Statistical Papers (2025) 66:111
https://doi.org/10.1007/s00362-025-01727-5

REGULAR ARTICLE

Estimation of the distribution function and quantiles through data integration

B. Cobo[1] · **S. Martínez**[2] · **M. Rueda**[3]

Received: 16 February 2024 / Revised: 2 February 2025

Abstract

Non-probability samples are increasingly as alternatives of probability samples to collecting detailed data from individuals. Non-probability sampling is a relatively inexpensive data source, although they require special treatment because the estimate may suffer from sample selection bias. In this paper, we consider methods for integrating a non-representative volunteer sample into a probability survey. We investigate several approaches to correcting non-probability sample selection bias in the estimation of the distribution function. We combine the estimators of the distribution function that correct the selection bias with the design unbiased estimators based on the probability sample. Our methodology for combining the voluntary and probability samples can be applied to other non-linear parameters. Empirical evidence of the improvements offered by the proposed methodology is provided in simulation settings.

Keywords Non-probability samples · Data integration · Survey sampling · Simulation

B. Cobo, S. Martínez, and M. Rueda have contributed equally to this work.

✉ S. Martínez
spuertas@ual.es

B. Cobo
beacr@ugr.es

M. Rueda
mrueda@ugr.es

[1] Department of Quantitative Methods for Economics and Business, University of Granada, Paseo de Cartuja, 7, 18011 Granada, Spain

[2] Department of Mathematics, University of Almería, Carretera Sacramento, s/n, La Cañada de San Urbano, 04120 Almería, Spain

[3] Department of Statistics and Operational Research, University of Granada, Av. de Fuente Nueva, s/n, 18071 Granada, Spain

Published online: 21 June 2025

1 Introduction

In finite population inference, probability sampling is usually used for obtaining a representative sample from the target population. However, official institutions that conduct surveys continually face increased demands for accuracy along with reduced resources, for example requests for efficient estimates for domains with small sample sizes. Multiple data sources are becoming increasingly available for statistical analyses and statistical offices face the increasing pressure to utilize other data sources as web surveys, mobile sensor data or social media data that provide timely data.

However, these data sources are nonprobability samples and makes its inferences prone to selection bias, mainly due to data-generating process. Self-selection bias can be very important, and invalidate the results when generalizing to the entire population specially if the estimated variables have some influence on the decision of the respondent to participate in the survey (Ferri-García and Rueda 2018). A clear example of the biases in the estimates produced by large non-probability surveys is presented in Bradley et al. (2021) who show this problem in the estimation of the first dose of COVID-19 Vaccine adoption with two large surveys: Delphi–Facebook (about 250,000 responses per week by using facebook newsfeed) and Census Households Press4 (around 75,000 every two weeks by using SMS and email as recruitment mode). These overestimated surveys adoption by 17 an 14 percentage points respectively. Their large sample sizes led to very small confidence intervals but very far from the correct values.

New bias correction techniques have been developed to infer parameters using data from non-probability sample. Some important methods are propensity score adjustment (Kim and Wang 2019), statistical matching (Rivers 2007), superpopulation modelling (Buelens et al. 2018) kernel smoothing methods (Wang et al. 2020), calibration adjustment (Ferri-García and Rueda 2018), combinations of these techniques (Chen et al. 2019) or bayesian approach (Rafei et al. 2022). The paper of Wu (2022) gives an idea of the state of the literature on the analysis of non-probability survey data. Some people have even come to believe that probability surveys could be phased out for the production of official statistics, although most authors believe that appropriate probability sampling methods should be used for real tests and non-probability sampling data is not a good substitute today (Beaumont 2020). In spite of their limitations, non-probability samples can be particularly useful to complement information from probability surveys in some situations. For example, they can be used in cases where the target population is a small sub-population unlikely to meet sample size requirements or when we are interested in non-demographic strata that cannot be taken into account in a sampling design. Even if the non-probability sample is biased, a very large volume of data can make the relative contribution of bias to the total error small, and can help not to exceed the total fixed error especially in those sub-populations where reliable estimates are needed and the probability survey includes few units.

There is recent literature on approaches to integrating surveys of different quality into a result to correct bias and reduce error. Several alternatives can be considered for integrating data obtained with probability and nonprobability samples. The simplest is the naive method, which consists of joint all data and all units are assigned the

same weight, but this method is rarely appropriate because nonprobability samples are not generally distributed proportionally across demographic or other important groups in the population. Alternatives were reviewed by several authors.

The pioneering work was that of Elliott and Haviland (2007) who propose a linear combination of the biased and unbiased estimators of the population mean where the weight of each sample takes into account the bias and error of each estimator. But in practice, bias and variations are unknown and must be estimated using available information of both samples and the authors show that large sizes of the probability sample are necessary for the method to be able to reduce the bias of the non-probability sample. Rueda et al. (2022) explore other alternatives that combine some of these ideas with the help of machine learning methods.

Other approach based on calibration weighting is done in Disogra et al. (2011) and requires a good selection of early adopter questions that are included in the two surveys. Kim and Tam (2021) developed new estimators for the population totals by stratifying the population into a nonprobability data stratum and a missing data stratum, and thus estimate the missing data stratum by using a probability sample. Authors also discuss how to improve the efficiency of the proposed estimator by using ratio and regression estimation.

Wiśniowski et al. (2020) consider a Bayesian approach for integrating a probability sample with a non-probability sample. The authors consider informative priors based on non-probability data and manage to reduce the variance and the mean square error of the estimators. Elliot (2009) proposes a new approach where probability and non-probability samples can be blended, and the resulting sample can be treated as a probability sample with new pseudo-weights. Robbins et al. (2020) define four estimators that integrate the two samples in a blended sample based on propensity score methods or on calibration weighting.

All these works are focused on the estimation of linear parameters, mainly totals or population means. The literature on sampling related to the estimation of functional parameters such as the distribution function is much scarcer. The issue of estimating the finite population distribution function arises when our interest lies in finding out the proportion of the values of the study variable which is less than or equal to some threshold. In certain situations, the need of cumulative distribution function is much more pertinent than totals and the means, since from convenient estimators of this function, we can estimate other relevant population parameters such as the Gini index (Goga and Ruiz-Gazen 2014), the reliability function (Acal et al. 2019) or specially population quantiles which are parameters of interest in many fields of research such as clinical chemistry (Bohn et al. 2019), atmospheric processes (Wilson et al. 2012), social science (Kimbro et al. 2011) and economics (Gelman et al. 2010) in which some measures and indicators depend on quantiles. Specifically, in economics it is very common to find variables with skewed distributions such as income, especially in studies of poverty and wage inequality, so measures based on quantile ratios are frequently used Burtless (1999); Jones and Weinberg (2000); Machin et al. (2003).

Although the distribution function is a particular case of a finite population mean of an indicator variable, there are some relevant aspects in its estimation that differs from the classical estimation of a population mean and hence when mean estimation techniques are applied directly to the estimation of a distribution function, could

suffer from several drawbacks (Silva and Skinner 1995). For instance, it is desirable that an estimator of a distribution function should have the properties of a distribution function but procedures that incorporate auxiliary information in the estimation of mean and totals can take values outside the range [0, 1] or do not fulfill the non-decreasing monotony, a property that is essential when estimating population quantiles. Additionally, the models specified for the study and auxiliary variables in a specific mean estimation technique cannot be applied with the corresponding indicator variables or the correlation between the indicator variables is usually weaker than that between the study and auxiliary variables, so the efficiency gain would be less than that obtained when these techniques are used to estimate a conventional mean. Consequently, it is necessary to establish specific techniques for the distribution function that take into account the particularities of its estimation.

There is enough work to improve the estimation of the distribution function and associated parameters such as quantiles and poverty measures with auxiliary information when working with probability samples, mainly using the calibration technique (see e.g. Martínez et al. (2010, 2020, 2012, 2015)). On the contrary, there are hardly any works that deal with techniques to reduce self-selection bias in the estimation of distribution functions.

Recently Rueda et al. (2022) establishes a general framework for making inference for the distribution function from non-probability surveys by applying techniques known as calibration, propensity score or statistical matching. The results of their simulation study show that self-selection bias can be greatly reduced particularly when using appropriate covariates and a valid machine learning technique. This paper consider the situation where the target variable has been observed in the non-probability sample only.

However, to our knowledge, very few, if any, studies have addressed the problem of estimation of the distribution function based on both: a probability and a nonprobability sample, when the study variables are measured in both samples. Our goal in this paper is to efficiently combine both the non-probability and probability samples to estimate the distribution function. Our contributions include a proposal for three data integration methods that will allow us to define different distribution function estimators and addressing the of properties for these estimators.

The rest of the article is organized as follows: In Sect. 2, we start by describing the estimation of the distribution function through rigorous mathematical notations. In Sect. 3 we propose a method that consider a natural post-stratified estimator of population means, similarly to the method used in Kim and Tam (2021). In Sect. 4, we explore some alternatives to the post-stratified estimator, by using the inverse probability weighting estimator (IPW) based on propensity scores for the non-probability sample. We then combine this IPW estimator with the unbiased estimator based on the probability sample in several ways. We discuss how to further improve the efficiency of the proposed integration estimators by using calibration in Sect. 5. Section 6 analyzes the conditions that must be met for the proposed estimators to be genuine distribution functions. Simulation studies are presented in Sect. 9. Finally, conclusions are drawn in Sect. 10.

2 Basic setup

Let U denote a finite population of size N, $U = \{1, \ldots, j, \ldots, N\}$. Let s_r be a probability sample of size n_r selected from U under a probability sampling design (s_r, p_r) with $\pi_j = \sum_{s_r \ni j} p_r(s_r)$ the first order inclusion probability for individual j. Let s_v be a non-probability (volunteer) sample of size n_v, self-selected from U. Let y be the variable of interest in the survey estimation. Our goal is the estimation of the distribution function $F_y(t)$ for the study variable y that can be defined as follows:

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{1}$$

where $\Delta(\cdot)$ denote the indicator function, given by:

$$\Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k. \end{cases}$$

The population distribution function, $F_y(t)$, can be estimated via the Horvitz-Thompson estimator:

$$\widehat{F}_{Yr}(t) = \frac{1}{N} \sum_{k \in s_r} d_k \Delta(t - y_k) \tag{2}$$

being $d_k = 1/\pi_k$. This estimator is design-unbiased of the distribution function and the design-based variance of this estimator is given by:

$$V_p(\widehat{F}_{Yr}(t)) = \frac{1}{N^2} \sum_{i,j \in U} \frac{\Delta(t - y_i)}{\pi_i} \frac{\Delta(t - y_j)}{\pi_j} (\pi_{ij} - \pi_i \pi_j). \tag{3}$$

where π_{ij} are the second order inclusion probabilities of the sampling design p_r. If $\pi_{ij} > 0 \; \forall (i, j)$, an unbiased estimator is given by:

$$\widehat{V}_p(\widehat{F}_{Yr}(t)) = \frac{1}{N^2} \sum_{i,j \in s_r} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\Delta(t - y_i)}{\pi_i} \frac{\Delta(t - y_j)}{\pi_j}. \tag{4}$$

$F_y(t)$ can be also estimated with the naive estimator based on the sample distribution function of y in s_v:

$$\widehat{F}_{Yv}(t) = \frac{1}{n_v} \sum_{j \in s_v} \Delta(t - y_j). \tag{5}$$

Let I_v be an indicator variable of an element being in s_v, this is

$$I_{vk} = \begin{cases} 1 & k \in s_v \\ 0 & k \notin s_v. \end{cases} \tag{6}$$

Rueda et al. (2022) show that this estimator is biased and that the bias and the mean squared error of the estimator are given by

$$B(\widehat{F}_{Yv}(t)) = E_R(\widehat{F}_{Yv}(t) - F_y(t)) = \frac{1}{f_v} E_R(Cov(I_v, \Delta(t-y)))$$

$$MSE(\widehat{F}_{Yv}(t)) = \frac{1}{f^2} E_R(Corr(I_v, \Delta(t-y))^2) Var(I_v) Var(\Delta(t-y)) =$$

$$= E_R(Corr(I_v, \Delta(t-y))^2) \times \left(\frac{1}{f_v} - 1\right) \times Var(\Delta(t-y))$$

where E_R denotes the expectation with respect to the random mechanism for I_v and $f_v = n_v/N$. So if $Corr(I_v, \Delta(t-y))$ exists there will be a self-selection bias.

