and Y. Tillé (2011). A direct bootstrap method for complex sampling designs from a finite population. Journal of the American Statistical Association 106(494), 534–543.

Fabrizi, E., N. Salvati, M. Pratesi, and N. Tzavidis (2014). Outlier robust model-assisted small area estimation. Biometrical Journal 56(1), 157–175.

Molina, I. and Y. Marhuenda (2015, jun). sae: An R package for small area estimation. The R Journal 7(1), 81–98.

Statistics Canada (2010). Survey of Household Spending 2006: Data Quality Indicators. `https://www150.statcan.gc.ca/n1/en/pub/62f0026m/62f0026m2010003-eng.pdf?st=HSMcu0Tt`. [Online. Accessed 26 November 2023.].