Statistical Methods & Applications
https://doi.org/10.1007/s10260-023-00740-y

ORIGINAL PAPER

# Integrating probability and big non-probability samples data to produce Official Statistics

Natalia Golini[1] · Paolo Righi[2]

Accepted: 6 December 2023

## Abstract

This paper introduces the pseudo-calibration estimators, a novel method that integrates a non-probability sample of big size with a probability sample, assuming both samples contain relevant information for estimating the population parameter. The proposed estimators share a structural similarity with the adjusted projection estimators and the difference estimators but they adopt a different inferential approach and informative setup. The pseudo-calibration estimators can be employed when the target variable is observed in the probability sample and, in the non-probability sample, it is observed correctly, observed with error, or predicted. This paper also introduces an original application of the jackknife-type method for variance estimation. A simulation study shows that the proposed estimators are robust and efficient compared to the regression data integration estimators that use the same informative setup. Finally, a further evaluation using real data is carried out.

**Keywords** Big data · Calibration weighting · Data integration · Missing at random · Model-based inference · Variance estimation

## 1 Introduction

In recent years, new data sources have emerged as a result of increased interactions with digital technologies by both citizens and business units, along with the growing capability of these technologies to generate digital trails. These sources, known as Big Data (BD) sources, encompass extensive amounts of digital information,

✉ Natalia Golini
natalia.golini@unito.it

Paolo Righi
parighi@istat.it

[1] Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Lungo Dora Siena 100 A, 10153 Turin, Italy

[2] Italian National Statistical Institute (Istat), Via Cesare Balbo, 16, 00184 Rome, Italy

Published online: 18 January 2024

including web surveys, search queries, website visits, social media activity, online purchases, self-reported administrative data sets, and other online interactions. BD sources typically comprise numerous records, often containing unstructured information, and are primarily generated for non-statistical purposes. They represent non-probability samples of the reference population. In many cases, they do not accurately represent the population of interest. Consequently, using them, for instance, to compute a simple mean of the observed values can lead to biased population mean estimates and erroneous conclusions, despite the large sample size (Bethlehem 2010; Vehovar et al. 2016; Meng 2018). Notwithstanding these limitations, BD sources offer quick, easy, and cost-effective alternatives for obtaining data. They are becoming increasingly relevant in research and, notably, they present challenging sources of information for producing Official Statistics.

The use of BD sources is leading to a paradigm shift for National Statistical Institutes (NSIs), transitioning from planned statistics achieved through a designed process to data-oriented or data-driven statistics. Traditionally, NSIs rely on a designed process for collecting statistical data. This involves identifying the target population and its records, defining the target variables, planning the sampling design, and using efficient estimators. In the data-driven approach, the primary focus is on choosing the estimator that is most suitable for the task based on the observed variables. The process involves using a specific data collection tool, usually a digital device, on a sub-population selected through an unknown sampling technique. Horrigan (2013) emphasizes the importance of creating transparent methodological documentation (metadata) describing how BD are used to construct any type of estimate. Citro (2014), Tam and Clarke (2015a), Pfeffermann (2015) address the methodological uses and challenges of BD sources in the production of Official Statistics. Many reports have developed suitable statistical frameworks (among others: EUROSTAT 2018; Japec et al. 2015) and quality frameworks (UNECE Big Data Quality Task Team 2014; United Nations 2019; EUROSTAT 2020) that outline the fundamental principles and guidelines for using BD sources in producing Official Statistics. Several papers focusing on the accuracy and reliability of BD sources emphasize the growing need to determine the conditions under which BD sources can provide valid inferences. In this regard, many authors agree with the necessity of using methods combining data from big non-probability and probability samples to not severely sacrifice the quality of the estimates (Beaumont 2020). Valliant (2020) and Rao (2021) provide insightful reviews of these methods. Kim (2022) offers an extensive review of data integration techniques for combining a probability sample with a non-probability sample when the study variable is only observed in the non-probability sample. Most methods assume that the variable of interest is available only in the non-probability sample, while other auxiliary variables are present in both samples.

In this work, we assume that the target variable is observed in the probability sample, while in the big non-probability sample, it is (a) observed correctly, (b) observed with error, or (c) predicted using covariates collected in the big non-probability sample. A real case study inspiring our research is the 2018 European Community survey data on ICT usage and e-commerce in enterprises, conducted annually by Istat. The ICT probability survey sample data can be combined with the internet data scraped from the enterprises' websites belonging to the ICT target population (big non-probability sample data). The target variables related to e-commerce functionalities, social media links, and presence of job advertisements can be observed, according to assumption (a), or predicted, according to assumption (c), using text-mining techniques on the scraped website data (Righi et al. 2019). By integrating this additional information with the ICT survey, one can significantly improve the accuracy of the estimates. Another real case illustrating the type of BD we consider in this paper is given in Tam (2015) and Tam and Clarke (2015b). In these papers, the use of remote sensing for agricultural statistics using geo-localized satellite imagery and other satellite data (e.g., moisture, temperature) is investigated. After transforming the images into structured data (for instance, the reflectance data from frequency bands), the target variables (land use, crop type, crop yield) are predicted by supervised machine learning classification techniques. A probability sample of geo-localized areas collecting ground truth data is used as a training set. Another example is given by Rueda et al. (2023) where an application of data integration techniques using a similar informative setup is provided. They consider a probability survey on the impact of the COVID-19 pandemic in Spain combined with a non-probability web-based survey. Both samples share the same questionnaire and measures.

In this paper, we introduce a novel class of estimators called pseudo-calibration (PC) estimators. They are based on big non-probability sample data, integrated with probability survey sample data and administrative or statistical registers. We also propose a variance estimation method based on the Delete-a-Group Jackknife technique (Kott 2001, 2006a). Specifically, we formalize the PC estimators initially introduced in an Istat technical report[1] and employed in Righi et al. (2019). The PC estimators are developed within a model-based framework, although an automatic calibration procedure, typical of model-assisted estimators, is carried out. We highlight that the proposed estimators have a similar structure to the *adjusted projection estimator* (Kim and Rao 2011) and the *difference estimators* (Breidt and Opsomer 2017), but a different inferential approach and informative setup. Furthermore, we show the analogies of the proposed estimators with the *doubly robust estimators* (Chen et al. 2020). Yet, we compare the proposed estimators with the *data integration estimators* proposed by Kim (2022), developed in the same informative setup. The data integration estimators utilize both a probability and non-probability sample from the reference population. The target variable is observed in both samples, but there is a possibility of inaccurate measurement in one of the samples. The PC and data integration estimators employ calibration techniques, which are well-established methods used by National Statistical Institutes (NSIs), making them suitable for producing Official Statistics. However, the calibration methods differ significantly between these two classes of estimators. Precisely, the PC estimators aim to compute the weights of units in the non-probability sample; the data integration estimators seek to compute the weights of the probability sample units according to a model-assisted approach. With few exceptions, the two classes of estimators produce different estimates of the target parameter.

[1] https://www.istat.it/it/files//2020/05/Tech_Report_ICT2018.pdf

The paper is structured as follows. Section 2 introduces the basic notation and the informative context. A brief introduction of the data integration estimators (Kim and Tam 2021) is in Sect. 3. Section 4 illustrates the novel class of PC estimators, and Sect. 5 shows the jackknife-type variance estimator. Section 6 presents the results of a Monte Carlo simulation on the performance of PC estimators, the comparison with the data integration estimators, and the accuracy of the jackknife-type variance estimator. Section 7 shows an application of the two classes of estimators on the motivating real survey data and BD source introduced above. Finally, some concluding remarks are in Sect. 8.

This paper is an extended version of the paper presented at the 51st Scientific Meeting of the Italian Statistical Society on June 2022 (Righi et al. 2022).

## 2 Informative context

We estimate the parameters of the finite target population using a big non-probability sample, where the values of the target variable may either be observed correctly, observed with error, or predicted. To ensure valid inferences, we assume the following: (i) there exists a reference survey, with a probability sample drawn from the target population, where the target variable is observed correctly; (ii) it is possible to identify which units in the probability sample also belong to the non-probability sample; (iii) in the big non-probability sample, a set of auxiliary variables related to the target variable are available.

Assumptions (i) and (ii) are necessary for implementing the proposed PC estimators when the values of the target variable are observed with error or predicted in the large non-probability sample. Assumption (iii) underlines that the non-probability sample can serve as a source to gather informative covariates for predicting the target variable.

### 2.1 Notation and basic setup

Let $\mathcal{U} = \{1, \dots, N\}$ denote the target population of size $N$, let $Y = \Sigma_{i=1}^{N} y_i$ be the target parameter and $y_i$ the observed value of the variable $\mathcal{Y}$ for the unit $i$. We have two independent samples from the finite population $\mathcal{U}$: a probability sample $\mathcal{S}_A$ of size $n_A$ and a big non-probability sample $\mathcal{S}_B$ of size $n_B$. For each unit $i \in \mathcal{S}_A$, we observe the values of a vector of auxiliary variables $\mathbf{x}_i$ and the target variable $y_i$. Within a design-based framework, $\hat{Y}_{HT,A} = \Sigma_{i \in A} d_i^A y_i$ stands as the design-unbiased Horvitz-Thompson estimator of $Y$, where $d_i^A = 1/\pi_i^A$ denotes the sampling weight and $\pi_i^A = Pr(i \in \mathcal{S}_A)$ is the first-order inclusion probability in $\mathcal{S}_A$.

In the big non-probability sample $\mathcal{S}_B$, the target variable $\mathcal{Y}$ can be observed correctly, with error, or predicted using a parametric or non-parametric model. In the first case, $y_i$ represents the observed value of $\mathcal{Y}$ for the unit $i$. In the latter two cases, the value of $\mathcal{Y}$ is denoted as $\tilde{y}_i$. We use the notation $y_i^*$ to indicate either $y_i$ or $\tilde{y}_i$.

**Table 1** Data available for $\mathcal{S}_A$ and $\mathcal{S}_B$

| Data | Target variable | Sampling weight | Representative? |
|---|---|---|---|
| $\{i \in \mathcal{S}_A\}$ | $y_i$ | $1/\pi_i^A$ | Yes |
| $\{i \in \mathcal{S}_B\}$ | $y_i^*$ | unknown | No |

We observe a vector of auxiliary variables $\mathbf{x}_i$ for each unit $i \in \mathcal{U}$ and an additional vector of auxiliary variables $\mathbf{x}_{i,B}$ for each unit $i \in \mathcal{S}_B$. When the target variable cannot be observed in $\mathcal{S}_B$, the vector $\mathbf{x}_{i,B}$ contains good predictors for it.