The mechanism of participation of the units in the non-probability sample is therefore fundamental in the behaviour of the estimators. Usually, three mechanisms are considered in the literature of this issue:

- Completely random participation mechanism: when *R* is independent of the variables under study and of the auxiliary variables. In this case the naive estimators without adjusting for the sampling process are not subject to selection biases.
- Ignorable participation mechanism: when *R* and the study variable is independent given the auxiliary variables. This mechanism holds if the set of covariates contains all predictors for the outcome that affect the possibility of being selected in sample s_v. This mechanism is similar to the Missing at random (MAR) mechanism used in non-response literature.
- Non-ignorable participation mechanism: *R* depends directly on *y*. This mechanism is similar to the NMAR mechanism in non-response.

Non ignorable participation mechanism is an important but difficult topic for analysis of non-probability survey samples and much of the existing literature, are based on the unverifiable assumption that the sampling mechanism for the non-probability sample is ignorable. In this paper we will follow the usual literature and also consider the case where the participation mechanism is ignorable.

3 A first proposal for data integration for the distribution function

First, we will consider a simple estimator proposal by decomposing the population distribution function into two parts: the part with the observed elements in the non-probability sample and a second part with the unobserved elements.

Proposition 1 *The proposed estimator defined as*

$$\overline{F}_{yP1}(t) = \frac{1}{N}\left(\sum_{k\in s_v}\Delta(t-y_k) + (N-n_v)\frac{\sum_{k\in s_r} d_k(1-I_{vk})\Delta(t-y_k)}{\sum_{k\in s_r} d_k(1-I_{vk})}\right) \quad (7)$$

is a asymptotically design-based estimator of the distribution function obtained from two samples.

Proof We consider the population distribution function decomposition:

$$F_y(t) = \frac{1}{N}\left(\sum_{k\in s_v}\Delta(t-y_k) + \sum_{k\in U-s_v}\Delta(t-y_k)\right) = \frac{1}{N}(n_v\widehat{F}_{Yv}(t) + (N-n_v)F_{U-s_v}(t))$$

being $F_{U-s_v}(t) = \frac{1}{N-n_v}\sum_{k\in U}(1-I_{vk})\Delta(t-y_k)$.

The unknown term $F_{U-s_v}(t)$ is thus estimated by the Hájek estimator $\frac{\sum_{k\in s_r} d_k(1-I_{vk})\Delta(t-y_k)}{\sum_{k\in s_r} d_k(1-I_{vk})}$ which is asymptotically unbiased of $F_{U-s_v}(t)$.

This estimator can be written in the form

$$\overline{F}_{yP1}(t) = \frac{1}{N}\sum_{s} w_{kT}\Delta(t-y_k) \quad (8)$$

where $s = s_r \cup s_v$, being $w_{kT} = 1$ for $k \in s_v$ and

$$w_{kT} = \frac{(N-n_v)d_k(1-I_{vk})}{\sum_{j\in s_r} d_j(1-I_{vj})} = \frac{(N-n_v)d_k(1-I_{vk})}{\widehat{N-n_v}},\ \text{for } k \in s_r$$

and $\widehat{N-n_v}$ denotes the estimator of $N-n_v$ obtained from the probability sample.

Note that this estimator has not considered any modeling for the probabilities of participating in the non-probability survey, and therefore is valid for any pattern of voluntariness, in contrast to the other estimators that will be presented later and that depend on the participation mechanism used (in general, the ignorable mechanism). The fundamental thing is that the probability sample is well designed so that it represents the population. □

4 A second proposal for data integration for the distribution function

In recent years there is a wide variety of frameworks to adjust for bias of estimators based on non-probability samples when the study variable is observed in the nonprobability sample only but auxiliary information can be obtained from an existing probability survey sample from the same population. The most common are as mass imputation, propensity score method (PSA), kernel weighting (KW), calibration weighting, and doubly robust estimation methods. In this section we will propose new estimators of the distribution function by combining these bias adjusted estima-

tors with the unbiased estimators based on the probability sample. To do so, we start by briefly describing the PSA and KW techniques that we will use for constructing weights for sample s_v to improve the representativeness of this sample.

4.1 Propensity score adjustment

The propensity score of the individual can be formulated, following notation in Chen et al. (2019), as the expected value of I conditional on his target variable and covariates' value:

$$\pi_k^v = E[I_{vk}|\mathrm{x}_k, y_k] = P(I_{vk} = 1|\mathrm{x}_k, y_k) \tag{9}$$

If the selection is ignorable, then $P(I_{vk}|\mathrm{x}_k, y_k) = P(I_{vk} = 1)$ and estimates obtained from s_v would be unbiased. In the rest of the paper we will consider the following strong ignorability condition: the sampling indicator I_{vk} of sample s_v and the study variable y are conditionally independent given x; i.e. $P(I_{vk}|\mathrm{x}_k, y_k) = P(I_{vk}|\mathrm{x}_k)$. We will also assume that the propensities are positive $\pi_k^v > 0$.

The propensity for an individual to take part on the non-probability survey is obtained by training a predictive model (often a logistic regression) on the dichotomous variable, I_{s_v}, which measures whether a respondent from the combination of both samples took part in the volunteer survey or in the reference survey. Covariates used in the model, x, are measured in both samples (in contrast to the target variable which is only measured in the non-probability sample), thus the formula to compute the propensity of taking part in the volunteer survey with a logistic model, π^v, can be displayed as

$$\pi^v(\mathrm{x}) = m(\gamma^T \mathrm{x}) \tag{10}$$

for some vector γ, as a function of the model covariates.

Using data from both samples we can estimate propensity scores by maximizing the pseudo-likelihood (Chen et al. 2019):

$$\tilde{l}(\gamma) = \sum_{s_v} log \frac{m(\gamma, \mathrm{x}_k)}{1 - m(\gamma, \mathrm{x}_k)} + \sum_{s_r} \frac{1}{\pi_{rk}} log(1 - m(\gamma, \mathrm{x}_k)). \tag{11}$$

The estimated propensities $\hat{\pi}_k^v = m(\hat{\gamma}, \mathrm{x}_k)$ obtained from the pseudomaximum likelihoor estimator $\hat{\gamma}$ are thus used to calculate new pseudo-weights, such as Valliant weight Valliant (2020) $w_k^V = \frac{1}{\hat{\pi}_k^v}$ for the non-probability sample. These weights are then used to compute two versions of the IPW estimator for the distribution function, depending on whether the population size N is known or not:

$$\overline{F}_{yIPW}(t) = \frac{1}{N} \sum_{k \in s_v} w_k^V \Delta(t - y_k)$$

and

$$\overline{F}_{yIPWH}(t) = \frac{1}{\sum_{k\in s_v} w_k^V} \sum_{k\in s_v} w_k^V \Delta(t - y_k).$$

The properties of the above estimators are developed under both the model for the propensity scores and the survey design for the probability sample. Under certain regularity conditions and assuming the logistic regression model for the propensity scores the IPW estimator $\overline{F}_{yIPW}(t)$ is asymptotically unbiased for the population total distribution and an asymptotic expression for its variance is given by:

$$V(\overline{F}_{yIPW}(t)) = \sum_U (\Delta(t - y_k)/\hat{\pi}_{ki} - \mathrm{b}_1^T \mathrm{x}_k)^2 (1 - \hat{\pi}_k^v)\hat{\pi}_k^v + \mathrm{b}_1^T D \mathrm{b}_1 \quad (12)$$

where

$$\mathrm{b}_1^T = \sum_U (1 - \hat{\pi}_k^v)\Delta(t - y_k)\mathrm{x}_k^T / \sum_U \hat{\pi}_k^v (1 - \hat{\pi}_k^v)\mathrm{x}_k \mathrm{x}_k^T$$

and $D = V_p(\sum_{k\in s_r} d_k \hat{\pi}_k^v \mathrm{x}_k)$.

The proof of this result can be obtained like that theorem 1 of Chen et al. (2019) but changing the values y_k by $\Delta(t - y_k)$.

Other types of weights based on estimated propensities have been formulated by various authors. A first example appears in the work of Schonlau and Couper (2017) where the weights adjust the volunteer sample to the population of the probability sample, rather than the complete population U: $w_k^{SC} = \frac{1-\hat{\pi}_k^v}{\hat{\pi}_k^v}$. Lee and Valliant (2009) developed other weights whereby the combined sample $s_v \cup s_r$ is grouped into g equally-sized strata of similar propensity scores from which an average propensity is calculated for each group. The final weights of the nonprobability sample to be applied in linear estimators are defined as: $w_k^L = f_c \cdot \frac{N}{n_v}, \quad k \in s_v, c \ni k$ where $f_c = \frac{\sum_{k\in s_r^c} d_k^r / \sum_{k\in s_r} d_k^r}{\sum_{j\in s_v^c} d_j^v / \sum_{j\in s_v} d_j^v}$, and s_r^c and s_v^c are the subset of individuals from the reference and the nonprobability sample respectively that belong to the c-th stratum. Alternative weights also based on groups are considered in Valliant and Dever (2011), $w_k^{VD} = \frac{n_{vc}}{\sum_{k\in s_v^c} \hat{\pi}_k^v}, \quad k \in s_v$ being n_{vc} the size of s_v^c.

4.2 Kernel weighting

Wang et al. (2020) propose construct kernel weights weighting the design weights from probability sample, according to similarity between the individuals from both samples

$$w_k^{KW} = \sum_{j\in s_r} w_{rj} k_{kj} \quad (13)$$

being $k_{kj} = \frac{K\{d_{kj}/h\}}{\sum_{k \in s_v} K\{d_{kj}/h\}}$, $d_{kj} = \widehat{\pi}_k^v - \widehat{\pi}_j^v, k \in s_v, j \in s_r$,$K\{.\}$ a kernel function centred at zero, and h the corresponding bandwidth.

4.3 Tree-based inverse propensity weighted

Chu and Beaumont (2019) propose a IPW estimator where the estimation of the propensity is based on an adaptation of the classification and regression trees (CART) algorithm. After the tree has been grown the nonprobability sample is partitioned into g homogeneous propensity groups (terminal nodes), s_{vg} and he propensity for each individual $i \in s_{vg}$ is estimated as: $\hat{\pi}_i^{TrIPW} = \frac{n_{vg}}{\sum_{j \in s_{rg}} d_j}$ Based on these propensities, the weights are defined as

$$w_k^{TrIPW} = \frac{1}{\hat{\pi}_k^{TrIPW}} \tag{14}$$

4.4 Integration estimators of the distribution function

In this subsection we propose a new method for obtain distribution function estimators by combining the probability and non-probability samples, normalizing the weights. We thus define the blended sample $s = s_v \cup s_r$ and we consider $\hat{w}_k = \frac{n_v}{nr+n_v} w_k^{TC}$, $(TC = V, SC, L, VD, KW, TrIPW)$ for units in s_v and $\hat{w}_k = \frac{n_r}{nr+n_v} d_k$ for units in s_r.

So, we define an integrated weighted estimator for the distribution function as: h

$$\overline{F}_{yP2}(t) = \frac{1}{N} \sum_{k \in s} \hat{w}_k \Delta(t - y_k). \tag{15}$$

Another option to normalize the weights is to consider the weights $\tilde{w}_k = \frac{1-\hat{\pi}_{vk}}{\pi_k}$ that take into account both the design weights and the pseudo-weights estimated for all units in the joint sample. From them the estimator can be define the estimator:

$$\overline{F}_{yP3}(t) = \frac{1}{N} \sum_{k \in s} \tilde{w}_k \Delta(t - y_k). \tag{16}$$

Other weights used by Robbins et al. (2020) to estimate means are

$$\ddot{w}_k = d_k * \frac{(\sum_r d_k)(\sum_v w_k^{2-TC})}{(\sum_r d_k)(\sum_v w_k^{2-TC}) + (\sum_r d_k^2)(\sum_v w_k^{TC})} \tag{17}$$

for $k \in s_r$ and

$$\ddot{w}_k = w_k^{TC} * \frac{(\sum_r d_k)(\sum_v w_k^{2-TC})}{(\sum_r d_k)(\sum_v w_k^{2-TC}) + (\sum_r d_k^2)(\sum_v w_k^{TC})} \tag{18}$$

for $k \in s_v$.