The probability of a unit being included in the big non-probability sample, say $\pi_i^B = Pr(i \in \mathcal{S}_B)$, is unknown. This probability is referred to as the propensity score. Let $\delta_i = I(i \in \mathcal{S}_B)$ be the indicator variable such that, $\delta_i = 1$ if $i \in \mathcal{S}_B$ and $\delta_i = 0$ if $i \notin \mathcal{S}_B$ $(i = 1, \dots, N)$. The propensity scores are given by $\pi_i^B = E_p(\delta_i \mid \mathbf{x}_i, y_i) = Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i)$, where $p$ refers to the model for generating $\mathcal{S}_B$.

Table 1 displays the data set available for the two samples and their representativeness.

As in Kim and Tam (2021) and Chen et al. (2020), we assume that units belonging to $\mathcal{S}_A$ can be recognized in $\mathcal{S}_B$. Therefore, it is possible to specify $\delta_i$ for each unit $i \in \mathcal{S}_A$.

## 3 Data integration estimators

The data integration (DI) estimators, developed by Kim and Tam (2021), provide a versatile tool for properly utilizing big non-probability samples in finite population inference. The big non-probability sample (BD source) is treated as a finite population of incomplete or inaccurate observations that can be used as auxiliary information. Thus, a calibration estimator can be directly used to adjust sampling weights for each $i \in \mathcal{S}_A$, to reproduce certain known population totals for both the target population $\mathcal{U}$ and a non-probability sample $\mathcal{S}_B$. In Kim and Tam (2021), the authors point out that if the fraction of the non-probability sample present in the finite population is not substantial, the efficiency gain achieved by the DI estimators is limited. Additionally, it is worth highlighting that making design-based inference is advantageous for NSIs, as they typically use this approach to produce Official Statistics.

The general form of the class of DI estimators is the Regression DI (RegDI) estimator which is defined as

$$\hat{Y}_{RegDI} = \sum_{i \in \mathcal{S}_A} w_i^A y_i, \tag{1}$$

where $\{w_i^A : i \in \mathcal{S}_A\}$ is the vector of calibrated weights. These weights are determined by solving the following optimization problem

$$\begin{cases} \min \sum_{i\in\mathcal{S}_A} Q(d_i^A, w_i^A)/q_i \\ \sum_{i\in\mathcal{S}_A} w_i^A \mathbf{x}_i = \mathbf{X} \end{cases}, \tag{2}$$

where $d_i^A$ represents the base sampling weight, $q_i$ is a known positive weight independent of $d_i^A$ and $\mathbf{X} = \sum_{i\in\mathcal{U}} \mathbf{x}_i$ is a vector of totals, including the totals of $\delta_i$ and $\delta_i y_i^*$. These totals are assumed to be known or possibly to be estimated by a large and accurate survey (e.g., Dever and Valliant 2010, 2016). The function $Q(\cdot)$ is a distance function that can be defined, for example, as

$$Q(d_i^A, w_i^A; q_i) = \sum_{i\in\mathcal{S}_A} \frac{d_i^A}{q_i}\left(\frac{w_i^A}{d_i^A} - 1\right)^2. \tag{3}$$

It is important to note that, in practice, uniform weighting ($q_i = 1$) is commonly used, although sometimes different weights are employed (Deville and Särndal 1992).

By specifying the terms of the RegDI estimator, one can derive various DI estimators. Kim and Tam (2021) gives insight into the specific estimators. Furthermore, if we use alternative distance functions, we can obtain the calibration data integration estimators according to the definition by Deville and Särndal (1992). Regression and calibration estimators are useful statistical tools for enhancing the precision of the sampling estimates and are commonly used to deal with unit non-response and the frame list under-coverage (Kott 2006b; Särndal and Lundström 2005).

***Remark 1*** A special case of (1) can be obtained considering the distance function (3) and setting $\mathbf{x}_i = \delta_i y_i^*$ and $q_i = \delta_i y_i^*$. We can define the RegDI estimator as

$$\hat{Y}_{RegDI} = \hat{Y}_{HT,A} + \frac{\hat{Y}_{HT,A}}{\hat{Y}^{*(B)}_{HT,A}}\left(Y^{*(B)} - \hat{Y}^{*(B)}_{HT,A}\right) = \frac{\hat{Y}_{HT,A}}{\hat{Y}^{*(B)}_{HT,A}} Y^{*(B)}, \tag{4}$$

where $\hat{Y}_{HT,A} = \sum_{i\in\mathcal{S}_A} d_i^A y_i$, $\hat{Y}^{*(B)}_{HT,A} = \sum_{i\in\mathcal{S}_A} d_i^A \delta_i y_i^*$ and $Y^{*(B)} = \sum_{i\in\mathcal{S}_B} y_i^*$.

***Remark 2*** The RegDI estimators utilize $y_i^*$ as auxiliary information in a design-based approach. They exhibit greater efficiency compared to the Horvitz-Thompson estimator when $y_i^*$ is correlated with the variable $y_i$ observed in $\mathcal{S}_A$, with maximum efficiency achieved when $y_i^* = y_i$. It is worth noting that in large-scale multi-purpose surveys, more than one target variable may be observed or predicted in $\mathcal{S}_B$. Consequently, the RegDI estimators could face excess calibration constraints, potentially making the calibration process unfeasible. Sampling errors may be notably large when these constraints are satisfied.

# 4 Pseudo-calibration estimators

## 4.1 Model-based estimators

In this section, we consider the case (a), i.e., where the target variable is observed in both samples.

Unlike the design-based approach, the model-based approach utilizes data from a big non-probability sample and directly estimates the finite population parameter. This is achieved by summing the observed target variable for $i \in \mathcal{S}_B$ and the target variable predicted for $i \notin \mathcal{S}_B$. In this case, inference can be made within a model-based framework. Prediction methods rely on defining a super-population model that generates the target variable $\mathcal{Y}$ (Valliant et al. 2000). Let's suppose that the finite population $(\mathbf{x}_i, y_i)$, for all $i \in \mathcal{U}$, can be viewed as a random sample from the model $y_i = \mu(\mathbf{x}_i) + \epsilon_i$, where $\mu(\cdot)$ can take a parametric or an unspecified non-parametric form, and $\epsilon_i$ is an independent variable with zero mean and variance $V(\epsilon_i) = v(\mathbf{x}_i)\sigma^2$, with the form of the variance function $v(\cdot)$ being known. This outcome model describes the dependence of the target variable on a vector of auxiliary variables $\mathbf{x}$. We can use this relationship to predict the values of the units not belonging to $\mathcal{S}_B$, provided that the $\mathbf{x}$ values are known for all $i \in \mathcal{U}$. In practice, we utilize the dataset of the pooled sample $\{(\mathbf{x}_i, y_i), i \in \mathcal{S}_B \cup \mathcal{S}_A\}$ to construct the outcome model and make predictions. Given a parametric outcome model, $y_i = \mu(\mathbf{x}_i;\boldsymbol{\beta}) + \epsilon_i$, and a consistent estimator $\hat{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta}$, we obtain the predictions as $\hat{y}_i = \mu(\mathbf{x}_i;\hat{\boldsymbol{\beta}})$. The estimator for the population total $Y$ is defined as $\hat{Y}_m = \sum_{i\in\mathcal{S}_B} y_i + \sum_{i\notin\mathcal{S}_B} \hat{y}_i$. If the assumption $E_\mu(y_i \mid \mathbf{x}_i, \delta_i) = E_\mu(y_i \mid \mathbf{x}_i)$ holds, we obtain unbiased model-based estimates. Finally, the model-based estimator can be defined as a weighted sum of the observed values, $\hat{Y}_m = \sum_{i\in\mathcal{S}_B} \omega_i y_i$, where $\omega_i$ are the appropriate weights representing the units not belonging to $\mathcal{S}_B$ (Valliant et al. 2000).

Another class of model-based estimators involves estimating the propensity scores. Since the selection mechanism of $\mathcal{S}_B$ is unknown, $\pi_i^B$ is estimated by a propensity score model exploting the dataset $\{(\delta_i, \delta_i y_i, \mathbf{x}_i), i \in \mathcal{V}\}$, where $\mathcal{V}$ is either $\mathcal{U}$ or $\mathcal{S}_B$. For example, let the propensity score model be parametric, $\pi_i^B = \pi(\mathbf{x}_i, y_i, \boldsymbol{\theta})$, and let $\hat{\boldsymbol{\theta}}$ be a consistent estimator of $\boldsymbol{\theta}$. The estimate of $\pi_i^B$ is then $\hat{\pi}_i^B = \pi(\mathbf{x}_i, y_i, \hat{\boldsymbol{\theta}})$. Once estimated $\pi_i^B$, the model-based estimator is given by $\hat{Y}_\pi = \sum_{i\in\mathcal{S}_B} y_i/\hat{\pi}_i^B$. In practice, $\boldsymbol{\theta}$ cannot be estimated when the model depends on the $y_i$ values since they are not observed for $i \notin \mathcal{S}_B$. Given the assumptions

A.1: the selection indicator $\delta_i$ and the target variable $y_i$ are independent given the vector of covariates $\mathbf{x}_i$;

A.2: $\pi_i^B > 0$ for all $i \in \mathcal{U}$;

A.3: the variables $\delta_i$ and $\delta_j$ are independent given $\mathbf{x}_i$ and $\mathbf{x}_j$ for $i \neq j$ with $i,j \in \mathcal{U}$,

then, by Chen et al. (2020), $\pi_i^B = Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i) = Pr(\delta_i = 1 \mid \mathbf{x}_i)$. This model corresponds to the Missing At Random mechanism (MAR) as defined by Rubin (1976) and Little and Rubin (2019). The MAR model parameters can be estimated using the dataset $\{(\delta_i, \mathbf{x}_i), i \in \mathcal{V}\}$. For example, one may opt for a logistic propensity score model and employ a maximum likelihood consistent estimator when $\mathcal{V} \equiv \mathcal{U}$. However, if $\mathcal{V} \equiv \mathcal{S}_B$, the (log)likelihood function cannot be completely computed. The method relies on the reference survey sample, collecting the $\mathbf{x}$ values for $i \in \mathcal{S}_A$. Afterwards, a pseudo-likelihood function can be defined, and the maximum pseudo-likelihood estimates of $\boldsymbol{\theta}$ can be computed (for details, refer to formula (4) in Chen et al. (2020)). Given the propensity score estimates, the inverse probability weighted estimator can be estimated as $\hat{Y}_{IPW} = \sum_{i \in \mathcal{S}_B} y_i / \hat{\pi}_i^B$ being $\hat{\pi}_i^B = \pi(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ (Kott 1994; Elliot and Valliant 2017). Chen et al. (2020) show that, assuming the logistic regression model for the propensity scores, under the regularity conditions A1–A3 and other reasonable conditions (C1-C6 specified in the supplementary materials), then $\hat{Y}_{IPW} - Y = O_p(n_B^{-1/2})$.