If we use these weights to weight the samples, we can formulate the following estimator of the distribution function:

$$\overline{F}_{yP4}(t) = \frac{1}{N} \sum_{k \in s} \ddot{w}_k \Delta(t - y_k). \tag{19}$$

5 Improving the efficiency of the data integration estimators by calibration

We now discuss how to further improve the efficiency of the proposed estimator by using calibration. Calibration approach introduce by Deville and Särndal (1992) is the most used technique for weights adjustment, aiming at ensuring consistency among estimates of different sample surveys and some improving the precision of estimators. After the calibration adjustment, we will hope that the sample s can resemble the population U. This technique of weighting was previously used to estimate population totals and means when information for two independent surveys from the same target population are available by Disogra et al. (2011) and Kim and Tam (2021).

Let z be a set of auxiliary variables related to y, whose the population total Z are known. Suppose that an initial set of weights $\{\omega_{kI}, k \in s\}$ is available for all units in the sample s. This system of weights could be a system of weights obtained with any method included in the previous sections. Given a pseudo distance $G(\cdot, \cdot)$ we try to find an estimator $\overline{F}_{yic}(t) = \frac{1}{N} \sum_{k \in s} w_k^{ic} \Delta(t - y_k)$ where the new set of calibrated weights w_k^{ic} for all $k \in s = s_v \cup s_r$ minimize the distance to the weights ω_{kI}:

$$\min_{\omega_k} \sum_{k \in s} G(w_{ck}, \omega_{kI}) \tag{20}$$

while respecting the following condition:

$$\sum_{k \in s} w_k^{ic1} \mathrm{z}_k = \mathrm{Z}. \tag{21}$$

Proposition 2 *If we consider the chi-square distance:*

$$\sum_{k \in s} \frac{(\omega_{kI} - w_{ck})^2}{w_{ck} q_k} \tag{22}$$

where we assume that q_k are positive constants not related to d_k, the calibration estimator is:

$$\overline{F}_{yic1}(t) = \frac{1}{N}\sum_{k\in s}\omega_{kI}\Delta(t-y_k) + \frac{\gamma_1}{N}\cdot\sum_{k\in s}\omega_{kI}q_k\mathrm{z}_k\Delta(t-y_k). \tag{23}$$

and the new weights are

$$w_k^{ic1} = \omega_{kI} + \gamma_1\cdot\omega_{kI}q_k\mathrm{z}_k \tag{24}$$

with

$$\gamma_1 = \left(\mathrm{Z} - \sum_{k\in s}\omega_{kI}\mathrm{z}_k\right)^T\left(\sum_{k\in s}\omega_{kI}q_k\mathrm{z}_k\mathrm{z}_k^T\right)^{-1}$$

Proof The demonstration of this result is similar to that of the paper Deville and Särndal (1992) by changing the variable y to $\Delta(t-y_k)$.

Finally, if we know the value of the auxiliary vector z_k for all $k\in U$, following Rueda et al. (2007) we can consider the following pseudo variable:

$$g_k = \widehat{\beta}'\mathrm{z}_k \text{ for } k = 1,2,\ldots N \tag{25}$$

$$\widehat{\beta} = \left(\sum_{k\in s}\omega_{kTk}\mathrm{z}_k\mathrm{z}_k^{'}\right)^{-1}\cdot\sum_{k\in s}\omega_{kTk}\mathrm{z}_k y_k \tag{26}$$

We consider P points $t_j \quad j = 1,2,\ldots,P$ with $t_1 < t_2 < \ldots t_P$ and we denote by $\mathrm{t} = (t_1,\ldots,t_P)$ and $\Delta(\mathrm{t}-g_k) = (\Delta(t_1-g_k),\ldots,\Delta(t_P-g_k))^T$.

Now we consider the pseudo variable g and the calibration procedure that replaces the weights ω_{kI} in the sample s by a new calibrated weights w_k^{ic2} with the minimization of (22) under the following condition:

$$\frac{1}{N}\sum_{k\in s}w_k^{ic2}\Delta(\mathrm{t}-g_k) = F_g(\mathrm{t}) \tag{27}$$

In this case, the calibration weights are:

$$w_k^{ic2} = \omega_{kI} + \frac{\gamma_2}{N}\cdot\omega_{kI}q_k\Delta(\mathrm{t}-g_k) \tag{28}$$

with

$$\gamma_2 = N^2 \cdot \left(F_g(\mathrm{t}) - \frac{1}{N} \sum_{k \in s} \omega_{kI} \Delta(\mathrm{t} - g_k) \right)^T \cdot \widehat{H}^{-1}$$

where

$$\widehat{H} = \left(\sum_{k \in s} \omega_{kI} q_k \Delta(\mathrm{t} - g_k) \Delta(\mathrm{t} - g_k)^T \right)$$

The new proposed calibration estimator is:

$$\overline{F}_{yic2}(t) = \frac{1}{N} \sum_{k \in s} \omega_{kI} \Delta(t - y_k) + \frac{\gamma_2}{N^2} \cdot \sum_{k \in s} \omega_{kI} q_k \Delta(\mathrm{t} - g_k) \Delta(t - y_k) \quad (29)$$

□

6 Properties

When estimating the distribution function $F_y(t)$, an important issue to consider is the compliance of the conditions of the distribution function by a new proposed estimator $\widehat{F}_y(t)$. The fulfillment of the distribution function conditions allows us to use a new estimator $\widehat{F}_y(t)$ in quantile estimation from the inverse function associated with the estimator $\widehat{F}_y(t)$. For an estimator $\widehat{F}_y(t)$ to be a true distribution function, it must meet the following conditions:

i) $\widehat{F}_y(t)$ is continuous on the right,
ii) $\widehat{F}_y(t)$ is monotone nondecreasing.
iii) a) $\lim_{t \to -\infty} \widehat{F}_y(t) = 0$ and b) $\lim_{t \to +\infty} \widehat{F}_y(t) = 1$,

However, not all the new estimators proposed in previous section meet all the above properties. In this section we will analyze whether each of the estimators proposed in the previous sections satisfy all the properties of the distribution function and if they do not satisfy any of the properties, we will propose alternatives to match the previous unfulfilled conditions. Since the two conditions i) and iii a) are clearly satisifed by all the proposed estimators, we are going to focus on analyzing the fulfillment of conditions ii) and iii b)

Firstly, concerning the new estimator $\overline{F}_{yP1}(t)$ satisfies the condition ii) because the associated weights w_{kT} are positive for all sample units. Concerning the new estimators in the family $\overline{F}_{yP2}(t)$, all of them satisfy the condition ii) if the associated weights ω_k^{TC} are positive for all sample units. This condition is fullfied by the weights ω_k^{TC} with $TC = V, SC, L, VD, KW$ and *TrIPW*.

Similarly, for the new estimators in the families $\overline{F}_{yP3}(t)$ and $\overline{F}_{yP4}(t)$ for $TC = V, SC, L, VD, KW$ and *TrIPW* also the corresponding weights are all positive so the condition ii) is guarantied. However, to meet the condition iii b) with the estimators $\overline{F}_{yP2}(t)$, $\overline{F}_{yP3}(t)$ and $\overline{F}_{yP4}(t)$, it is necessary to divide the resulting estimator by the sum of its weights while estimator $\overline{F}_{yP1}(t)$ always fulfills this condition.

Finally, regarding calibration estimators, following (Rueda et al. 2007) the estimator $\overline{F}_{yic2}(t)$ fullfils the condition ii) if $q_k = c$ for all sample units. In general, the calibration weights for the calibration estimator $\overline{F}_{yic1}(t)$ are not positive for all sample units with the chi-square distance (22) and therefore condition ii) cannot be guaranteed. To meet this condition, there are several calibration methods for obtaining nonnegative weights (see e.g Deville et al. (1993); Kott and Liao (2012)). One of the calibration mechanisms considered to obtain non-negative weights is the so-called raking procedure Deville et al. (1993). Based on this method, we can consider the corresponding calibration process with the distance function given by:

$$G_s(\omega_k, \omega_{kI}) = \sum_{k \in s} \frac{1}{q_k} \left(\omega_k \log \frac{\omega_k}{\omega_{kI}} - \omega_k + \omega_{kI} \right). \tag{30}$$

With this distace measure, the calibration weights have the form:

$$\omega_k = \omega_{kI} \exp\left(\gamma_1 \cdot q_k \mathrm{z}_k\right)$$

and we can avoid negative calibrated weights Deville et al. (1993), so the calibration estimator $\overline{F}_{yic1}(t)$ can fulfill the condition ii).

Following Kott and Liao (2012), another way to avoid negative calibrated weights with the estimator $\overline{F}_{yic1}(t)$ is the logistic-response model. Following this calibration method, we can consider the calibrated weights given by:

$$\omega_k = \omega_{kI} \cdot (1 + exp(\gamma_1 \cdot z_k)) \tag{31}$$

and find a vector γ_1 to satisfy

$$\sum_{k \in s} \omega_k z_k = \sum_{k \in s} \omega_{kI} \cdot (1 + exp(\gamma_1 \cdot z_k)) = \mathrm{Z}$$

Since the calibrated weights (31) obtained with this method are positive, the version of the estimator $\overline{F}_{yic1}(t)$ obtained with them satisfies property ii).

Regarding condition iii b), none of the calibrated estimators meets this property in general. To meet this condition with the estimator $\overline{F}_{yic1}(t)$, we can add an auxiliary variable $z_k^* = 1$ for $k \in U$ in the auxiliary vector z_k. In the case of the calibration estimator $\overline{F}_{yic2}(t)$, following (Rueda et al. 2007), we can guarantee the condition iii b) by taking t_P sufficiently large.

Table 1 Bias and mean square error of the proposed estimators considering various sizes of the non-probability sample in the considered scenarios

	nb=250		nb=500		nb=1000	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
	Scenario 1					
$P1$	0,0026	0,0007	0,0024	0,0007	0,0024	0,0007
$P2\_V$	0,0549	0,0007	0,0748	0,0009	0,0846	0,0009
$P3$	0,0444	0,0007	0,0451	0,0007	0,0425	0,0007
$P4\_V$	0,0544	0,0007	0,0744	0,0009	0,0844	0,0009
	Scenario 2					
$P1$	0,0015	0,0007	0,0017	0,0007	0,0022	0,0007
$P2\_V$	0,0021	0,0004	0,0025	0,0003	0,0033	0,0003
$P3$	0,1707	0,0104	0,1715	0,0102	0,1713	0,0101
$P4\_V$	0,0018	0,0005	0,0022	0,0003	0,0033	0,0003
	Scenario 3					
$P1$	0,0010	0,0007	0,0154	0,0006	0,0017	0,0008
$P2\_V$	0,1304	0,0040	0,1728	0,0067	0,2057	0,0091
$P3$	0,1608	0,0099	0,1720	0,0110	0,1570	0,0097
$P4\_V$	0,1228	0,0036	0,1667	0,0063	0,1994	0,0086
	Scenario 4					
$P1$	0,0013	0,0009	0,0011	0,0008	0,0017	0,0008
$P2\_V$	0,0130	0,0005	0,0190	0,0004	0,0232	0,0003
$P3$	0,0290	0,0007	0,0295	0,0006	0,0288	0,0005
$P4\_V$	0,0158	0,0004	0,0217	0,0004	0,0252	0,0003

7 Application of new estimators in quantile and poverty measures estimation

Nowadays, studies on poverty, wage inequality and social exclusion are issues of main priority for economic research (Darvas 2019; Jones and Weinberg 2000; Meyer and Sullivan 2012) and for official institutions, governments and society (Guio et al. 2021; Eurostat Products Datasets 2022; Shrider et al. 2021) being a key aspect the development of indexes to measure wage inequality (Eurostat Experimental statistics 2022). In economics research, percentile ratios such as P90/P10; P95/P20 and P80/P20 (Jones and Weinberg 2000); P50/P5 and P50/P25 (Dickens and Manning 2004), P50/P10 (Burtless 1999) or P95/P50 (Machin et al. 2003) have been widely considered by previous studies for measuring the wage inequality. Likewise, percentile ratios measures are usually used by organizations to measure income inequality, such as the percentile ratios ratios P95/P20; P95/P50; P90/P10; P80/P50; P80/P20 and P20/P50 used for the US Census Bureau (Shrider et al. 2021) or the P80/P20 percentile ratio by employed Eurostat to assess the wage inequality in the European Union (Eurostat Products Datasets 2022). Given the relevance of wage inequality measures, in this section we focus on estimating the poverty measures based on percentile ratios.

Let a finite population $U = \{1, \ldots, N\}$ with distribution function $F_y(t)$ given by (1), and let α a value such that $0 < \alpha < 1$, the population α-quantile of y is given by:

Table 2 Bias and mean square error of the proposed estimators adding auxiliary information in the considered scenarios

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
$P1$	0,0009	0,0007	0,0020	0,0007	0,0019	0,0007	0,0011	0,0008
$P1\_C1$	0,0065	0,0004	0,0052	0,0004	0,0068	0,0004	0,0037	0,0005
$P1\_C21$	0,0012	0,0007	0,0016	0,0007	0,0011	0,0007	0,0009	0,0008
$P1\_C22$	0,0015	0,0006	0,0013	0,0006	0,0013	0,0006	0,0010	0,0008
$P2\_V$	0,0712	0,0008	0,0025	0,0003	0,1738	0,0067	0,0190	0,0004
$P2\_V\_C1$	0,0599	0,0005	0,0034	0,0002	0,0156	0,0003	0,0034	0,0001
$P2\_V\_C21$	0,0679	0,0007	0,0015	0,0003	0,0019	0,0003	0,0156	0,0003
$P2\_V\_C22$	0,0640	0,0007	0,0013	0,0002	0,0019	0,0002	0,0131	0,0002
$P3$	0,0425	0,0007	0,1701	0,0100	0,1609	0,0099	0,0295	0,0006
$P3\_C1$	0,0472	0,0004	0,0043	0,0002	0,0142	0,0003	0,0040	0,0002
$P3\_C21$	0,0421	0,0006	0,0033	0,0005	0,0257	0,0007	0,0111	0,0003
$P3\_C22$	0,0415	0,0006	0,0033	0,0004	0,0131	0,0006	0,0096	0,0003
$P4\_V$	0,0709	0,0008	0,0022	0,0003	0,1667	0,0062	0,0217	0,0004
$P4\_V\_C1$	0,0596	0,0005	0,0034	0,0002	0,0145	0,0003	0,0037	0,0001
$P4\_V\_C21$	0,0676	0,0007	0,0016	0,0003	0,0018	0,0003	0,0178	0,0003
$P4\_V\_C22$	0,0637	0,0007	0,0013	0,0002	0,0017	0,0002	0,0150	0,0002

$$Q_y(\alpha) = \inf\{t : F_y(t) \geq \alpha\} = F_y^{-1}(\alpha) \tag{32}$$

and given two values $0 < \alpha_2 < \alpha_1 < 1$, the percentile ratio $R(\alpha_1, \alpha_2)$ is define as follow:

$$R(\alpha_1, \alpha_2) = \frac{Q_y(\alpha_1)}{Q_y(\alpha_2)} = \frac{F_y^{-1}(\alpha_1)}{F_y^{-1}(\alpha_2)} \tag{33}$$

Given an estimator $\widehat{F}_y(t)$ of the distribution function that satisfies the distribution function properties, a generic procedure for obtaining an estimator for $Q_y(\alpha)$ can be established as follows:

$$\widehat{Q}_y(\alpha) = \inf\{t : \widehat{F}_y(t) \geq \alpha\} = \widehat{F}_y^{-1}(\alpha) \tag{34}$$

Consequently, the percentile ratio $R(\alpha_1, \alpha_2)$ can be estimated as follows:

$$\widehat{R}(\alpha_1, \alpha_2) = \frac{\widehat{Q}_y(\alpha_1)}{\widehat{Q}_y(\alpha_2)} \tag{35}$$

Since the distribution function estimators $\overline{F}_{yP1}(t)$, $\overline{F}_{yP2}(t)$, $\overline{F}_{yP3}(t)$, $\overline{F}_{yP4}(t)$, $\overline{F}_{yic1}(t)$ and $\overline{F}_{yic2}(t)$ allow us to integrate the information provided by a non-representative volunteer sample with the information provided from a probability sample and since in the previous section we have established the necessary conditions for the proposed estimators to fulfill all the properties of the distribution function, through the generic mechanism described above we can estimate $Q_y(\alpha)$ as $R(\alpha_1, \alpha_2)$ integrating information from a non-probability sample and a probability sample, thus obtaining the corresponding quantile estimators $\overline{Q}_{yP1}(\alpha)$, $\overline{Q}_{yP2}(\alpha)$, $\overline{Q}_{yP3}(\alpha)$, $\overline{Q}_{yP4}(\alpha)$, $\overline{Q}_{yic1}(\alpha)$ and $\overline{Q}_{yic2}(\alpha)$ and the corresponding percentile ratio estimators $\overline{R}_{yP1}(\alpha_1, \alpha_2)$, $\overline{R}_{yP2}(\alpha_1, \alpha_2)$, $\overline{R}_{yP3}(\alpha_1, \alpha_2)$, $\overline{R}_{yP4}(\alpha_1, \alpha_2)$, $\overline{R}_{yic1}(\alpha_1, \alpha_2)$ and $\overline{R}_{yic2}(\alpha_1, \alpha_2)$.

8 Variance estimation

Variance estimation for the proposed estimators is a challenging problem. Variance estimation under the sample $s = s_r \cup s_v$ is difficult because involves at least two different sources of variation: taking into account the two random mechanisms, and the probabilities of the conditional expectation, we have

$$V(\overline{F}_y(t)) = V_p E_R(\overline{F}_y(t)) + E_p V_R(\overline{F}_y(t))$$

where *R* stands for the model of the selection mechanism for the sample s_v and *p* refers to the probability sampling design for s_r.

Table 3 Bias and mean square error of the estimators considering the different weights in the considered scenarios

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
$P2\_V$	0,0724	0,0008	0,0024	0,0003	0,1742	0,0068	0,0190	0,0004
$P2\_SC$	0,0600	0,0008	0,0009	0,0007	0,1610	0,0065	0,0358	0,0006
$P2\_LV$	0,0574	0,0008	0,0097	0,0012	0,3630	0,0339	0,0227	0,0005
$P2\_VD$	0,0721	0,0008	0,0028	0,0005	0,1877	0,0080	0,0182	0,0003
$P2\_KW$	0,0571	0,0008	0,0152	0,0006	0,3445	0,0318	0,0274	0,0005
$P2\_TrIPW$	0,0229	0,0008	0,0009	0,0005	0,0592	0,0016	0,0028	0,0007
$P4\_V$	0,0721	0,0008	0,0022	0,0003	0,1667	0,0062	0,0217	0,0004
$P4\_SC$	0,0590	0,0008	0,0010	0,0006	0,1316	0,0044	0,0404	0,0007
$P4\_LV$	0,0563	0,0008	0,0077	0,0010	0,1338	0,0055	0,0257	0,0005
$P4\_VD$	0,0718	0,0008	0,0026	0,0005	0,1755	0,0070	0,0209	0,0003
$P4\_KW$	0,0558	0,0008	0,0105	0,0006	0,0779	0,0025	0,0309	0,0005
$P4\_TrIPW$	0,0222	0,0008	0,0008	0,0004	0,0351	0,0009	0,0032	0,0007

Plug-in estimators can be used to construct variance estimators for all the required components but it is not a simple issue. In practice, the use of jackknife and bootstrap techniques (Wolter 2007) in the variance estimation for nonlinear parameters should be more advantageous because of their wide applicability for different cases and conditions. Valliant (2020) and Chen et al. (2019) use resampling techniques to obtain estimators of the variances of linear parameters from non-probability samples. Rueda et al. (2022) use a jackniffe technique to obtain estimates of variance and confidence intervals for means and proportions from a combination of probability and non-probability survey data. These authors propose a grouped jackknife in which both the random mechanisms are repeated for each replicate at the expense of some extra computation. Kim and Tam (2021) developed a procedure to apply the bootstrap method for variance estimation for the Mass Imputation technique and established the consistency of this bootstrap variance estimator of MI estimator under certain conditions. These methodology can be used in our context to estimate the variances of the proposed distribution function estimators but formal proof of the consistency of such a replication estimator does not exist. Depending on the estimator used and the sources of randomness that are present, a specific resampling method has to be developed.

In the simulation study in the next section we have used the bootstrap method adapted to our estimators. We denote by $\overline{F}_y(t)$ any of the estimators proposed in the preceding sections, we use the following algorithm:

1. For each iteration bootstrap ($b = 1, \ldots, B$) extract a simple random sample with replacement of size n_r from s_r, obtaining a bootstrap replicate denoted by ($d_{vi}^{(b)}$, y_i, $\mathrm{x}_i^{(b)}$)
2. Extract a simple random sample with replacement of size n_v from s_v obtaining bootstrap $(y_j^{(b)}, \mathrm{x}_j^{(b)})$
3. From the bootstrap samples $s_v^{(b)}$ and $s_r^{(b)}$ calculate the corresponding estimator $\overline{F}_y(t)^{(b)}$
4. Compute the bootstrap variance estimators as

$$\hat{V}_B(\overline{F}_y(t)) = \frac{1}{B}\sum_{b=1}^{B}(\overline{F}_y^{(b)}(t) - \overline{F}_y(t))^2.$$

9 Simulation study

We carry out a simulation study in which we are going to study the previously proposed estimators. We are going to use the data called BigLucy from the TeachingSampling R package Gutiérrez Rojas (2020). This data set contains the financial variables of 85396 industrial enterprises of a city in a particular fiscal year. Specifically we are going to use the variable Income (the total ammount of a company's earnings (or profit) in the previous fiscal year. It is calculated by taking revenues and adjusting for the cost of doing business) as a variable of interest throughout the entire simulation

Table 4 Bias and mean square error of the estimators considering the different machine learning techniques in the considered scenarios

	glm		gbm		nnet		kknn	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
	Scenario 1							
$P2\_V$	0,0724	0,0008	0,0471	0,0006	0,0674	0,0008	0,0727	0,0008
$P3$	0,0435	0,0007	0,0101	0,0007	0,0335	0,0007	0,2387	0,0178
$P4\_V$	0,0721	0,0008	0,0460	0,0006	0,0665	0,0008	0,0714	0,0008
	Scenario 2							
$P2\_V$	0,0024	0,0003	0,0036	0,0003	0,0033	0,0004	0,0048	0,0003
$P3$	0,1705	0,0101	0,1475	0,0079	0,1647	0,0096	0,0815	0,0028
$P4\_V$	0,0022	0,0003	0,0034	0,0003	0,0031	0,0004	0,0047	0,0003
	Scenario 3							
$P2\_V$	0,1742	0,0068	0,1699	0,0062	0,1795	0,0071	0,1795	0,0068
$P3$	0,1614	0,0100	0,1267	0,0068	0,1237	0,0068	0,1861	0,0139
$P4\_V$	0,1667	0,0062	0,1692	0,0062	0,1746	0,0067	0,1776	0,0067
	Scenario 4							
$P2\_V$	0,0190	0,0004	0,0202	0,0004	0,0168	0,0003	0,0124	0,0003
$P3$	0,0295	0,0006	0,0394	0,0008	0,0233	0,0004	0,1892	0,0116
$P4\_V$	0,0217	0,0004	0,0230	0,0003	0,0172	0,0003	0,0139	0,0003

study. As variables when performing the calibration we consider Level (the industrial companies are discrimitnated according to the taxes declared. There are small, medium and big companies), Employees (the total number of persons working for the company in the previuos fiscal year), Taxes (the total ammount of a company's income tax), SPAM (indicates if the company uses the Internet and WEBmail options in order to make self-propaganda), ISO (indicates if the company is certified by the International Organization for Standardization) and as covariates to estimate the propensities, all of the above plus the Years variable (the age of the company). In the case of the probability sample we draw it through simple random sampling, and in the case of the non-probability sample we will consider three scenarios, the first (Scenario 1) in which the probability of selection is

$$\pi^v = \frac{e^{Taxes}}{1 + e^{Taxes}}$$

the second (Scenario 2) in which we consider a stratified sample of the SPAM variable, assigning 80% to the "no" category and 20% to the "yes" category, and the third scenario (Scenario 3) in which we select a simple random sample from the companies that have more than 63 workers. Finally, we consider a fourth scenario (Scenario 4) in which the probability of selection depends on the variable of interest.