### 4.2 Pseudo-calibration estimators when the target variable is observed in $\mathcal{S}_B$

In the case (a), we derive the PC estimators from the inverse probability weighted estimator. In this case, the maximum pseudo-likelihood estimator is replaced by a consistent estimator based on unbiased estimating functions. Consider the following class of estimating equations

$$\sum_{i \in \mathcal{U}} \delta_i \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) - \sum_{i \in \mathcal{U}} \pi(\mathbf{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{5}$$

where $\boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta})$ is a predefined smooth function of $\boldsymbol{\theta}$ that ensures the system (5) has a unique solution.

When $\mathbf{x}_i$ is known for each $i \in \mathcal{U}$ and $\boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) = \pi(\mathbf{x}_i, \boldsymbol{\theta})^{-1} \mathbf{x}_i$, the system (5) becomes the conventional calibration equations

$$\sum_{i \in \mathcal{S}_B} w_i^B \mathbf{x}_i = \sum_{i \in \mathcal{U}} \mathbf{x}_i, \tag{6}$$

where $w_i^B = 1/\pi(\mathbf{x}_i, \boldsymbol{\theta})$.

When $\mathbf{x}_i$ can only be observed for the units belonging to $\mathcal{S}_A$, Chen et al. (2020) propose to replace $\sum_{i \in \mathcal{U}} \pi(\mathbf{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta})$ with $\sum_{i \in \mathcal{S}_A} d_i^A \pi(\mathbf{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta})$ in (5), obtaining the class of estimating equations,

$$\sum_{i \in \mathcal{S}_B} \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) - \sum_{i \in \mathcal{S}_A} d_i^A \pi(\mathbf{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{7}$$

When $\boldsymbol{h}(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i$, the system (7) simplifies to

$$\sum_{i \in \mathcal{S}_B} w_i^B \mathbf{x}_i = \sum_{i \in \mathcal{S}_A} d_i^A \mathbf{x}_i, \tag{8}$$

where the calibration is based on the estimated totals from the reference survey.

We can obtain the solution to (6) or (8) through the standard calibration process (Deville and Särndal 1992). The weights $w_i^B$ are determined by solving the optimization problem

$$\begin{cases} \min \sum_{i\in\mathcal{S}_B} Q(d_i^B, w_i^B; q_i) \\ \sum_{i\in\mathcal{S}_B} w_i^B \mathbf{x}_i = \mathbf{X}^* \end{cases}, \tag{9}$$

where $Q(\cdot)$ is a convex distance function, which may take the same form as illustrated in (3), replacing $d_i^A$ and $w_i^A$ by $d_i^B$ and $w_i^B$, respectively. Additionally, the summation is indexed for $i \in \mathcal{S}_B$. Here, $d_i^B$ represent the base sampling weights, $w_i^B$ are the calibration weights, and $\mathbf{X}^*$ is a vector of known totals, denoted as $\mathbf{X}$, or estimated totals, denoted as $\hat{\boldsymbol{X}}$, derived from an accurate reference survey (Dever and Valliant 2010, 2016).

At first glance, the PC estimators may appear to be a slight variation of the DI indicators proposed by Kim and Tam (2021). However, there are key distinctions: the PC estimators operate on the propensity score of the non-probability sample, whereas the DI estimators work on the weights of the probability sample; the inference for the PC estimators is based on the outcome and a propensity score model, while the inference for the DI estimators is based on a model-assisted approach; the calibration constraints of the PC estimators do not use the target variable(s), while the calibration constraints of the DI estimators are strictly target variable-dependent. Since $\mathcal{S}_B$ is not a probability sample, the propensity scores $\pi_i^B$ and the base sampling weights $d_i^B = 1/\pi_i^B$ for the units in $\mathcal{S}_B$ are unknown. Nevertheless, we make two alternative assumptions. The first is that we plan a census, but the frame list of $\mathcal{S}_B$ under-covers the target population $\mathcal{U}$. Then, $\pi_i^B = 1$ for all $i \in \mathcal{S}_B$ and the base sampling weights $d_i^B$ are adjusted to account for the under-coverage bias through a calibration estimator (Little and Rubin 2019; Kott 2006b). The second assumption is that in the absence of information about the process generating $\mathcal{S}_B$, the maximum likelihood estimate of $\pi_i^B$ is $n_B/N$ for all $i \in \mathcal{S}_B$, and $d_i^B = d^B$ with $d^B = N/n_B$. After observing the sample and the auxiliary variables within it, we improve the estimates of $\pi_i^B$. In the space of possible final weight vectors, we look for the vector closest to the initial value of $d^B$, reducing the variability of $w_i^B$ as much as possible. Furthermore, Remark 7 shows that by setting $d^B = N/n_B$ for all $i \in \mathcal{S}_B$, the RegDI estimator in (4) can be expressed as a special case of PC estimator. Eventually, the two proposed guesses provide the same vector of calibrated weights and, in general, we achieve the same solution when using $d_i^B = d^B$ regardless of the value of $d^B$.

The general expression of PC estimators is given by

$$\hat{Y}_{PC} = \sum_{i\in\mathcal{S}_B} w_i^B y_i. \tag{10}$$

The PC estimators ensure that the weighted distribution of the non-probability sample across auxiliary variables aligns with the distribution of those variables in the target population. They offer a simple and direct implementation and utilize well-established and widely used statistical calibration tools in NSIs.

***Remark 3*** (Justification of the optimization problem). We solve the calibration equations, $\sum_{i\in\mathcal{S}_B} w_i^B \mathbf{x}_i = \mathbf{X}^*$, by setting up the optimization problem (9). In the special case of $d_i^B = 1$, we encounter a frame list under-coverage problem for $\mathcal{S}_B$. Solving the optimization problem given in (9) is a commonly used strategy to address this issue. The basic idea is to limit the variability of $w_i^B$ and, through the choice of specific distance functions, prevent the occurrence of very large or negative values of $w_i^B$. With $d_i^B = N/n_B$, the optimization starts from the simple propensity score mean model, which does not incorporate any auxiliary variables. Then, we enhance the model by introducing explanatory auxiliary variables, aiming to find the model closest to the parsimonious mean model, obtaining the final weights.

***Remark 4*** The pseudo-calibrated estimate converges to the true value of the target parameter as $n_B \to N$, but its accuracy is sensitive to potential failures of the propensity score model, such as the violation of the MAR assumption, especially when dealing with small sample sizes. This is particularly notable in the case of sub-population estimates. A possible approach to enhancing the robustness of the PC estimators is to integrate a prediction model for the target variable and use a doubly robust estimator. When $\mathbf{x}_i$ is known for each $i \in \mathcal{U}$, we have

$$\hat{Y}_{DR1} = \sum_{i\in\mathcal{S}_B} w_i^B (y_i - \hat{y}_i) + \sum_{i\in\mathcal{U}} \hat{y}_i. \tag{11}$$

When $\mathbf{x}_i$ is known for $i \in \mathcal{S}_B$ or $i \in \mathcal{S}_A$, we have

$$\hat{Y}_{DR2} = \sum_{i\in\mathcal{S}_B} w_i^B (y_i - \hat{y}_i) + \sum_{i\in\mathcal{S}_A} d_i^A \hat{y}_i. \tag{12}$$

Chen et al. (2020) show the theoretical properties of (11) and (12).

***Remark 5*** We can test the MAR assumption of the propensity score model (MAR model) using the dataset $\{(\delta_i, y_i, \mathbf{x}_i, ), i \in \mathcal{S}_A\}$.

***Remark 6*** The $\hat{Y}_{PC}$ estimator, with $\mathbf{x}_i = x_i$ (where $x_i$ is a scalar), a distance function in the form given in (3), and $q_i = x_i$, can be expressed as

$$\hat{Y}_{PC} = \sum_{i\in\mathcal{S}_B} y_i d_i^B \left( \frac{X}{\hat{X}_B} \right), \tag{13}$$

where $X = \sum_{i\in\mathcal{U}} x_i$ and $\hat{X}_B = \sum_{i\in\mathcal{S}_B} x_i d_i^B$. $\hat{Y}_{PC}$ in (13) is the PC ratio estimator. Note we only use the $x_i$ values for $i \in \mathcal{S}_B$. Moreover, we can obtain a second version of the PC ratio estimator by replacing $X$ with $\hat{X}$.

***Remark 7*** In the case (a), the RegDI estimator in (4) can be reformulated as

$$\hat{Y}_{RegDI} = \frac{N}{n_B} Y^{(B)} \left( \frac{n_B \hat{Y}_{HT,A}}{N \hat{Y}^{(B)}_{HT,A}} \right) = \sum_{i\in\mathcal{S}_B} y_i \frac{N}{n_B} \left( \frac{\hat{Y}_{HT,A}}{\hat{Y}_{HT,AB}} \right), \tag{14}$$

where $Y^{(B)} = \sum_{i \in \mathcal{S}_B} y_i$, $\hat{Y}_{HT,A} = \sum_{i \in \mathcal{S}_A} d_i^A y_i$, $\hat{Y}_{HT,A}^{(B)} = \sum_{i \in \mathcal{S}_A} d_i^A \delta_i y_i$ and $\hat{Y}_{HT,AB} = \sum_{i \in \mathcal{S}_A \cap \mathcal{S}_B} y_i d_i^A \frac{N}{n_B}$. The RegDI estimator in (14) is equivalent to the PC ratio estimator in (13) when $d_i^B = N/n_B$. In this case, the PC ratio estimator incorporates $\mathcal{Y}$ as an auxiliary variable and performs calibration based on the unknown total population $X \equiv Y$. Then, $Y$ is replaced with $\hat{Y}_{HT,A}$. Similarly, $\hat{X}_B$ is replaced with $\hat{Y}_{HT,AB}$.