$$\pi^v = \frac{e^{Taxes+0{,}05Income}}{1 + e^{Taxes+0{,}05Income}}$$

In this scenario we consider that the probability sample is obtained by stratified sampling design with uniform allocation in which the stratification variable was the

Table 5 Variance, confidence intervals, coverage and lenght scenario 1 considering the different quantiles and estimators

Quantile	Method	Theoretical Variance	Variance	Lower bound	Upper bound	Coverage	Length
0,1	$P1$	0,00038	0,00036	0,0994	0,1007	0,9348	0,0736
0,1	$P2\_V$	0,00011	0,00011	0,0807	0,0811	0,5173	0,0401
0,1	$P3$	0,00016	0,00015	0,0871	0,0877	0,7729	0,0479
0,1	$P4\_V$	0,00011	0,00011	0,0808	0,0812	0,5193	0,0402
0,2	$P1$	0,00064	0,00063	0,1980	0,2005	0,9430	0,0982
0,2	$P2\_V$	0,00019	0,00021	0,1677	0,1685	0,3971	0,0563
0,2	$P3$	0,00029	0,00031	0,1787	0,1799	0,7719	0,0687
0,2	$P4\_V$	0,00019	0,00021	0,1678	0,1686	0,4043	0,0565
0,25	$P1$	0,00078	0,00075	0,2512	0,2541	0,9440	0,1069
0,25	$P2\_V$	0,00025	0,00026	0,2177	0,2187	0,4308	0,0630
0,25	$P3$	0,00039	0,00039	0,2303	0,2318	0,7770	0,0776
0,25	$P4\_V$	0,00025	0,00026	0,2179	0,2189	0,4328	0,0631
0,3	$P1$	0,00086	0,00083	0,3001	0,3033	0,9430	0,1128
0,3	$P2\_V$	0,00029	0,00030	0,2664	0,2675	0,4745	0,0682
0,3	$P3$	0,00046	0,00047	0,2796	0,2815	0,8147	0,0848
0,3	$P4\_V$	0,00029	0,00030	0,2665	0,2677	0,4745	0,0683
0,4	$P1$	0,00094	0,00094	0,4000	0,4037	0,9470	0,1204
0,4	$P2\_V$	0,00036	0,00038	0,3702	0,3717	0,6507	0,0762
0,4	$P3$	0,00059	0,00061	0,3836	0,3860	0,8982	0,0969
0,4	$P4\_V$	0,00036	0,00038	0,3704	0,3718	0,6517	0,0763
0,5	$P1$	0,00097	0,00099	0,4985	0,5023	0,9552	0,1230
0,5	$P2\_V$	0,00042	0,00043	0,4774	0,4791	0,8147	0,0810
0,5	$P3$	0,00072	0,00073	0,4894	0,4923	0,9399	0,1058
0,5	$P4\_V$	0,00042	0,00043	0,4775	0,4792	0,8157	0,0810
0,6	$P1$	0,00097	0,00094	0,6083	0,6119	0,9470	0,1199
0,6	$P2\_V$	0,00046	0,00045	0,5949	0,5966	0,8992	0,0828
0,6	$P3$	0,00087	0,00084	0,6056	0,6089	0,9389	0,1133
0,6	$P4\_V$	0,00046	0,00045	0,5949	0,5967	0,9002	0,0830
0,7	$P1$	0,00085	0,00082	0,7054	0,7086	0,9369	0,1119
0,7	$P2\_V$	0,00042	0,00040	0,6978	0,6994	0,9297	0,0780
0,7	$P3$	0,00080	0,00075	0,7070	0,7099	0,9338	0,1075
0,7	$P4\_V$	0,00042	0,00040	0,6979	0,6994	0,9297	0,0781
0,75	$P1$	0,00074	0,00074	0,7509	0,7538	0,9430	0,1062
0,75	$P2\_V$	0,00036	0,00034	0,7451	0,7464	0,9277	0,0726
0,75	$P3$	0,00065	0,00063	0,7533	0,7557	0,9460	0,0980
0,75	$P4\_V$	0,00036	0,00034	0,7451	0,7465	0,9308	0,0726
0,8	$P1$	0,00059	0,00063	0,7995	0,8020	0,9491	0,0983
0,8	$P2\_V$	0,00030	0,00029	0,7957	0,7968	0,9379	0,0663
0,8	$P3$	0,00051	0,00051	0,8026	0,8046	0,9440	0,0883
0,8	$P4\_V$	0,00030	0,00029	0,7957	0,7969	0,9389	0,0663
0,9	$P1$	0,00032	0,00035	0,8999	0,9013	0,9633	0,0734
0,9	$P2\_V$	0,00015	0,00016	0,8995	0,9002	0,9521	0,0492
0,9	$P3$	0,00025	0,00027	0,9032	0,9042	0,9511	0,0647
0,9	$P4\_V$	0,00015	0,00016	0,8995	0,9002	0,9521	0,0490

Table 6 Variance, confidence intervals, coverage and lenght scenario 2 considering the different quantiles and estimators

Quantile	Method	Theoretical Variance	Variance	Lower bound	Upper bound	Coverage	Length
0,1	$P1$	0,00037	0,00035	0,0998	0,1012	0,9277	0,0734
0,1	$P2\_V$	0,00014	0,00015	0,1005	0,1011	0,9564	0,0476
0,1	$P3$	0,00033	0,00034	0,1158	0,1172	0,8894	0,0722
0,1	$P4\_V$	0,00014	0,00015	0,1005	0,1011	0,9532	0,0475
0,2	$P1$	0,00063	0,00062	0,1992	0,2016	0,9511	0,0978
0,2	$P2\_V$	0,00027	0,00028	0,1990	0,2000	0,9468	0,0650
0,2	$P3$	0,00063	0,00065	0,2338	0,2363	0,7574	0,0999
0,2	$P4\_V$	0,00027	0,00027	0,1991	0,2001	0,9436	0,0648
0,25	$P1$	0,00073	0,00074	0,2521	0,2550	0,9500	0,1065
0,25	$P2\_V$	0,00031	0,00033	0,2530	0,2543	0,9543	0,0715
0,25	$P3$	0,00072	0,00079	0,2941	0,2972	0,7053	0,1098
0,25	$P4\_V$	0,00031	0,00033	0,2530	0,2543	0,9521	0,0715
0,3	$P1$	0,00081	0,00082	0,3006	0,3039	0,9511	0,1125
0,3	$P2\_V$	0,00034	0,00038	0,3003	0,3018	0,9617	0,0764
0,3	$P3$	0,00079	0,00091	0,3528	0,3564	0,6000	0,1181
0,3	$P4\_V$	0,00034	0,00038	0,3004	0,3019	0,9617	0,0764
0,4	$P1$	0,00092	0,00094	0,3983	0,4020	0,9426	0,1199
0,4	$P2\_V$	0,00041	0,00046	0,4006	0,4024	0,9628	0,0835
0,4	$P3$	0,00096	0,00110	0,4657	0,4700	0,4979	0,1297
0,4	$P4\_V$	0,00041	0,00046	0,4005	0,4023	0,9606	0,0836
0,5	$P1$	0,00095	0,00098	0,4974	0,5012	0,9521	0,1225
0,5	$P2\_V$	0,00044	0,00050	0,4978	0,4998	0,9617	0,0875
0,5	$P3$	0,00108	0,00124	0,5859	0,5907	0,2894	0,1380
0,5	$P4\_V$	0,00045	0,00050	0,4979	0,4999	0,9606	0,0878
0,6	$P1$	0,00090	0,00093	0,6071	0,6107	0,9553	0,1195
0,6	$P2\_V$	0,00046	0,00052	0,6080	0,6100	0,9521	0,0892
0,6	$P3$	0,00112	0,00132	0,7126	0,7178	0,1617	0,1426
0,6	$P4\_V$	0,00048	0,00052	0,6080	0,6100	0,9543	0,0896
0,7	$P1$	0,00076	0,00081	0,7053	0,7084	0,9553	0,1114
0,7	$P2\_V$	0,00044	0,00047	0,7070	0,7088	0,9543	0,0842
0,7	$P3$	0,00108	0,00126	0,8242	0,8292	0,0660	0,1390
0,7	$P4\_V$	0,00045	0,00046	0,7069	0,7087	0,9553	0,0843
0,75	$P1$	0,00071	0,00073	0,7508	0,7537	0,9553	0,1057
0,75	$P2\_V$	0,00037	0,00042	0,7526	0,7542	0,9553	0,0790
0,75	$P3$	0,00096	0,00117	0,8771	0,8816	0,0309	0,1341
0,75	$P4\_V$	0,00038	0,00040	0,7525	0,7541	0,9500	0,0786
0,8	$P1$	0,00060	0,00062	0,7990	0,8015	0,9543	0,0978
0,8	$P2\_V$	0,00030	0,00036	0,8009	0,8023	0,9564	0,0729
0,8	$P3$	0,00081	0,00108	0,9353	0,9396	0,0074	0,1286
0,8	$P4\_V$	0,00031	0,00034	0,8008	0,8021	0,9543	0,0720
0,9	$P1$	0,00035	0,00035	0,9008	0,9021	0,9415	0,0727
0,9	$P2\_V$	0,00017	0,00022	0,9042	0,9051	0,9340	0,0556
0,9	$P3$	0,00049	0,00085	1,0520	1,0554	0,0000	0,1143
0,9	$P4\_V$	0,00017	0,00019	0,9040	0,9047	0,9351	0,0537

Table 7 Variance, confidence intervals, coverage and lenght scenario 3 considering the different quantiles and estimators

Quantile	Method	Theoretical Variance	Variance	Lower bound	Upper bound	Coverage	Length
0,1	$P1$	0,00036	0,00035	0,0995	0,1009	0,9360	0,0734
0,1	$P2\_V$	0,00010	0,00011	0,0811	0,0815	0,5360	0,0410
0,1	$P3$	0,00021	0,00021	0,0854	0,0862	0,7740	0,0571
0,1	$P4\_V$	0,00010	0,00011	0,0819	0,0824	0,5710	0,0417
0,2	$P1$	0,00059	0,00063	0,1988	0,2013	0,9550	0,0979
0,2	$P2\_V$	0,00018	0,00020	0,1505	0,1513	0,0700	0,0548
0,2	$P3$	0,00038	0,00041	0,1700	0,1716	0,6670	0,0793
0,2	$P4\_V$	0,00019	0,00021	0,1527	0,1535	0,1020	0,0560
0,25	$P1$	0,00073	0,00074	0,2516	0,2545	0,9540	0,1065
0,25	$P2\_V$	0,00023	0,00025	0,1931	0,1941	0,0380	0,0615
0,25	$P3$	0,00048	0,00050	0,2149	0,2168	0,5920	0,0871
0,25	$P4\_V$	0,00025	0,00026	0,1958	0,1968	0,0700	0,0624
0,3	$P1$	0,00083	0,00083	0,3007	0,3040	0,9470	0,1125
0,3	$P2\_V$	0,00028	0,00030	0,2373	0,2385	0,0420	0,0674
0,3	$P3$	0,00056	0,00056	0,2560	0,2582	0,5140	0,0926
0,3	$P4\_V$	0,00030	0,00030	0,2402	0,2414	0,0600	0,0677
0,4	$P1$	0,00095	0,00094	0,3998	0,4035	0,9440	0,1199
0,4	$P2\_V$	0,00034	0,00039	0,3212	0,3228	0,0230	0,0767
0,4	$P3$	0,00066	0,00067	0,3386	0,3412	0,3470	0,1016
0,4	$P4\_V$	0,00036	0,00038	0,3249	0,3263	0,0310	0,0759
0,5	$P1$	0,00095	0,00097	0,4986	0,5024	0,9570	0,1222
0,5	$P2\_V$	0,00041	0,00047	0,4071	0,4090	0,0130	0,0837
0,5	$P3$	0,00075	0,00075	0,4211	0,4240	0,1980	0,1073
0,5	$P4\_V$	0,00041	0,00044	0,4113	0,4130	0,0160	0,0818
0,6	$P1$	0,00091	0,00093	0,6068	0,6104	0,9580	0,1193
0,6	$P2\_V$	0,00046	0,00056	0,5083	0,5105	0,0210	0,0899
0,6	$P3$	0,00078	0,00080	0,5108	0,5139	0,0670	0,1107
0,6	$P4\_V$	0,00046	0,00049	0,5129	0,5148	0,0120	0,0859
0,7	$P1$	0,00081	0,00081	0,7045	0,7076	0,9440	0,1114
0,7	$P2\_V$	0,00046	0,00060	0,6002	0,6026	0,0120	0,0912
0,7	$P3$	0,00071	0,00071	0,5880	0,5908	0,0070	0,1045
0,7	$P4\_V$	0,00044	0,00048	0,6052	0,6071	0,0050	0,0848
0,75	$P1$	0,00073	0,00073	0,7495	0,7524	0,9360	0,1059
0,75	$P2\_V$	0,00045	0,00061	0,6491	0,6515	0,0210	0,0907
0,75	$P3$	0,00061	0,00062	0,6233	0,6257	0,0000	0,0976
0,75	$P4\_V$	0,00042	0,00044	0,6541	0,6558	0,0070	0,0820
0,8	$P1$	0,00059	0,00063	0,7983	0,8008	0,9470	0,0980
0,8	$P2\_V$	0,00043	0,00062	0,7046	0,7070	0,0360	0,0899
0,8	$P3$	0,00050	0,00052	0,6587	0,6608	0,0000	0,0897
0,8	$P4\_V$	0,00039	0,00040	0,7094	0,7110	0,0030	0,0784
0,9	$P1$	0,00035	0,00035	0,9001	0,9015	0,9380	0,0727
0,9	$P2\_V$	0,00038	0,00064	0,8342	0,8367	0,0900	0,0852
0,9	$P3$	0,00031	0,00032	0,7323	0,7335	0,0000	0,0698
0,9	$P4\_V$	0,00031	0,00030	0,8381	0,8393	0,0230	0,0675

Table 8 Variance, confidence intervals, coverage and lenght scenario 4 considering the different quantiles and estimators

Quantile	Method	Theoretical variance	Variance	Lower bound	Upper bound	Coverage	Length
0,1	$P1$	0,00047	0,00049	0,0988	0,1008	0,9386	0,0859
0,1	$P2\_V$	0,00013	0,00014	0,0967	0,0972	0,9305	0,0460
0,1	$P3$	0,00015	0,00018	0,0955	0,0962	0,9305	0,0518
0,1	$P4\_V$	0,00012	0,00013	0,0963	0,0968	0,9245	0,0443
0,2	$P1$	0,00080	0,00092	0,1976	0,2012	0,9507	0,1181
0,2	$P2\_V$	0,00022	0,00026	0,1945	0,1955	0,9416	0,0632
0,2	$P3$	0,00028	0,00036	0,1919	0,1933	0,9325	0,0737
0,2	$P4\_V$	0,00020	0,00024	0,1939	0,1948	0,9335	0,0604
0,25	$P1$	0,00097	0,00111	0,2502	0,2546	0,9537	0,1301
0,25	$P2\_V$	0,00027	0,00032	0,2464	0,2476	0,9486	0,0700
0,25	$P3$	0,00036	0,00045	0,2430	0,2447	0,9366	0,0830
0,25	$P4\_V$	0,00025	0,00029	0,2456	0,2468	0,9446	0,0666
0,3	$P1$	0,00108	0,00127	0,2988	0,3038	0,9617	0,1393
0,3	$P2\_V$	0,00031	0,00037	0,2944	0,2958	0,9476	0,0751
0,3	$P3$	0,00042	0,00053	0,2903	0,2924	0,9325	0,0904
0,3	$P4\_V$	0,00028	0,00033	0,2935	0,2948	0,9386	0,0711
0,4	$P1$	0,00117	0,00154	0,3974	0,4034	0,9758	0,1538
0,4	$P2\_V$	0,00035	0,00045	0,3911	0,3928	0,9486	0,0831
0,4	$P3$	0,00050	0,00070	0,3857	0,3884	0,9335	0,1032
0,4	$P4\_V$	0,00031	0,00040	0,3900	0,3915	0,9436	0,0780
0,5	$P1$	0,00119	0,00175	0,4954	0,5023	0,9819	0,1637
0,5	$P2\_V$	0,00035	0,00051	0,4880	0,4900	0,9527	0,0884
0,5	$P3$	0,00055	0,00084	0,4812	0,4845	0,9406	0,1133
0,5	$P4\_V$	0,00031	0,00044	0,4867	0,4885	0,9456	0,0820
0,6	$P1$	0,00099	0,00193	0,6043	0,6118	0,9960	0,1720
0,6	$P2\_V$	0,00033	0,00055	0,5949	0,5970	0,9627	0,0920
0,6	$P3$	0,00055	0,00098	0,5867	0,5906	0,9466	0,1225
0,6	$P4\_V$	0,00029	0,00046	0,5933	0,5951	0,9517	0,0840
0,7	$P1$	0,00080	0,00198	0,7015	0,7093	0,9980	0,1745
0,7	$P2\_V$	0,00029	0,00054	0,6929	0,6950	0,9688	0,0905
0,7	$P3$	0,00049	0,00097	0,6849	0,6887	0,9658	0,1218
0,7	$P4\_V$	0,00025	0,00044	0,6915	0,6932	0,9587	0,0818
0,75	$P1$	0,00074	0,00197	0,7469	0,7547	0,9980	0,1741
0,75	$P2\_V$	0,00027	0,00050	0,7400	0,7420	0,9748	0,0878
0,75	$P3$	0,00044	0,00089	0,7329	0,7364	0,9718	0,1167
0,75	$P4\_V$	0,00024	0,00041	0,7388	0,7404	0,9587	0,0790
0,8	$P1$	0,00065	0,00194	0,7965	0,8041	0,9980	0,1727
0,8	$P2\_V$	0,00024	0,00047	0,7907	0,7926	0,9819	0,0844
0,8	$P3$	0,00038	0,00081	0,7848	0,7880	0,9859	0,1114
0,8	$P4\_V$	0,00022	0,00037	0,7897	0,7911	0,9678	0,0752
0,9	$P1$	0,00037	0,00184	0,8974	0,9046	0,9990	0,1682
0,9	$P2\_V$	0,00014	0,00037	0,8962	0,8976	0,9960	0,0752
0,9	$P3$	0,00024	0,00064	0,8922	0,8947	0,9960	0,0992
0,9	$P4\_V$	0,00013	0,00027	0,8958	0,8968	0,9930	0,0644

Table 9 Ratio P90/P10 considering 4 scenarios and different estimators proposed

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
$P1$	0,3274	0,5460	0,3316	0,5515	0,3230	0,5400	0,3442	0,5608
$P1\_C1$	0,3283	0,5474	0,3308	0,5501	0,3222	0,5388	0,3425	0,5580
$P1\_C21$	0,3272	0,5457	0,3310	0,5504	0,3225	0,5393	0,3432	0,5592
$P1\_C22$	0,3278	0,5467	0,3314	0,5511	0,3225	0,5393	0,3426	0,5582
$P2\_V$	0,3277	0,5465	0,3303	0,5492	0,3268	0,5464	0,3443	0,5609
$P2\_V\_C1$	0,3282	0,5472	0,3301	0,5489	0,3228	0,5398	0,3430	0,5589
$P2\_V\_C21$	0,3277	0,5464	0,3300	0,5488	0,3227	0,5397	0,3439	0,5604
$P2\_V\_C22$	0,3280	0,5469	0,3303	0,5493	0,3227	0,5397	0,3438	0,5602
$P3$	0,3272	0,5457	0,3251	0,5410	0,3296	0,5508	0,3450	0,5621
$P3\_C1$	0,3282	0,5473	0,3302	0,5491	0,3224	0,5391	0,3430	0,5590
$P3\_C21$	0,3273	0,5458	0,3308	0,5501	0,3220	0,5385	0,3437	0,5600
$P3\_C22$	0,3277	0,5465	0,3314	0,5512	0,3229	0,5399	0,3435	0,5597
$P4\_V$	0,3277	0,5465	0,3304	0,5494	0,3266	0,5462	0,3443	0,5610
$P4\_V\_C1$	0,3282	0,5472	0,3302	0,5490	0,3227	0,5397	0,3430	0,5590
$P4\_V\_C21$	0,3277	0,5464	0,3301	0,5490	0,3227	0,5397	0,3440	0,5605
$P4\_V\_C22$	0,3280	0,5469	0,3304	0,5495	0,3227	0,5397	0,3440	0,5605

county of the company (100 counties in total). In summary, in the first scenario the probability of participating is a logistic function of the auxiliary variable taxes and we are in the case of an ignorable mechanism where all units are eligible. In the second case, the probability of participating depends on the SPAM variable, but the function is not continuous. In the third there is also a coverage bias since there are units that cannot be selected, and in the fourth we consider a mechanism that depends on the variable of interest, that is, we consider a non-ignorable mechanism.

For each sample, estimations of the distribution function $F(t)$ were obtained by each of the estimators included in the simulation study, at 11 different points, namely the quantiles for $\alpha = 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8$ and 0.9.

To compare the results we will use the average of the relative bias and the average of the relative mean squared error:

$$RB(t) = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{F}(t)_b - F_y(t)}{F_y(t)}, \qquad MSE(t) = \frac{1}{B}\sum_{b=1}^{B}(\hat{F}(t)_b - F_y(t))^2$$

$$AVRB(t) = \frac{1}{11}\sum_{q=1}^{11}|RB(t_q)|, \qquad AVMSE(t) = \frac{1}{11}\sum_{q=1}^{11}|MSE(t_q)|$$

In the first simulation study we are going to compare the first 4 proposed estimators ($P1, P2, P3, P4$), considering the Valliant method as a technique to obtain the weights from the propensities. We consider a sample size for the probability sample of 250, and in the case of stratified sampling $n_A = 300$ by rounding up the sizes of the strata, and for the non-probability sample of n_B=250, 500 and 1000 and we compare the values obtained from the bias and the mean square error obtaining the results that can be seen in the Table 1.

Looking at Table 1 we can see that in scenario 1 all the proposed estimators work very well, obtaining a small mean square error in all cases regardless of the sample size. If we look at the value of the bias, the best estimator is $P1$. In scenario 2 we see that the best estimator is the first one obtaining the lowest value of the mean square error $P2$ and $P4$. In scenario 3 we find that the first estimator is the best, both considering the bias and the mean square error. Finally, in scenario 4 we see that the best estimator is the first one obtaining the lowest values of the mean square error $P2$ and $P4$.