### 4.3 Pseudo-calibration estimators when the target variable is not observed in $\mathcal{S}_B$

When the target variable is not observed in $\mathcal{S}_B$, we could be in the cases (b) or (c). In the case (b), we observe the target variable with error, i.e., $\tilde{y}_i$ is generated by a measurement error model as $\tilde{y}_i = e(y_i) + \epsilon_i$, where $e(\cdot)$ is a *method* for estimating $\tilde{y}_i$, $\epsilon_i$ such that are independent error terms with zero mean and variance $V(\epsilon_i) = v(y_i)\sigma^2$. In the case (c) we predict its values according to a prediction model as $\tilde{y}_i = m(\mathbf{z}_i) + \epsilon_i$, where $m(\cdot)$ is a *method* for predicting $\tilde{y}_i$, $\mathbf{z}_i' = (\mathbf{x}_{i,B}', \mathbf{x}')$ such that $\epsilon_i$ are independent error terms with zero mean and variance $V(\epsilon_i) = v(\mathbf{z}_i)\sigma^2$. In both the cases, we can use the probability survey sample data, where the target variable is observed, to build the methods. Concerning the prediction methods, $m(\cdot)$ can belong to a very broad class of supervised prediction methods, encompassing both parametric and non-parametric methods, as well as machine learning techniques such as kernel methods, regression-tree (Hastie et al. 2001) and random forest (Breiman 2001). Non-parametric methods can be useful with high-dimensional and unstructured data, a scenario often encountered in big non-probability sources.

A real example of the case (c) is estimating the number of websites offering specific services, such as e-commerce. In this case, we can employ a web-scraping technique to collect text documents from the websites, perform text analysis, and then predict the presence of functionalities and services on the website using supervised machine learning techniques (Righi et al. 2019). Supervised machine learning methods learn from a labelled training set, which consists of predictors ($\mathbf{z}_i$) and their corresponding target values ($y_i$). After training, the method can be employed to make predictions on new, unseen data ($\tilde{y}_i$). We assume to observe the target variable in $\mathcal{S}_A$, where $\mathcal{S}_A \subset \mathcal{S}_B$ or $\mathcal{S}_A \cap \mathcal{S}_B \neq \emptyset$ (see Sect. 4.3.1). We train $m(\cdot)$ on the dataset $\{(y_i, \mathbf{z}_i) : i \in \mathcal{S}_A \cap \mathcal{S}_B\}$ to obtain $\hat{m}(\cdot)$. Then, we make deterministic predictions with $\bar{\tilde{y}}_i = \hat{m}(\mathbf{z}_i)$ or random predictions with $\hat{\tilde{y}}_i = \bar{\tilde{y}}_i + \hat{\epsilon}_i$, where $\hat{\epsilon}_i$ represents the estimated random error terms. Plugging the deterministic predictions in (10), we obtain the projection pseudo-calibration estimator, $\hat{Y}_{PC}^P$, similar to the *projection estimator* proposed by Kim and Rao (2011). Plugging the random predictions in (10), we obtain

$$\hat{Y}_{PC}^P = \sum_{i \in \mathcal{S}_B} w_i^B \hat{\tilde{y}}_i. \tag{15}$$

If the prediction method is misspecified or fails to capture the true relationship between the predictors and the target variable, then the estimates produced by (15) are biased. In cases where $\mathcal{S}_A \subset \mathcal{S}_B$, we introduce a correction term, defining the difference pseudo-calibration estimator in the case (c) as

$$\hat{Y}_{PC}^{D} = \sum_{i \in \mathcal{S}_B} w_i^B \hat{\hat{y}}_i + \sum_{i \in \mathcal{S}_A} d_i^A (y_i - \hat{\hat{y}}_i). \tag{16}$$

In the case (b), we can define an estimator similar to $\hat{Y}_{PC}^{D}$ by replacing $\hat{\hat{y}}_i$ with $\tilde{y}_i$ in (16).

***Remark 8*** The estimator $\hat{Y}_{PC}^{D}$ shares similarities with the estimator the $\hat{Y}_{DR2}$ described in (12). Notice how $\hat{Y}_{PC}^{D}$ reverses the roles of probability and non-probability samples compared to the $\hat{Y}_{DR2}$.

***Remark 9*** The estimator $\hat{Y}_{PC}^{D}$ shares a similar structure with the adjusted *projection estimator* proposed by Kim and Rao (2011) and the *difference estimator* developed by Breidt and Opsomer (2017), both defined in the model-assisted framework.

***Remark 10*** The asymptotic properties of (16) when $\hat{m}(\cdot)$ is used instead of $m(\cdot)$ are outlined in Breidt and Opsomer (2017). They provide conditions under which the differences $(\tilde{y}_i - \hat{\hat{y}}_i)$ can be considered negligible for many parametric and non-parametric methods. This theory is developed in the model-assisted framework, which is the context of (16) when considering the big non-probability sample frame list affected by under-coverage. Additionally, Chen et al. (2020) offer insights into the asymptotic properties in the model-based framework, particularly when the propensity score model is the logistic model, and the outcome model is parametric.

***Remark 11*** Given $\hat{Y}_{RegDI}$ in (4), where $y_i^* = \hat{\hat{y}}_i$, and assuming $m(\cdot)$ such that $E_m(\hat{\hat{y}}_i) = y_i$, then,

$$E_m(\hat{Y}_{RegDI}) = Y^{(B)} \frac{\hat{Y}_{HT,A}}{\hat{Y}_{HT,A}^{(B)}} + \text{Term of minor order}. \tag{17}$$

When we consider $\hat{Y}_{PC}^{D}$, with the first term defined as in (13), and $w_i^B$ being independent of $y_i$, we have that $E_m(\hat{Y}_{RegDI}) \approx E_m(\hat{Y}_{PC}^{D})$.

### 4.3.1 Pseudo-calibration estimators when $\mathcal{S}_A \cap \mathcal{S}_B \neq \mathcal{S}_A$

In some cases, it may happen that $\mathcal{S}_A \not\subset \mathcal{S}_B$ and $\mathcal{S}_A \cap \mathcal{S}_B \neq \emptyset$, meaning that for certain units in $\mathcal{S}_A$, we cannot observe $\mathbf{x}_B$. This condition generally arises when we plan $\mathcal{S}_A$ independently from $\mathcal{S}_B$, but other practical reasons may also lead to this situation. For instance, in the previous example of business statistics regarding the services and functionalities of enterprise websites, the condition $\mathcal{S}_A \not\subset \mathcal{S}_B$ and $\mathcal{S}_A \cap \mathcal{S}_B \neq \emptyset$ arises when we select enterprises in $\mathcal{S}_A$ that implement anti-scraping techniques to block automatic scraping procedures on their websites. Consequently, these enterprises cannot belong to $\mathcal{S}_B$. We handle this situation as a non-response problem and replace $d_i^A$ in (16) with $f_i^A$, which are the final adjusted weights, since $\mathcal{S}_A$ is not fully included in $\mathcal{S}_B$. The difference pseudo-calibration estimator (16) can be rewritten in this case as

$$\hat{Y}^D_{PC} = \sum_{i\in\mathcal{S}_B} w_i^B \hat{\tilde{y}}_i + \sum_{i\in\mathcal{S}_A\cap\mathcal{S}_B} f_i^A (y_i - \hat{\tilde{y}}_i). \tag{18}$$

## 5 Variance estimation

We estimate the variance of PC estimators using a jackknife-type method based on an adjusted version of the Delete-a-Group Jackknife (DAGJK) method (Kott 2001, 2006a), which is suitable for handling huge sample sizes. The DAGJK method offers computational advantages over the traditional Jackknife technique. It is well-suited for complex sampling strategies involving stratified design, several sampling phases, adjustment for non-response, calibration and composite estimation (Kott 2001). The variance estimation is asymptotically unbiased when the target parameter is a smooth function of the stratum means. However, it is not possible to guarantee that the DAGJK quantile variance estimation is unbiased.

The DAGJK method defines $G$ random replication groups drawn from the parent sample, i.e., $\mathcal{S}_B$ and $\mathcal{S}_A$. Then, $G$ estimation processes are carried out using the sampled data, excluding the units of one replication group. For the $g-$th ($g = 1, \dots, G$) replicated estimate, the method computes a weight for each unit, $w_i^{B(g)}$ and $d_i^{A(g)}$ based respectively on the $w_i^B$ and $d_i^A$ weights adjusted by the exclusion of the units in the group $g$. When $y_i^* = y_i$, the DAGJK variance estimation is

$$v(\hat{Y}_{PC}) = \frac{G-1}{G} \sum_{g=1}^{G} (\hat{Y}^{(g)}_{PC} - \hat{Y}_{PC})^2, \tag{19}$$

with $\hat{Y}^{(g)}_{PC} = \sum_{i\in\mathcal{S}_B} w_i^{B(g)} y_i$, being $w_i^{B(g)} = 0$ when the unit $i$ belongs to group $g$.

Let $y_i^* = \bar{\tilde{y}}_i$ be a deterministic prediction. In this case, we add a variability correction term into (19) and rewrite the DAGJK variance estimator as

$$\begin{aligned} v(\hat{Y}^D_{PC}) = \frac{G-1}{G} \Bigg\{ & \sum_{g=1}^{G} \Bigg( \sum_{i\in\mathcal{S}_B} w_i^{B(g)} \bar{\tilde{y}}_i - \hat{Y}^P_{PC} \Bigg)^2 \\ & + \sum_{g=1}^{G} \Bigg[ \sum_{i\in\mathcal{S}_A} d_i^{A(g)} (y_i - \bar{\tilde{y}}_i) - \sum_{i\in\mathcal{S}_A} d_i^A (y_i - \bar{\tilde{y}}_i) \Bigg]^2 \Bigg\}, \end{aligned} \tag{20}$$

being $w_i^{B(g)} = 0$ and $d_i^{A(g)} = 0$ when the unit $i$ belongs to group $g$. Finally, let $y_i^* = \hat{\tilde{y}}_i + \hat{\epsilon}_i$ be a random prediction, where $\hat{\epsilon}_i$ is the estimated error term obtained from the dataset $\{(y_i, \mathbf{z}_i) : i \in \mathcal{S}_A \cap \mathcal{S}_B\}$. In the case of observations with errors in $\mathcal{S}_B$, we use the dataset $\{(y_i, \tilde{y}_i) : i \in \mathcal{S}_A \cap \mathcal{S}_B\}$ to estimate the measurement error model and the relative error term. In both cases, the DAGJK method involves a random generation of $y_i^*$ values in each group, replacing the $\bar{\tilde{y}}_i$ in (20).

We evaluate (19) and (20) in the simulation study presented in the next section.

## 6 Simulation study

Following the simulation study 1) by Kim and Tam (2021), we generate a finite population, $\mathcal{U}$, of size $N = 1,000,000$. The response variable, $\mathcal{Y}$, is given by the following model:

$$y_i = 3 + 0.7(x_i - 2) + \eta_i,$$

where $x_i \sim \mathcal{N}(2, 1)$, $\eta_i \sim \mathcal{N}(0, 0.51)$ and $\eta_i$ is independent of $x_i$. Next, we generate a contaminated version of $y_i$ as follows:

$$\tilde{y}_i = 2 + 0.9(y_i - 3) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 0.5^2)$ and $\epsilon_i$ is independent of $y_i$.

Additionally, we generate the auxiliary variable $\boldsymbol{\xi}$ such that $cor(\mathbf{x}, \boldsymbol{\xi}) = 0.5$, which is given by:

$$\xi_i = \frac{cov(\mathbf{x}, \boldsymbol{\xi})}{var(\mathbf{x})} x_i + v_i,$$

where $v_i \sim \mathcal{N}(0, 1)$ and $v_i$ is independent of $x_i$. We define $\Xi_1 = \sum_{i \in U} \xi_{1i}$ and $\Xi_2 = \sum_{i \in U} \xi_{2i}$, where $\xi_{1i} = 1$ if $\xi_i \leq 1$ and 0 otherwise, and $\xi_{2i} = 1$ if $\xi_i > 1$ and 0 otherwise. We use these variables for prediction and calibration purposes.

We select two samples: $\mathcal{S}_A$ and $\mathcal{S}_B$, representing the probability and the non-probability samples, respectively. $\mathcal{S}_A$ is a simple random sample of size $n_A = 1000$, while $\mathcal{S}_B$ is selected by a different probability sampling of size $n_B = 500,000$. The latter is obtained by creating two strata in $\mathcal{U}$: stratum 1 consists of units with $x_i \leq 2$, while stratum 2 consists of those with $x_i > 2$. We define $X_1 = \sum_{i \in U} x_{1i}$ and $X_2 = \sum_{i \in U} x_{2i}$, where $x_{1i} = 1$ if $x_i \leq 2$ and 0 otherwise, and $x_{2i} = 1$ if $x_i > 2$ and 0 otherwise. Within each stratum, we independently select $n_{B1} = 300,000$ and $n_{B2} = 200,000$ observations, respectively, through simple random sampling. The target parameter is the finite population mean of $\mathcal{Y}$. This sampling procedure implies that the sample mean of $\mathcal{S}_B$ is smaller than the population mean.

The simulation study examines the first two scenarios proposed in Kim and Tam (2021):

1. Scenario I: we observe $y_i$ in both samples;
2. Scenario II: we observe $y_i$ in $\mathcal{S}_A$ and $\tilde{y}_i$ in $\mathcal{S}_B$.

The indicator variable $\delta_i$ is observed in both $\mathcal{S}_B$ and $\mathcal{S}_A$. Therefore, if $\delta_i = 1$ in $\mathcal{S}_A$, we have both $y_i$ and $\tilde{y}_i$.

### 6.1 Estimators

The simulation study considers two benchmark estimators:

1. Mean $\mathcal{S}_A = \frac{1}{n_A}\sum_{i\in\mathcal{S}_A} y_i$,
2. Mean $\mathcal{S}_B = \frac{1}{n_B}\sum_{i\in\mathcal{S}_B} y_i$,

and compares two classes of estimators integrating $\mathcal{S}_A$ and $\mathcal{S}_B$.

The first class of estimators considers the RegDI methods proposed by Kim and Tam (2021):

1. RegDI: regression data integration estimator of the form (1) with calibration equation
$$\sum_{i\in\mathcal{S}_A} w_i^A(1,\delta_i,\delta_i y_i) = \sum_{i\in U}(1,\delta_i,\delta_i y_i) = (N, n_B, Y_B^*).$$
2. RegDI$_{(X_1,X_2)}$: regression data integration estimator of the form (1) with calibration equation
$$\sum_{i\in\mathcal{S}_A} w_i^A(1,\delta_i,\delta_i y_i, x_{1i}, x_{2,1}) = \sum_{i\in U}(1,\delta_i,\delta_i y_i, x_{1i}, x_{2,1}) = (N, n_B, Y_B^*, X_1, X_2).$$
3. RegDI$_{(\Xi_1,\Xi_2)}$: regression data integration estimator of the form (1) with calibration equation
$$\sum_{i\in\mathcal{S}_A} w_i^A(1,\delta_i,\delta_i y_i, \xi_{1i}, \xi_{2,1}) = \sum_{i\in U}(1,\delta_i,\delta_i y_i, \xi_{1i}, \xi_{2,1}) = (N, n_B, Y_B^*, \Xi_1, \Xi_2).$$

The second class of estimators includes the PC estimators. For Scenario I, we consider the following:

1. PC$_{(X_1,X_2)}$: pseudo-calibration estimator of the form (10) with calibration equation
$$\sum_{i\in\mathcal{S}_B} w_i^B(x_{1i}, x_{2i}) = \sum_{i\in U}(x_{1i}, x_{2,1}) = (X_1, X_2).$$
2. PC$_{(\Xi_1,\Xi_2)}$: pseudo-calibration estimator of the form (10) with calibration equation
$$\sum_{i\in\mathcal{S}_B} w_i^B(\xi_{1i}, \xi_{2i}) = \sum_{i\in U}(\xi_{1i}, \xi_{2,1}) = (\Xi_1, \Xi_2).$$

For Scenario II, we consider the following estimators:

1. Difference Mean $\mathcal{S}_B = \frac{1}{n_B}\sum_{i\in\mathcal{S}_B}\tilde{y}_i + \frac{1}{N}\sum_{i\in\mathcal{S}_A} d_i^A(y_i - \tilde{y}_i)$: the sample mean of predictions in $\mathcal{S}_B$, corrected by the weighted residuals calculated in $\mathcal{S}_A$.
2. PC$^D_{(X_1,X_2)}$: pseudo-calibration estimator of the form (16) with calibration equation

**Table 2** Results of the five estimators for the simulation study based on design-based Monte Carlo simulations of size 1000

| Scenario | Estimator | Bias | SE | RMSE |
|---|---|---|---|---|
| I | Mean $\mathcal{S}_A$ | 0.00 | 0.031 | 0.031 |
| | Mean $\mathcal{S}_B$ | − 0.11 | 0.001 | 0.113 |
| | RegDI | 0.00 | 0.022 | 0.022 |
| | $\text{RegDI}_{(X_1,X_2)}$ | 0.00 | 0.021 | 0.021 |
| | $\text{RegDI}_{(\Xi_1,\Xi_2)}$ | 0.00 | 0.022 | 0.022 |
| II | Mean $\mathcal{S}_A$ | 0.00 | 0.031 | 0.031 |
| | Mean $\mathcal{S}_B$ | − 1.10 | 0.001 | 1.101 |
| | RegDI | 0.00 | 0.024 | 0.024 |
| | $\text{RegDI}_{(X_1,X_2)}$ | 0.00 | 0.022 | 0.022 |
| | $\text{RegDI}_{(\Xi_1,\Xi_2)}$ | 0.00 | 0.024 | 0.024 |

*RMSE* root mean squared error, *SE* standard error

**Table 3** Results of the six estimators for the simulation study based on Monte Carlo populations of size 1000

| Scenario | Estimator | Bias | SE | RMSE |
|---|---|---|---|---|
| I | Mean $\mathcal{S}_B$ | − 0.11 | 0.001 | 0.112 |
| | $\text{PC}_{(X_1,X_2)}$ | 0.00 | 0.001 | 0.001 |
| | $\text{PC}_{(\Xi_1,\Xi_2)}$ | − 0.10 | 0.001 | 0.098 |
| II | Mean $\mathcal{S}_B$ | − 1.10 | 0.001 | 1.101 |
| | Difference Mean $\mathcal{S}_B$ | − 0.11 | 0.021 | 0.108 |
| | $\text{PC}^D_{(X_1,X_2)}$ | 0.00 | 0.021 | 0.022 |
| | $\text{PC}^D_{(\Xi_1,\Xi_2)}$ | − 0.09 | 0.021 | 0.096 |

*RMSE* root mean squared error, *SE* standard error

$$\sum_{i\in\mathcal{S}_B} w_i^B(x_{1i},x_{2i}) = \sum_{i\in U}(x_{1i},x_{2,1}) = (X_1,X_2).$$

3. $\text{PC}^D_{(\Xi_1,\Xi_2)}$: pseudo-calibration estimator of the form (16) with calibration equation

$$\sum_{i\in\mathcal{S}_B} w_i^B(\xi_{1i},\xi_{2i}) = \sum_{i\in U}(\xi_{1i},\xi_{2,1}) = (\Xi_1,\Xi_2).$$

## 6.2 Results

The performance of each estimator is evaluated through the bias (Bias), the standard error (SE), and the root mean squared error (MSE) given by the Monte Carlo process.

The two classes of estimators employ different inference approaches: a design-based approach for the RegDI estimators, where the $y_i$ values are treated as fixed, and a model-based approach for the PC estimators, where the variable $\mathcal{Y}$ is considered random. For the RegDI estimators, we generate 1000 Monte Carlo samples for

**Table 4** Standard error estimates of four PC estimators for the simulation study based on 1000 Monte Carlo populations of size and on Delete-a-Group Jackknife estimator using 100 random groups

| Scenario | Estimator | SE MC | SE DAGJK |
|---|---|---|---|
| I | $PC_{(X_1,X_2)}$ | 0.00099 | 0.00102 |
| | $PC_{(\Xi_1,\Xi_2)}$ | 0.00108 | 0.00114 |
| II | $PC^D_{(X_1,X_2)}$ | 0.02137 | 0.02314 |
| | $PC^D_{(\Xi_1,\Xi_2)}$ | 0.02133 | 0.02362 |

The SE DAGJK values are the means of 1000 DAGJK estimates

both $\mathcal{S}_A$ and $\mathcal{S}_B$ from the finite population. We simulate 1000 Monte Carlo populations for the PC estimators and draw a single sample for each population for $\mathcal{S}_A$ and $\mathcal{S}_B$.

Table 2 shows the simulation study results for the design-based estimators. The results for the first three estimators (i.e., Mean $\mathcal{S}_A$, Mean $\mathcal{S}_B$, RegDI) are identical to the ones presented in Kim and Tam (2021) (see pg. 394, Table 2). As discussed in Kim and Tam (2021), Mean $\mathcal{S}_A$ and the RegDI estimators are unbiased in both scenarios. In contrast, Mean $\mathcal{S}_B$ estimator is always biased due to the selection bias in sample $\mathcal{S}_B$. The RegDI estimators have lower RMSE values. In particular, the RegDI$_{(X_1,X_2)}$ estimator, not considered in Kim and Tam (2021), has the lowest standard error. On the other hand, the RegDI$_{(\Xi 1,\Xi_2)}$ estimator, which employs calibration variables not strictly related to the $y_i$ values, leads to an inflation in the standard error (SE).