In the Appendix you can see the graphs of the 4 scenarios in which the boxplots of the 4 estimators proposed for each of the quantiles considered are represented. In these graphics we have represented with horizontal lines the α values to see what estimator of the 4 considered best fits. In scenario 1, we see how $P1$ fits very well in all cases, but has a greater dispersion in its values compared to the rest of the estimators. Starting at the value $\alpha = 0.5$, all the estimators fit very well, with lower variability for higher values of α. In scenario 2 we find that the estimators $P1$, $P2$, and $P4$ approximate the values of α very well, however $P3$ greatly overestimates the value, also having greater variability in its values. In scenario 3 we find a large number of outliers, especially in the central values of α. In this case $P1$ is the only estimator that closely approximates the values of α, since the rest of the estimators underestimate

Table 10 Ratio P80/P20 considering 4 scenarios and different estimators proposed

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE	AVRB	AVMSE
$P1$	0,3002	0,5035	0,2939	0,5001	0,2865	0,4899	0,3169	0,5188
$P1\_C1$	0,3018	0,5060	0,2915	0,4963	0,2848	0,4872	0,3137	0,5137
$P1\_C21$	0,2999	0,5030	0,2943	0,5007	0,2863	0,4896	0,3150	0,5157
$P1\_C22$	0,3009	0,5047	0,2948	0,5015	0,2862	0,4895	0,3140	0,5140
$P2\_V$	0,3008	0,5043	0,2935	0,4995	0,2929	0,5007	0,3169	0,5191
$P2\_V\_C1$	0,3016	0,5056	0,2922	0,4975	0,2851	0,4878	0,3146	0,5154
$P2\_V\_C21$	0,3007	0,5042	0,2933	0,4990	0,2851	0,4877	0,3163	0,5180
$P2\_V\_C22$	0,3012	0,5051	0,2937	0,4998	0,2854	0,4883	0,3161	0,5177
$P3$	0,2999	0,5030	0,2837	0,4845	0,2971	0,5066	0,3182	0,5210
$P3\_C1$	0,3017	0,5058	0,2919	0,4970	0,2850	0,4876	0,3147	0,5155
$P3\_C21$	0,3000	0,5031	0,2935	0,4994	0,2858	0,4889	0,3159	0,5173
$P3\_C22$	0,3007	0,5044	0,2944	0,5010	0,2871	0,4911	0,3156	0,5168
$P4\_V$	0,3008	0,5043	0,2935	0,4996	0,2926	0,5002	0,3169	0,5191
$P4\_V\_C1$	0,3016	0,5056	0,2921	0,4973	0,2851	0,4877	0,3147	0,5156
$P4\_V\_C21$	0,3007	0,5042	0,2934	0,4992	0,2851	0,4878	0,3165	0,5183
$P4\_V\_C22$	0,3012	0,5051	0,2938	0,5000	0,2855	0,4883	0,3164	0,5182

it. On scenario 4 we find that all estimators work very well, being $P4$ the one that presents a variability somewhat higher than the rest in all cases.

In the second simulation study we compare the 4 estimators, the same as in the previous case, but this time we also introduce auxiliary information and calculate the proposed calibration estimators ($C1$, $C21$ and $C22$). In the case of $C21$ proposed calibration estimator we are going to use the values of the median and the maximun and $C22$ we use the quartiles and the maximum. In this case we set the size of the non-probability sample at 500. The results can be seen in the Table 2.

In general, for all the scenarios considered, if we observe the values of the bias and the mean square error, we see that using calibration we obtain lower values, that is, we obtain more efficient estimators, as we expected.

In the third simulation study we are going to compare the different techniques for calculating weights through propensities, such as Valliant weights (V), Schonlau and Couper weights (SC), Lee weights (L), Valliant and Dever weights (VD), kernel weighting (KW) and tree-based inverse propensity weighting (TrIPW) using a general linear model. The results of the bias and mean square error can be seen in the Table 3.

Regarding the different weights considered to carry out the estimates, we see that they all work quite well, obtaining a lower bias value in TrIPW in all scenarios. If we look at the value of the mean square error we see that in scenario 1 there are hardly any differences between all the weights considered. For scenario 2 the weight of Valliant is slightly lower, for scenario 3 the weight of TrIPW and for scenario 4 the weight of Valliant and Dever.

In the fourth simulation study we are going to compare various machine learning techniques, such as gradient boosting machine (gbm), neural networks (nnet) and k-nearest neighbors (kknn) in addition to the general linear model (glm) considering how techniques for calculating weights Valliant weights. The results of the bias and mean square error can be seen in the Table 4

If we focus on scenario 1 we see that machine learning techniques work slightly better than glm, the opposite occurs in scenario 2 where glm is slightly better than the machine learning techniques considered. For scenario 3 and 4, both the machine learning and glm techniques work quite well.

We also performed a simulation study with the purpose of analyzing the resampling variance. In tables 5, 6, 7 and 8 we can see the theoretical variance obtained with the Monte Carlo method and the variance obtained through the bootstap method along with the resampling confidence intervals, coverage and length of these of each of the quantiles and proposed methods in each of the scenarios.

In view of the results in scenario 1, we see that the resampling variance is quite similar to the theoretical one. In the case of $P1$ we obtain a greater variance value, so the interval length and coverage are higher. From $\alpha = 0.5$ we get a coverage of more than 80% for all estimators. On scenario 2, the theoretical variance is similar in all estimates except for $P3$ that we find a greater difference, even making the confidence interval for $\alpha = 0.9$ exceeds the value of 1. In this case, the coverage is quite good for the other three estimators for all values of α. In scenario 3, we obtain quite similar theoretical and resampling variance values, except in $P2$. The coverage in this case is good for the estimator $P1$ but for the rest it is quite low. Finally, in scenario

4, both variances are similar, finding the greatest differences in $P1$ when the value of α increases. In this scenario, we found very good coverage, obtaining values greater than 92% in all cases.

Finally, we carry out a simulation study in which we will calculate the percentile ratios in the different proposed estimators. We start with P90/P10 of which we can see the results in the Table 9 and in the Table 10 we can see the results of the ratio P80/P20

In view of the results for the case of the P90/10 ratio we can see that for all the proposed scenarios and estimators, including the calibration estimators, we obtain very similar bias values, the same occurs for the case of the mean square error. In the case of the P80/20 ratio we found very similar values throughout all the estimators proposed for the bias and the mean square error in each of the scenarios. If we compare the scenarios we can see that in scenario 2 and 3 we obtain slightly lower values of the bias and the mean square error.

10 Conclusions

Probability methods are well established by statistical offices and researchers as one of the main tools for survey data collection, yet new data sources have emerged in recent years that could be considered to improve probability survey estimates because of their ease, speed and cost of data collection. Thus, the convenience of integrating data obtained through both sampling procedures arises.

Data integration is a new field of study with a wide range of prospective research subjects. In this paper we have addressed the problem of estimation of the distribution function based on both: a probability and a nonprobability sample, when the study variables are measured in both samples. As a result, sampling variance affects the probability sample, whereas selection bias affects the non-probability sample. Our goal is to efficiently combine both the non-probability and probability samples to estimate the distribution function. To do so, we have proposed several methods of estimating the distribution function based on different methodologies, which give rise to different estimators, and we have studied the properties that each estimator has so that they are genuine distribution functions and can thus be used as a basis for defining complex estimators such as quantiles and poverty measures.

Of the proposed estimators, the first one, is simpler as it does not have to calculate the propensities to carry out the estimation. This estimator performs quite well in all the simulations carried out. By introducing auxiliary variables and carrying out calibration estimators, we see that in general they work better than if we do not take them into account. When comparing the different weights considered in the simulation study we see that the results of all of them are very similar, finding certain improvements in the weight of Valliant and TrIPW. We also did not find many differences in terms of the machine learning technique used, but in certain scenarios they work better than general linear regression models.

Among the proposed estimators, the first always satisfies all the distribution function properties. The rest of the proposed estimators also satisfy the distribution function properties under mild conditions. This allows the estimators proposed in this

study to be directly used in quantile and percentile ratio estimation and therefore the techniques proposed here can be used in the estimation of the measurement of wage inequality. The simulation study shows that the proposed estimators are also a reliable alternative for estimating wage inequality combining the information from both the non-probability and probability samples.

This study has some limitations. The proposed data integration methods employ explicit assumptions for the outcome regression model or sample selection model. Failure to meet these assumptions can lead to significant problems of both bias and variability in the estimates. The ignorability assumption is the most crucial assumption for the validity of the estimators based on PSA and SM adjustment although all other assumptions are also involved (Wu 2022). In practice, this assumption cannot be verified from sample data and non-probability sources often have very difficult participation mechanisms, so you should be very careful when using these methods if you are unsure of the behaviour of the selection mechanism.

Appendix

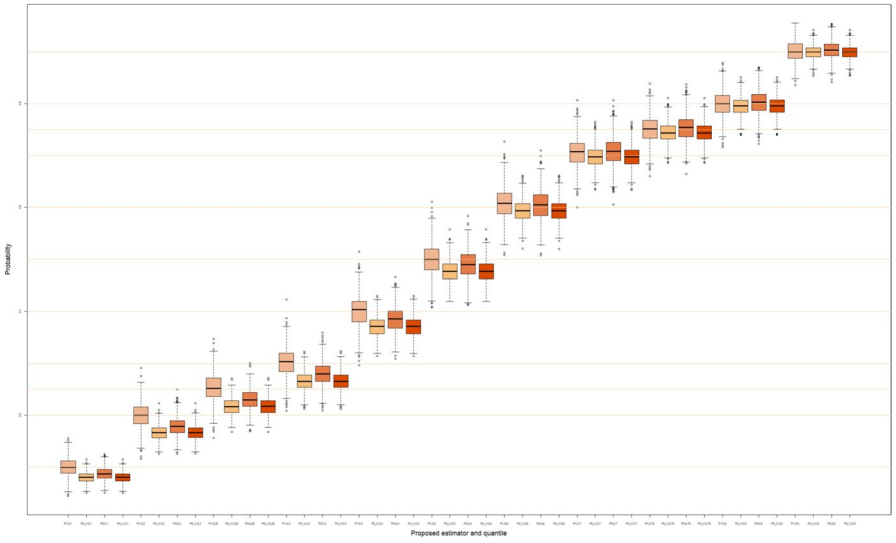

Fig. 1 Boxplots of the 4 estimators proposed for each of the quantiles considered in the scenario 1

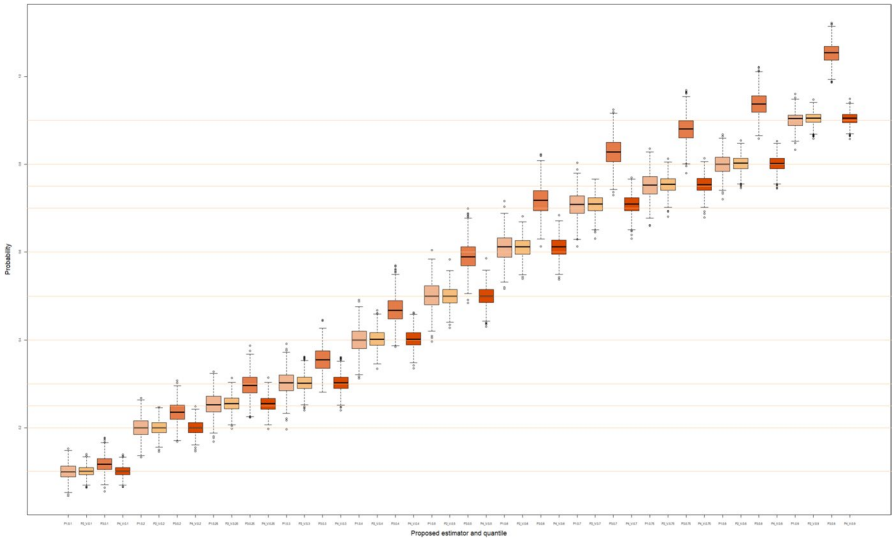

Fig. 2 Boxplots of the 4 estimators proposed for each of the quantiles considered in the scenario 2

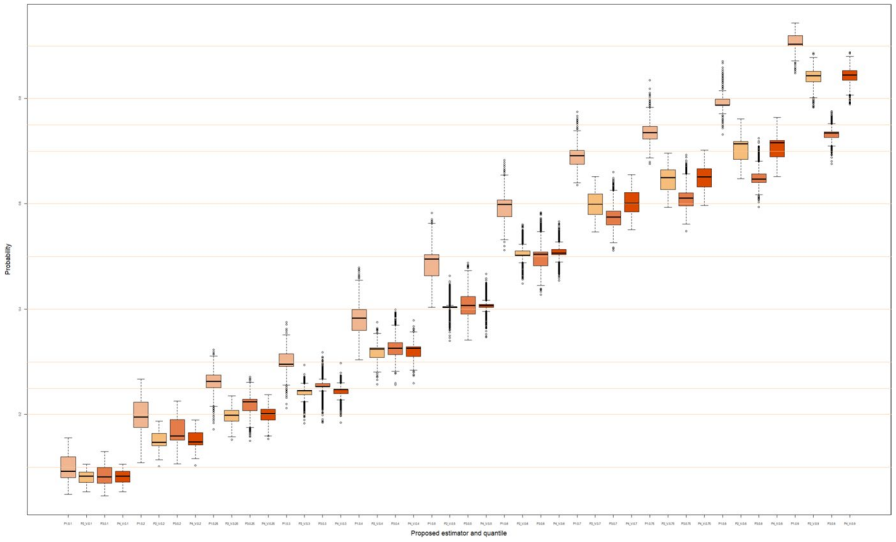

Fig. 3 Boxplots of the 4 estimators proposed for each of the quantiles considered in the scenario 3