Table 3 shows the simulation study results of the model-based estimators. The Mean $\mathcal{S}_B$ estimator remains seriously biased due to the selection bias in $\mathcal{S}_B$. Focusing on Scenario I, the PC$_{(X_1,X_2)}$ estimator shows competitiveness compared to the RegDI estimators, while the PC$_{(\Xi_1,\Xi_2)}$ estimator is affected by the use of a slightly wrong propensity score model, implicitly defined by the $\xi_1$ and $\xi_2$ variables. In Scenario II, as shown in Table 3, the Mean $\mathcal{S}_B$ estimator increases the bias. In this case, indeed, both the outcome model for $\tilde{y}_i$ and the propensity score model for $w_i^B$ come into play. By utilizing the outcome model, the Difference Mean $\mathcal{S}_B$ estimator significantly reduces the bias. Scenario II does not present results for the projection-type estimators, PC$^P_{(X_1,X_2)}$ and PC$^P_{(\Xi_1,\Xi_2)}$, as they exclusively rely on the propensity score model. The PC$^P_{(X_1,X_2)}$ and PC$^P_{(\Xi_1,\Xi_2)}$ estimators exhibit bias levels closer to the Mean $\mathcal{S}_B$ estimator than the Difference Mean $\mathcal{S}_B$ estimator. The PC$^D_{(X_1,X_2)}$ estimator uses both models and remains competitive with the RegDI estimators. The PC$^D_{(\Xi_1,\Xi_2)}$ estimator reduces the bias compared to the Difference Mean $\mathcal{S}_B$ estimator. However, it is still more biased than the RegDI estimators.

The second part of the simulation study is devoted to the variance estimator presented in Sect. 5. We use the DAGJK method with 100 random groups. Table 4 shows the standard error estimates of the PC estimators using either $(X_1, X_2)$ or $(\Xi_1, \Xi_2)$ in both scenarios. The DAGJK values represent the mean values of the DAGJK estimates computed on 1000 samples of the Monte Carlo simulation. In Scenario I, the DAGJK variance estimates are close to the Monte Carlo variances. As expected, in Scenario II we slightly overestimate the DAGJK variance estimates.

## 7 An application to the European community survey data on ICT usage and e-commerce in enterprises

We implement the RegDI and PC estimators using the 2018 European Community Survey data on ICT usage and e-commerce in enterprises. This ICT survey is conducted yearly by Istat and by other member states of the EU. Additionally, we consider internet data scraped from enterprises' websites that fall within the ICT target population. The primary objective of the ICT survey is to supply users with indicators related to internet connectivity and usage, encompassing aspects such as website usage, social media engagement, and cloud computing. The survey's target population refers to enterprises with ten or more employees working in the industry and non-financial market services. The population frame is the Italian Business Register (Asia), which was last updated two years before the survey's reference period. For the 2018 ICT survey, this population comprises 199,914 units. The ICT survey considers a stratified simple random sampling design with strata given by four classes of number of persons employed (0–9; 10–19; 20–249; 250 or more), economic activities (24 Nace groups) and geographical breakdown (21 administrative regions at NUTS 2 level). The strata, including the fourth size class (enterprises with 250 and more persons employed), are taken entirely. The number of units within these strata is 3342. For the 2018 ICT survey, the sample of respondents consists of 22,097 units. The survey posed questions to enterprises, including whether a) the website enables online ordering, reservations, or bookings and b) there are links to social media on the website. We assign specific variable names, WEBORD ($\mathcal{Y}_1$) and WEBSM ($\mathcal{Y}_2$), to these two questions, respectively. The current ICT survey estimator employs a calibration method, which considers the number of enterprises and persons employed based on economic activity, size class, and administrative region, according to a complex combination of these variables. We use Internet data as a big non-probability sample (i.e., a big data source). This process starts with text documents collected through web scraping from the enterprise's websites. Specifically, we have gathered 93,848 scraped websites representing the units falling in $\mathcal{S}_B$. It is worth mentioning that the total number of websites in the target population is unknown. The ICT survey estimates approximately 134,655.82 enterprises with a relative error of about 1%. The web-scraping step returns information retrieval for the WEBSM variable. That means that we observe the variable with $y_{2i} = 1$ when the website has a link to social media and with $y_{2i} = 0$ otherwise. Using the text document of each website, we predict the WEBORD variable using a machine learning technique (Random Forest) as described in Bianchi et al. (2020) and Bruni and Bianchi (2020). We use a deterministic prediction for the WEBORD, meaning we use the estimated probability that the website incorporates functionalities for online ordering, reservations, or bookings. Further insights into the ICT survey, web scraping, and machine learning procedure can be found in Righi et al. (2019).

### 7.1 Estimators

We compare a simplified version of the estimator used by Istat for the ICT survey, denoted as T0, with four different RegDI estimators (RegDI.1, RegDI.2, RegDI.3, RegDI.4) and three other PC estimators (PC.1, PC.2, PC.3) for the population total.

T0 is a calibration estimator of the form (1) that uses the number of enterprises and employed persons for four enterprise-size classes (0–9; 10–19; 20–249; +249) and for three macro-regions (aggregation of NUTS 2 regions, North, Centre and South) as known totals. We set $\mathbf{x}_i = (1\boldsymbol{\lambda}_i', e_i\boldsymbol{\lambda}_i')'$, where $e_i$ is the number of employed persons in unit $i$, and $\boldsymbol{\lambda}_i = (\lambda_{i(0-9)}, \lambda_{i(10-19)}, \lambda_{i(20-249)}, \lambda_{i(+249)}, \lambda_{i(North)}, \lambda_{i(Centre)}, \lambda_{i(South)})'$, where the generic element of $\boldsymbol{\lambda}_i$, $\lambda_{i(d)}$, is equal to 1 if $i$ belongs to one of four enterprise-size classes or one of three macro-regions, and equal to 0 otherwise. For example, if the enterprise $i$ has ten employed persons and is located in southern Italy, then $\boldsymbol{\lambda}_i = (0, 1, 0, 0, 0, 0, 1)'$. Then, the calibration equation is

$$\sum_{i\in\mathcal{S}_A} w_i^A(1\lambda_i', e_i\lambda_i') = \sum_{i\in\mathcal{U}}(N_{0-9}, N_{10-19}, N_{20-249}, N_{+249}, N_{Centre}, N_{North}, N_{South},$$
$$E_{0-9}, E_{10-19}, E_{20-249}, E_{+249}, E_{Centre}, E_{North}, E_{South}),$$

where $N_d$ and $E_d$ are the total number of enterprises and employed persons, respectively, in each enterprise-size class and macro-region. As a result, the estimates of the total population and the seven sub-populations using T0 can be derived as

$$\hat{Y}_{j,T0} = \sum_{i\in\mathcal{S}_A} w_i^A y_{ji} \quad \text{and} \quad \hat{Y}_{j,T0(d)} = \sum_{i\in\mathcal{S}_A} w_i^A y_{ji}\lambda_{i(d)} \quad \text{for} \quad j = 1, 2. \tag{21}$$

The calibration variables are $(\mathbf{x}_i', \delta_i\boldsymbol{\lambda}_i')'$, $(\mathbf{x}_i', \delta_i\boldsymbol{\lambda}_i', \delta_i y_{1i}\boldsymbol{\lambda}_i')'$, $(\mathbf{x}_i', \delta_i\boldsymbol{\lambda}_i', \delta_i y_{2i}\boldsymbol{\lambda}_i')'$ and $(\mathbf{x}_i', \delta_i\boldsymbol{\lambda}_i', \delta_i y_{1i}\boldsymbol{\lambda}_i', \delta_i y_{2i}\boldsymbol{\lambda}_i')'$ for the RegDI.1, RegDI.2, RegDI.3 and RegDI.4 estimators, respectively. It follows that the estimates of the population total and the seven sub-populations using the RegDI estimators have the same form of (21) but different calibration weights.

The PC.1 estimator calibrates the weights using the same totals as T0. It corresponds to the estimator (10) for WEBSM and the estimator (15) for WEBORD. Additionally, for the WEBORD total, we implement the PC.2 and PC.3 estimators following the formula (18). In the PC.2 estimator, the sampling calibrated weights are adjusted by the factor $f_{2i}^A = \sum_{\mathcal{S}_A}\phi_i / \sum_{\mathcal{S}_A}\delta_i$, with $\phi_i = 1$ when the enterprise has the website and $\phi_i = 0$ otherwise. The PC.3 estimator uses the factor $f_{3i}^A = \sum_{\mathcal{S}_A}\phi_i w_i^A / \sum_{\mathcal{S}_A}\delta_i w_i^A$, where $w_i^A$ is the calibrated sampling weight of the ICT survey estimator (T0). It follows that the estimates of the population total and the seven sub-populations using the PC.1 estimator can be derived as

$$\hat{Y}_{1,PC.1}^P = \sum_{i\in\mathcal{S}_B} w_i^B \hat{\tilde{y}}_{1i} \quad \text{and} \quad \hat{Y}_{1,PC.1(d)}^P = \sum_{i\in\mathcal{S}_B} w_i^B \hat{\tilde{y}}_{1i}\lambda_{i(d)},$$
$$\hat{Y}_{2,PC.1} = \sum_{i\in\mathcal{S}_B} w_i^B y_{2i} \quad \text{and} \quad \hat{Y}_{2,PC.1(d)} = \sum_{i\in\mathcal{S}_B} w_i^B y_{2i}\lambda_{i(d)}.$$

**Table 5** Results of the considered estimators at the national level

| Estimator | Variable | Total | CI(95%)<br>Lower bound | CI(95%)<br>Upper Bound | Estimate $\notin$<br>T0 CI(95%) | CV |
|---|---|---|---|---|---|---|
| T0 | WEB | 134,655.82 | 131,831.46 | 137,480.18 | | 1.07% |
| | WEBORD | 26,451.41 | 24,473.67 | 28,429.14 | | 3.81% |
| | WEBSM | 68,221.35 | 65,157.69 | 71,285.01 | | 2.29% |
| RegDI.1 | WEBORD | 27,150.30 | 25,092.20 | 29,208.40 | | 3.87% |
| | WEBSM | 70,520.33 | 67,388.36 | 73,652.30 | | 2.27% |
| RegDI.2 | WEBORD | 27,387.05 | 25,806.85 | 28,967.25 | | 2.94% |
| | WEBSM | 70,684.85 | 67,577.39 | 73,792.32 | | 2.24% |
| RegDI.3 | WEBORD | 28,313.23 | 26,225.65 | 30,400.82 | | 3.76% |
| | WEBSM | 77,021.37 | 74,646.39 | 79,396.34 | ** | 1.57% |
| RegDI.4 | WEBORD | 27,541.93 | 25,989.47 | 29,094.39 | | 2.88% |
| | WEBSM | 77,022.19 | 74,647.43 | 79,396.96 | ** | 1.57% |
| PC.1 | WEBORD | 30,120.58 | 29,956.38 | 30,284.78 | ** | 0.27% |
| | WEBSM | 79,123.88 | 78,625.52 | 79,622.24 | ** | 0.31% |
| PC.2 | WEBORD | 26,860.18 | 25,740.40 | 28,361.63 | | 2.47% |
| PC.3 | WEBORD | 26,817.45 | 26,009.59 | 27,625.31 | | 1.54% |

CI(95%), Confidence Interval (CI) computed at the 95% confidence level; $**$, estimates outside the 95% CIs of T0; CV, Coefficient of Variation

The estimates of the population total and the seven sub-populations using the PC.2 and PC.3 estimators can be obtained as

$$\hat{Y}^{D}_{1,PC.l} = \sum_{i\in\mathcal{S}_B} w_i^B \hat{\hat{y}}_{1i} + \sum_{i\in\mathcal{S}_A} f_{li}^A (y_{1i} - \hat{\hat{y}}_{1i}) \quad (\text{for} \quad l = 2,3) \quad \text{and}$$

$$\hat{Y}^{D}_{1,PC.l(d)} = \sum_{i\in\mathcal{S}_B} w_i^B \tilde{y}_{1i}\lambda_{i(d)} + \sum_{i\in\mathcal{S}_A} f_{li}^A (y_{1i} - \hat{\hat{y}}_{1i})\lambda_{i(d)} \quad (\text{for} \quad l = 2,3).$$

### 7.2 Results

Table 5 shows the estimates for the totals at the national level. We can observe that the RegDI.1 estimator does not affect the Coefficient of Variation (CV) of the estimates compared to T0. On the other hand, the RegDI.2 and RegDI.3 estimators reduce the CV for the variable involved in the calibration. Only when we apply the RegDI.4 estimator we observe a substantial decrease in CV for both the WEBORD and WEBSM variables. These results highlight a crucial crossroads in a multi-purpose survey. The choices are twofold: (1) make a massive calibration, risking either the non-attainment of the optimal solution to the optimization problem or the inflation of variance due to excessively small or large final weights; (2) omit certain target variables from the calibration process and risking to compromise in the enhancement of their estimation accuracy. The estimates of WEBSM given by RegDI.3 and RegDI.4 fall outside the 95% Confidence Interval (CI) of T0. Since

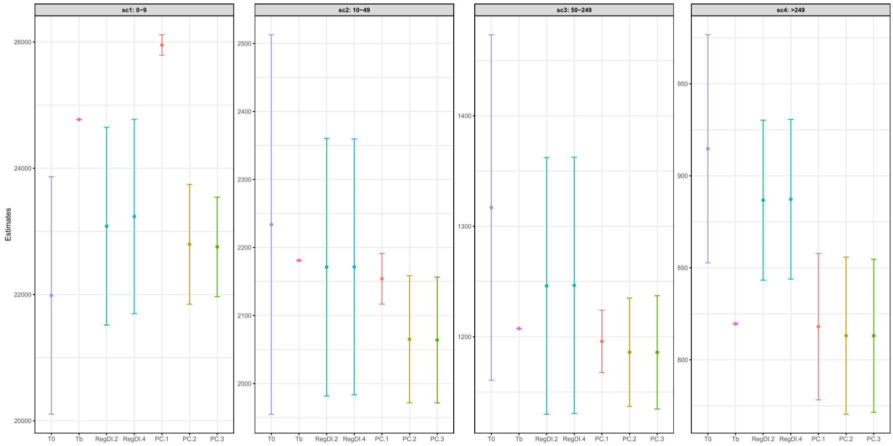

**Fig. 1** Estimator CIs (95%) by size class for WEBORD total

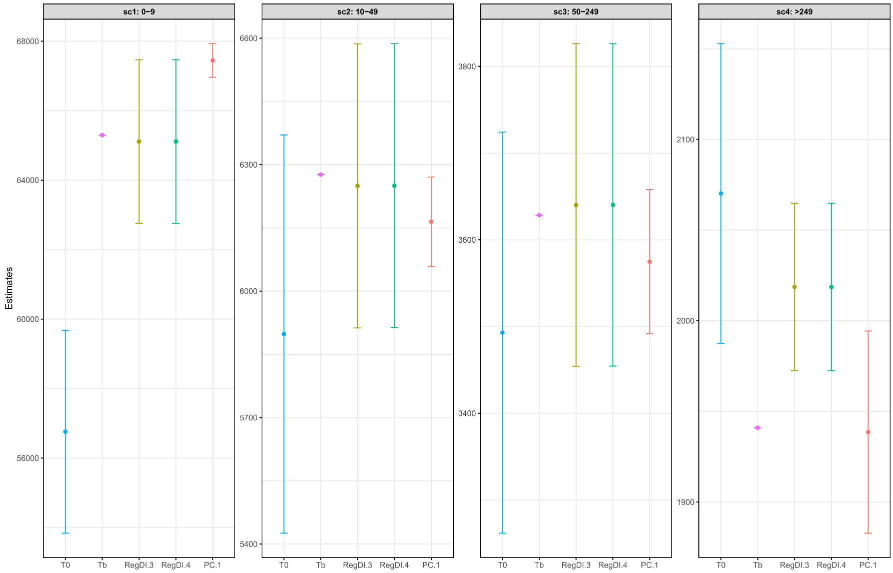

**Fig. 2** Estimator CIs (95%) by size class for WEBSM total

the RegDI.3 and RegDI.4 estimators are unbiased, the findings suggest that the T0 estimate has such a large error that it considerably underestimates the WEBSM total.

The analysis of the PC estimators reveals some important findings. The PC.1 estimates fall outside the 95% CI of T0. While we know the PC.1-WEBORD estimate can be biased since it bypasses the correction term, the PC.1-WEBSM estimate appears different from the corresponding T0 estimate. Nevertheless, the 95% CI of the PC.1-WEBSM estimator overlaps the CI of the RegDI.3 and RegDI.4 estimates. The consistency among these three estimates suggests that the PC.1-WEBSM

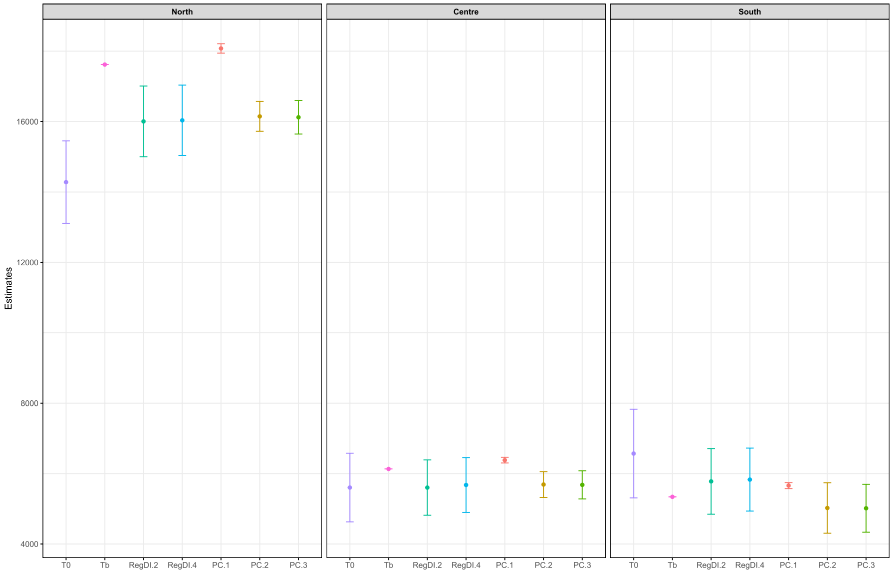

**Fig. 3** Estimator CIs (95%) by macro-regions for WEBORD total

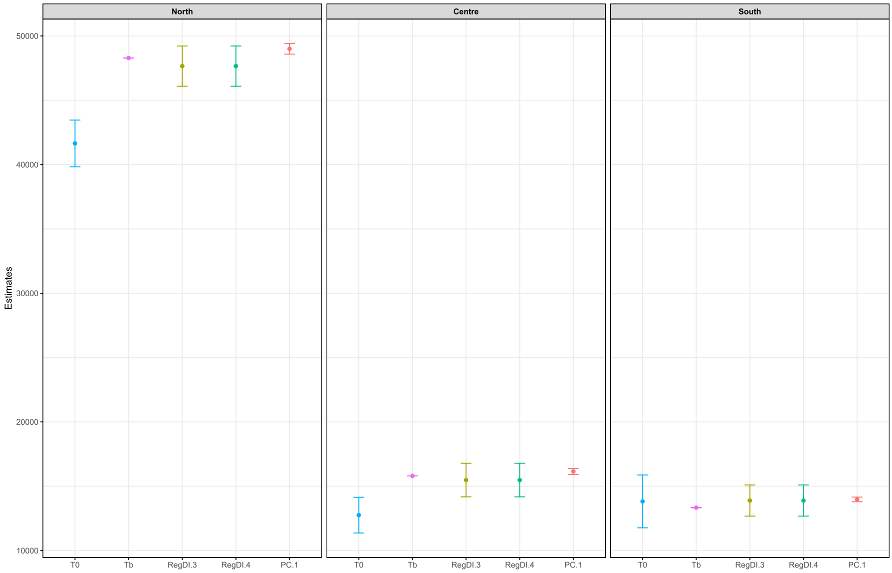

**Fig. 4** Estimator CIs (95%) by macro-regions for WEBSM total

estimator is unbiased. The CI of PC.1-WEBSM estimator is much narrower compared to the CIs of RegDI. We apply the pseudo-calibration difference estimators, PCE.2 and PCE.3, for the WEBORD total estimate. The PCE.2 and PCE.3 estimates fall within the 95% CI of the T0 estimate, and their 95% CIs overlap the

**Table 6** Coefficient of variation of the estimators for size classes by the macro-region domain of WEBORD total

| Domain | Average CV(%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | T0 | PC.2 | PC.3 | RegDI.1 | RegDI.2 | RegDI.3 | RegDI.4 |
| Group 1 | 12.91 | 6.18 | 6.11 | 13.59 | 14.36 | 13.95 | 14.54 |
| Group 2 | 7.50 | 3.73 | 3.75 | 7.85 | 6.26 | 7.68 | 6.23 |

Group 1: domains with a number of units in the interval [344; 547]

Group 2: domains with a number of units in the interval [1558; 8299]

**Table 7** Coefficient of variation of the estimators for size classes by macro-region domain of WEBSM total

| Domain | Average CV(%) | | | | | |
|---|---|---|---|---|---|---|
| | T0 | PC.1 | RegDI.1 | RegDI.2 | RegDI.3 | RegDI.4 |
| Group 1 | 8.00 | 3.18 | 8.47 | 8.58 | 9.16 | 9.26 |
| Group 2 | 4.78 | 1.05 | 7.02 | 4.75 | 3.59 | 3.59 |

Group 1: domains with a number of units in the interval [344; 547]

Group 2: domains with a number of units in the interval [1558; 8299]

RegDI-WEBORD estimators' 95% CIs. This suggests that we may have effectively adjusted for the bias in the measurement error of the big data target variable. The CV of the PC.3 estimator is smaller than that of T0 and is roughly equivalent to the CVs of the others RegDI.3 and RegDI.4.