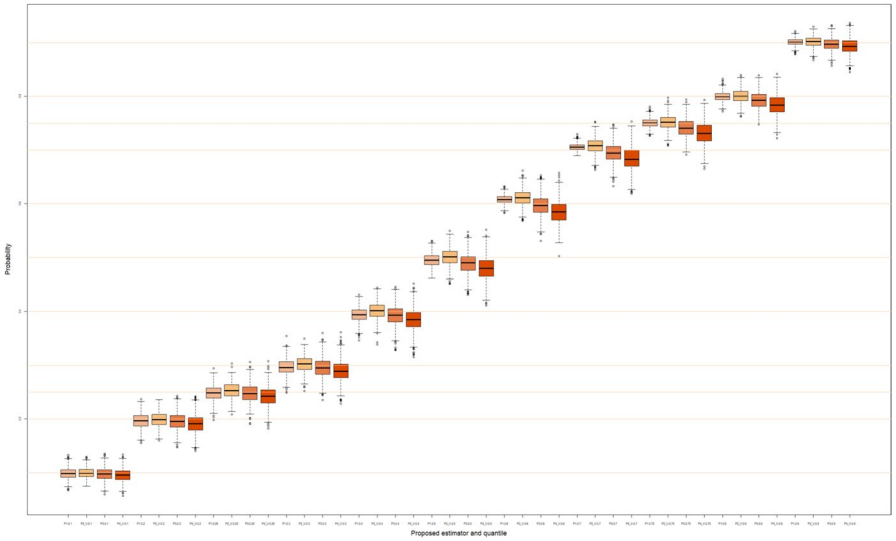

Fig. 4 Boxplots of the 4 estimators proposed for each of the quantiles considered in the scenario 4

Appendix

- **Machine Learning models used in the simulationsGradient Boosting Machine**: works as an ensemble of weak classifiers. Boosting is an iterative process that trains subsequent models giving more importance to the data for which previous models failed. This idea can be interpreted as an optimization problem Breiman (1997) and, therefore, it is suitable for the gradient descent algorithm Friedman (2001). This algorithm allows us to converge towards the minimum value of a function (usually a loss function) by an iterative process.

$$\hat{\pi}_{vi} = v^T J\left(\mathrm{x}_i\right), i \in s_v$$

  - $J\left(\mathrm{x}_i\right)$ stands for a matrix of terminal nodes of m decision trees used for the boosting.
  - v is a vector representing the weight of each tree.

- **k-Nearest Neighbors**: "one of the most fundamental and simple classification methods" Peterson (2009). The algorithm doesn't need training, as simply averages the value of the target variable for the k individuals closer to the estimated individual (its k nearest neighbors), given a certain distance dependent on the covariates.

$$\hat{\pi}_{vi} = \frac{\sum_{j \in s_i^k} y_j}{k}, i \in s_v$$

– $s_i^k = \left\{ j \in s / d(\mathrm{x}_i, \mathrm{x}_j) \leq d(\mathrm{x}_i, \mathrm{x}_{(k)}) \right\}$ and $x_{(1)}, \ldots, x_{(n-1)}$ are, respectively, the closest and the furthest individual to x_i according to the distance d. Choosing the right k is important for the proper performance of the algorithm.

- **Neural networks**: the inputs follow an iterative process through one or more hidden layers until reaching the last layer, which produces the final output. The weights are initialized randomly and then optimized via gradient descent with the backpropagation algorithm Rumelhart et al. (1986) which looks for the minimization of a function, usually of a lost function.

$$\hat{\pi}_{vi} = g\left(\sum_{i=1}^{L} v_i f_i(\cdot) + b \right)$$

– g and f_i stand for the activation functions
– v_i are the weights of the i-th neuron of the hidden layer
– b is the activation threshold

Acknowledgements The research was partially supported by MCIN/AEI /10.13039/501100011033, PDC2022-133293-I00, Spain, Strategic Action in Health (DTS23/00032, Spain) and from IMAG-María de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.

References

Acal C, Ruiz-Castro JE, Aguilera AM, Jiménez-Molinos F, Roldán JB (2019) Phase-type distributions for studying variability in resistive memories. J Comput Appl Math 345:23–32

Beaumont JF (2020) Are probability surveys bound to disappear for the production of official statistics? Surv Methodol 46(1):1–29

Bohn MK, Higgins V, Kavsak P, Hoffman B, Adeli K (2019) High-sensitivity generation 5 cardiac troponin t sex-and age-specific 99th percentiles in the caliper cohort of healthy children and adolescents. Clin Chem 65(4):589–591

Bradley VC, Kuriwaki S, Mea Isakov (2021) Unrepresentative big surveys significantly overestimated us vaccine uptake. Nature 600:695–700. https://doi.org/10.1038/s41586-021-04198-4

Breiman L (1997) Arcing the edge. Tech Report. 486

Buelens B, Burger J, Brakel JA (2018) Comparing inference methods for non-probability samples. Int Stat Rev 86(2):322–343

Burtless G (1999) Effects of growing wage disparities and changing family composition on the us income distribution. Eur Econ Rev 43(4–6):853–865

Chen Y, Li P, Wu C (2019) Doubly robust inference with nonprobability survey samples. J Am Stat Assoc 115(532):2011–2021

Chu KCK, Beaumont JF (2019) The use UF classification trees to reduce selection bias for a non-probability sample with help from a probability sample, Calgary, AB, Canada. In Proceedings of the Survey Methods Section: SSC Annual Meeting

Darvas Z (2019) Why is it so hard to reach the Eu's poverty target? Soc Indic Res 141(3):1081–1105

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87(418):376–382. https://doi.org/10.1080/01621459.1992.10475217

Deville J, Särndal C, Sautory O (1993) Generalized raking procedures in survey sampling. J Am Stat Assoc 88:1013–1020
Dickens R, Manning A (2004) Has the national minimum wage reduced UK wage inequality? J R Stat Soc A Stat Soc 167(4):613–626
Disogra C, Cobb CL, Chan EK, Dennis JM (2011) Calibrating non-probability internet samples with probability samples using early adopter characteristics. In Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM)
Elliot MR (2009) Combining data from probability and non-probability samples using pseudo-weights. Surv Prac. https://doi.org/10.29115/SP-2009-0025
Elliott M, Haviland A (2007) Use of a web-based convenience sample to supplement a probability sample. Surv Methodol 33:211–215
Eurostat Experimental statistics: income inequality and poverty indicators. https://ec.europa.eu/eurostat/web/experimental-statistics/income-inequality-and-poverty-indicators. Online (2022)
Eurostat Products Datasets: inequality of income distribution S80/S20 income quintile share ratio - EU-SILC and ECHP surveys. https://ec.europa.eu/eurostat/web/products-datasets/-/ilc_pns4. Online (2022)
Ferri-García R, Rueda MdM (2018) Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. SORT 1:159–162
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat, 1189–1232
Gelman A, Kenworthy L, Su Y-S (2010) Income inequality and partisan voting in the united states. Soc Sci Quart, 1203–1219
Goga C, Ruiz-Gazen A (2014) Efficient estimation of non-linear finite population parameters by using non-parametrics. J R Stat Soc Ser B Stat Methodol 76(1):113–140
Guio A-C, Marlier É, Nolan B (2021) Improving the understanding of poverty and social exclusion in Europe. Publications Office of the European Union, Luxembourg
Gutiérrez Rojas, H.A (2020) Selection of samples and parameter estimation in finite population. https://CRAN.R-project.org/package=TeachingSampling (Version 4.1.1)
Jones AF, Weinberg DH (2000) The changing shape of the nation's income distribution, 1947–1998 vol. 204. Department of Commerce, Economics and Statistics Administration, US
Kim JK, Wang Z (2019) Sampling techniques for big data analysis. Int Stat Rev 87:177–191
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
Kimbro RT, Brooks-Gunn J, McLanahan S (2011) Young children in urban areas: links among neighborhood characteristics, weight status, outdoor play, and television watching. Soc Sci Med 72(5):668–676
Kott P.S, Liao D (2012) Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. Surv Res Methods 6:105–111
Lee S, Valliant R (2009) Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociol Methods Res 37:319–343
Machin S, Manning A, Rahman L (2003) Where the minimum wage bites hard: introduction of minimum wages to a low wage sector. J Eur Econ Assoc 1(1):154–180
Martínez S, Rueda M, Arcos A, Martínez H (2010) Optimum calibration points estimating distribution functions. J Comput Appl Math 233(9):2265–2277. https://doi.org/10.1016/j.cam.2009.10.011
Martínez S, Rueda M, Arcos A, Martínez H, Muñoz JF (2012) On determining the calibration equations to construct model-calibration estimators of the distribution function. Revista Matemática Complutense 25(1):87–95. https://doi.org/10.1007/s13163-010-0058-z
Martínez S, Rueda MdM, Martínez H, Arcos A (2015) Determining P optimum calibration points to construct calibration estimators of the distribution function. J Comput Appl Math 275:281–293
Martínez S, Rueda M, Illescas M (2020) The optimization problem of quantile and poverty measures estimation based on calibration. J Comput Appl Math, 113054
Meyer BD, Sullivan JX (2012) Identifying the disadvantaged: official poverty, consumption poverty, and the new supplemental poverty measure. J Econo Perspect 26(3):111–36
Peterson L (2009) K-nearest neighbor. Scholarpedia 4(2):1883
Rafei A, Elliott MR, Flannagan CAC (2022) Robust and efficient Bayesian inference for non-probability samples. Preprint at arXiv: 2203.14355 (2022)
Rivers D (2007) Sampling for web surveys. In: Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA

Robbins MW, Ghosh-Dastidar B, Ramchand R (2020) Blending probability and nonprobability samples with applications to a survey of military caregivers. J Surv Statist Methodol 9(5):1114–1145. https://doi.org/10.1093/jssam/smaa037 (https://academic.oup.com/jssam/article-pdf/9/5/1114/41727173/smaa037.pdf)

Rueda M, Martínez S, Martínez H, Arcos A (2007) Estimation of the distribution function with calibration methods. J Statist Plann Inference 137(2):435–448

Rueda MDM, Amo S, Cobo B, Castro-Martín L, Ferri-García R (2022a) 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. Biometric J. https://doi.org/10.1002/bimj.202200035

Rueda M.d.M, Martínez-Puertas S, Castro-Martín L (2022b) Methods to counter self-selection bias in estimations of the distribution function and quantiles. Mathematics 10:4726. https://doi.org/10.3390/math10244726

Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536

Schonlau M, Couper MP (2017) Options for conducting web surveys. Stat Sci 32(2):279–292

Shrider EA, Kollar M, Chen F, Semega J, et al (2021) Income and poverty in the United States: 2020. US Census Bureau, Current Population Reports (P60-273)

Silva PN, Skinner CJ (1995) Estimating distribution functions with auxiliary information using poststratification. J Off Stat 11(3):277

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

Valliant R, Dever JA (2011) Estimating propensity adjustments for volunteer web surveys. Sociol Methods Res 40(1):105–137

Wang L, Graubard BI, Katk HA, Li Y (2020) Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. J R Stat Soc A Stat Soc 183(3):1293–1311. https://doi.org/10.1111/rssa.12564

Wilson R, Fleming ZL, Monks P, Clain G, Henne S, Konovalov I, Szopa S, Menut L (2012) Have primary emission reduction measures reduced ozone across Europe? an analysis of European rural background ozone trends 1996–2005. Atmos Chem Phys 12(1):437–454

Wiśniowski A, Sakshaug J.W, Pérez Ruiz D.A, Blom A.G (2020) Integrating probability and nonprobability samples for survey inference. J Surv Statist Methodol 8(1):120–147. https://doi.org/10.1093/jssam/smz051 (https://academic.oup.com/jssam/article-pdf/8/1/120/33387851/smz051.pdf)

Wolter K (2007) Introduction to variance estimation. Springer, New York, p 427

Wu C (2022) Statistical inference with non-probability survey samples. Surv Methodol 48(2):284