We compare the estimates of WEBORD and WEBSM totals by size class domains (Figs. 1 and 2) and macro-regions domains (Figs. 3 and 4).

In Figs. 1, 2, 3 and 4, it is evident that the RegDI estimator 95% CIs always overlap the 95% CI of the T0 estimates except for WEBSM estimates in the domain $0-9$ size class or North macro-region. We underlined this evidence for the national estimate. While the 95% CIs of the RegDI estimators appear similar in length, they are slightly narrower than the 95% CI of the T0 estimate for some domains (such as the size class $0-9$ for WEBORD and WEBSM). The PC estimators give the shortest intervals. As expected, in certain domains, the WEBSM estimates significantly deviate from those of T0 (specifically, in the $0-9$ size class, Center and North macro-regions) (see Figs. 2 and 4). Regarding the WEBORD totals, the PC.1 estimates fall outside the CIs of T0 and frequently deviate significantly from those generated by the RegDI.1 estimator, which utilizes the same auxiliary variables. This outcome is anticipated because the PC.1 estimator ignores the correction term, and the prediction method (i.e., random forest technique) could fail to accurately capture the true relationship between the predictors and the target variable in specific domains. Consequently, the PC.1 estimator can be biased. The difference estimator adjusts the PC.1-WEBORD estimates that fall within the 95% CIs of the T0 estimate, or at least produces 95% CIs that overlap the 95% CIs of the RegDI.3 and RegDI.4 estimators. Figures 1 and 2 also include the $T_b$ estimator, which is a naïve PC estimator defined as $(\hat{N}_W/n_B)\sum_{S_B} y_i^*$, where $\hat{N}_W$ is the survey-based estimate of the number of units with a website. Finally, it is noteworthy that the estimates from PC.2 and PC.3 differ (though not significantly) from those generated by the RegDI.4 estimator, which incorporates all auxiliary variables. We interpret these findings as indicative of intrinsic distinctions arising from the utilization of information within the two classes of estimators.

Tables 6 and 7 investigate the sampling errors of the estimators of the cross-classified domain size class by macro-region (12 domains) by the average CV (%) for WEBORD and WEBSM total estimate, respectively. We categorize the domains into two groups: six domains with a sample size between 344 and 547 units (Group 1) and six domains with a sample size between 1558 and 8299 sample units (Group 2).

Tables 6 and 7 show that the PC estimators are more efficient. Among the RegDI estimators, those in Group 2 (large domains) are more efficient than T0. Conversely, the average CVs (%) for the RegDI estimators in Group 1 (small domains) are greater than the CV of T0. We attribute this to the increased number of calibration constraints, resulting in some units having extreme weights, which in turn leads to higher variance estimates. This effect of extreme weights is more pronounced in domains with small sample sizes.

## 8 Discussion

The PC estimators integrate data from various sources, including probability and big non-probability samples, administrative records, or statistical registers. They are applicable when the target variable is observed in the probability sample and, in the big non-probability sample, it is (a) observed correctly, (b) observed with error, or (c) predicted using highly correlated auxiliary variables. In the case (a), the PC estimators are inverse probability weighted estimators (Chen et al. 2020). In the cases (b) and (c), they represent a novel class of estimators with forms akin to the adjusted projection estimator (Kim and Rao 2011) and difference estimators (Breidt and Opsomer 2017) developed within the model-assisted framework. In these cases, the PC estimators reverse the roles of probability and non-probability samples in the estimation process compared to the doubly robust estimators.The paper outlines the RegDI estimators (Kim and Tam 2021), as they share the same informative context (cases (a) and (b)) required by the PC estimators. Both the PC and RegDI estimators employ calibration techniques, albeit with distinct approaches. The use of calibration tools is standard in the inferential context of data integration estimators, while it is less conventional in the context of PC estimators, although it has been previously suggested in the literature (Lee and Valliant 2009). With few exceptions, one of which is shown in the paper, the RegDI and PC estimators yield different point estimates. A jackknife-type variance estimator is introduced for the PC estimators suitable for large-scale datasets. Moreover, a comparative analysis is conducted between the PC and RegDI estimators, both defined within the same informative framework. This analysis leverages a Monte Carlo simulation and an experiment using real data from the ICT enterprise survey and information scraped from enterprise websites. The PC estimators show competitiveness with the RegDI estimators in Monte Carlo simulations, with the jackknife-type variance estimates very close to the Monte Carlo variances. In the experiment with real data, PC estimates are not significantly different from the RegDI estimates, and the confidence intervals are narrower than those of the RegDI.

**Funding** Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declarations

**Conflict of interest** All authors declare that they have no conflicts of interest.

## References

Beaumont JF (2020) Are probability surveys bound to disappear for the production of official statistics. Surv Methodol 46(1):1–28

Bethlehem J (2010) Selection bias in web surveys. Int Stat Rev 78(2):161–188

Bianchi G, Bri R, Scalfati F (2020) Identifying e-commerce in enterprises by means of text mining and classification algorithms Hindawi. Math Probl Eng 2018:1–8. https://doi.org/10.1155/2018/7231920

Breidt FJ, Opsomer JD (2017) Model-assisted survey estimation with modern prediction techniques. Stat Sci 32:190–205

Breiman L (2001) Random forests. Mach Learn 45:5–32

Bruni R, Bianchi G (2020) Website categorization: a formal approach and robustness analysis in the case of e-commerce detection. Expert Syst Appl 142(113):001

Chen Y, Li P, Wu C (2020) Doubly robust inference with nonprobability survey samples. J Am Stat Assoc 115(532):2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Citro CF (2014) From multiple modes for surveys to multiple data sources for estimates. Surv Methodol 40(2):137–162

Dever J, Valliant R (2010) A comparison of variance estimators for post-stratification to estimated control totals. Surv Methodol 36:45–56

Dever J, Valliant R (2016) General regression estimation adjusted for undercoverage and estimated control totals. J Surv Stat Methodol 4:289–318

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87:367–382

Elliot M, Valliant R (2017) Inference for nonprobability samples. Stat Sci 32:249–264

EUROSTAT (2018) Report describing the quality aspects of big data for official statistics. In: Work Package 8 Quality Deliverable 8.2. ESSnet Big Data

EUROSTAT (2020) Deliverable k3: Revised version of the quality guidelines for the acquisition and usage of big data. In: Workpackage K Methodology and quality. ESSnet Big Data II

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer New York Inc., New York

Horrigan MW (2013) Big data: A perspective from the BLS. AMSTAT News January:25–27

Japec L, Kreuter F, Berg M et al (2015) Big data in survey research: AAPOR task force report. Public Opin Q 79(4):839–880. https://doi.org/10.1093/poq/nfv039

Kim JK (2022) A gentle introduction to data integration in survey sampling. Surv Stat 85:19–29

Kim JK, Rao JNK (2011) Combining data from two independent surveys: a model-assisted approach. Biometrika 99(1):85–100

Kim JK, Tam SM (2021) Data integration by combining big data and survey sample data for finite population inference. Int Stat Rev 89(2):382–401

Kott PS (1994) A note on handling nonresponse in sample surveys. J Am Stat Assoc 89(426):693–696

Kott PS (2001) Delete-a-group jackknife. J Off Stat 17(4):521–526

Kott PS (2006) Delete-a-group variance estimation for the general regression estimator under Poisson sampling. J Off Stat 22(4):759–767

Kott PS (2006) Using calibration weighting to adjust for nonresponse and coverage errors. Surv Methodol 32(2):133

Lee S, Valliant R (2009) Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociol Methods Res 37(3):319–343

Little RJA, Rubin DB (2019) Statistical analysis with missing data, 3rd edn. Wiley, Hoboken

Meng XL (2018) Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. Ann Appl Stat 12(2):685–726

Pfeffermann D (2015) Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. J Surv Stat Methodol 3(4):425–483

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

Righi P, Bianchi G, Nurra A et al (2019) Integration of survey data and big data for finite population inference in official statistics: statistical challenges and practical applications. Stat Appl XVII(2):135–158

Righi P, Golini N, Bianchi G (2022) Big data and official statistics: some evidence. In: Balzanella A, Bini M, Cavicchia C et al (eds) Book of short the papers: 51st scientific meeting of the Italian statistical society. Pearson, Hoboken, pp 723–734

Rubin DB (1976) Inference and missing data. Biometrika 63:581–590

Rueda MDM, Pasadas-del-Amo S, Rodríguez BC et al (2023) Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a survey on the impact of the COVID-19 pandemic in Spain. Biom J 65(2):2200035

Särndal CE, Lundström S (2005) Estimation in surveys with nonresponse. John Wiley & Sons, Hoboken

Tam SM (2015) A statistical framework for analysing big data. Surv Stat 72:36–51

Tam SM, Clarke F (2015) Big data, official statistics and some initiatives by the Australian Bureau of Statistics. Int Stat Rev 83(3):436–448

Tam SM, Clarke F (2015) Big data, statistical inference and official statistics—methodology research papers. Australian Bureau of Statistics, Canberra

UNECE Big Data Quality Task Team (2014) A suggested framework for the quality of big data. Deliverables of the UNECE Big Data Quality Task Team

United Nations (2019) United Nations National Quality Assurance Frameworks Manual for Official Statistics. United Nations publication

Valliant R (2020) Comparing alternatives for estimation from nonprobability samples. J Surv Stat Methodol 8(2):231–263

Valliant R, Dorfman AH, Royall RM (eds) (2000) Finite population sampling and inference: a prediction approach. Wiley Series in Survey Methodology

Vehovar V, Toepoel V, Steinmetz S (2016) Non-probability sampling, vol 1. The Sage handbook of survey methods