AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-025-00530-9

ORIGINAL PAPER

Evaluation of available techniques and their combinations to address selection bias in nonprobability surveys

Jorge Luis Rueda-Sánchez[1] · **Ramón Ferri-García**[1] · **María del Mar Rueda**[1] · **Beatriz Cobo**[2]

Received: 7 May 2024 / Accepted: 19 May 2025

Abstract

New survey methodologies that often produce nonprobability samples have recently become very important. However, estimates from nonprobability samples can be subject to selection bias, which is primarily caused by the lack of coverage and the respondent's ability to decide whether or not to participate in the survey. In such cases, inclusion probabilities can be zero or unknown. When this happens, the estimators normally used in sample surveys are useless, and we must employ methods to reduce this bias. There is a wide variety of techniques to achieve this which depend on the auxiliary information available, but no study has determined which is better among all. In this paper, we briefly explain most of these methods and conduct an extended study to compare their performances. We will study superpopulation models, which require knowledge of the auxiliary variables of all individuals in the population, linear calibration, which requires the population totals of the covariates, and several techniques that use a reference probability sample, such as propensity score adjustment, propensity-adjusted probability prediction, Kernel Weighting, Statistical Matching and, Doubly Robust estimators. In addition, we compare their performance using linear regression or XGBoost as a predictive model, and the design weights in estimating inclusion probabilities or not, and with or without prior variables selection. The study was performed using five different datasets to determine which technique provides accurate and reliable estimates from nonprobability samples.

Keywords Inference · Nonprobability samples · Survey sampling · Selection bias · Variable selection · Weighted models

Jorge Luis Rueda-Sánchez, Ramón Ferri-García, María del Mar Rueda, and Beatriz Cobo have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 13 June 2025

1 Introduction

Probability sampling constitutes a standard procedure for obtaining reliable estimates of population figures since the works of Jerzy Neyman et al. (1934); Horvitz and Thompson (1952). For many years, face-to-face surveys, postal surveys and telephone surveys were the main procedures used to ensure a probability sampling in the study of a population. However, new questionnaire administration and data collection methods are proving themselves as being able to provide larger sample sizes in less time and at lower costs, making them attractive to researchers. In the majority of cases, the samples are drawn following a nonprobability sampling. Volunteering on social media websites, analysis of social media users, online opt-in panels, and e-mail surveys are some examples. Such surveys have even been considered for the production of official statistics in the last few years (Beaumont 2020).

In a probability sampling design, each unit of the target population U must have a known inclusion probability. Nonprobability samples fail to comply with this condition because they have no sampling design that enables the computation of selection probabilities. These samples often imply selection bias caused by various mechanisms, such as coverage error or self-selection bias, which can be problematic if the characteristics of the potentially covered population, $U_{pc} \subseteq U$ (from which the nonprobability samples are drawn) differ from those of U (Elliott and Valliant 2017).

In some scenarios, this selection bias can be mitigated, completely or partially. Methods for mitigating selection bias in nonprobability samples depend on the amount of auxiliary information available. We primarily distinguish between three possible situations, and we consider the variable of interest, y, from which we want to estimate a parameter, θ_y, is only available for individuals in the nonprobability sample and not for any other unit of the (unsampled) population. The usual scenario is to have access to only the population totals for some covariates measured in the nonprobability sample. The population totals can be official figures from regional statistical offices, or estimated figures from probability samples with (ideally) little to no bias. In such cases, the usual approach is to perform calibration reweighting as defined in Deville and Särndal (1992). An alternative to the original formulation is the use of penalized methods such as ridge calibration Chen et al. (2002) and generalized calibration Kott (2006); Haziza and Beaumont (2017). Normally, population figures for certain variables (age, gender, education level, etc.) are easier to access than other sources of auxiliary information, making these methods a common choice for survey researchers.

Occasionally, a reference sample drawn using a probability sampling design is available to be used as a source of auxiliary information. This reference probability sample and the nonprobability sample must have measured a common set of covariates to assess the similarities between both samples. These covariates are auxiliary variables that are good as long as they are related to the variable of interest, or the selection mechanism, and the variable of interest (see Ferri-García and Rueda (2022) for a literature review on the matter). The reference

samples can come from multiple sources, such as official statistics microdata, government surveys, or even censuses Schonlau and Couper (2017); Elliott and Valliant (2017). The advantage of using this information is that reference samples enable the use of relevant covariates (which are related to the variable of interest) for which population totals might be rarely available. Methods for adjusting nonprobability samples using information from reference probability samples can be divided into pseudodesign-based methods which aim to estimate participation probability, such as propensity score adjustment (Lee 2006), propensity-adjusted probability prediction (Elliott and Valliant 2017), or Kernel Weighting (Kern et al. 2021), model-based methods that aim to estimate the (unmeasured) values of the variable of interest in the reference sample, such as Statistical Matching Rivers (2007) or Mass Imputation Kim et al. (2021), and a combination of both approaches, such as Doubly Robust estimators (Chen et al. 2020) and the combination of PSA and Statistical Matching proposed in Castro-Martín et al. (2022).

In some cases, a census of the full population for a set of covariates (also available in the nonprobability sample) might be available. This is the most uncommon case, although it might not be as rare for certain populations (university students, associations, etcetera). For example, it is common in ecological studies to consider areal units or sites in a landscape as population units, thus allowing researchers to access the auxiliary information of these units via satellite images Boyd et al. (2024). Censuses can even be constructed if the cross-count population totals are known for combinations of variables; however it is rare that an acceptable number of variables is available for such combinations. In the cases where a complete census is available, the estimators based on the superpopulation modeling theory formulated in Royall and Herson (1973) can be applied. These estimators include model-based approach Royall (1970), model-assisted approach Cassel et al. (1976), and model-calibrated approach Wu and Sitter (2001). They are typically applied in probability sampling contexts; however, their application in the nonprobability sampling context has shown promising results Buelens et al. (2018); Ferri-García et al. (2021).

The proposed methods can be applied in various situations, and they can also be improved with usual statistical and data science techniques, such as data preprocessing, variable selection or weight manipulation. This work compares the performance of these methods and their possible improvements using a simulation study that employs several pseudopopulations and several predictive methods where applicable. For this purpose, we introduce the methods in the following subsections.

1.1 Calibration and model-based adjustments

1.1.1 Calibration

Consider a target population U of size N, with a variable of interest y and a linear parameter that we want to estimate from it; for the sake of simplicity, we assume that this parameter is the population mean, $\overline{Y}$. Here, let s be a probability sample with a given sampling design such that each element $k \in s$ has an associated inclusion probability, $\pi_k = 1/d_k$. The element d_k is known as the design weight of

the k-th unit of the sample. The usual Horvitz-Thompson and Hajek estimators can be written as

$$\hat{\bar{Y}}^{HT} = \frac{\sum_{i \in s} d_i y_i}{N} \qquad \hat{\bar{Y}}^{Hajek} = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i}. \tag{1.1}$$

Let $\mathbf{x}$ be an auxiliary vector associated with y. We assume that population totals for each variable in the vector are known; that is, $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$ is known. Calibration reweighting aims to find a new set of weights, w_k, that minimizes the distance with d_k for $k \in s$, given a distance function $G(.,.)$ such that

$$\min_{w_k} \sum_{k \in s} G(w_k, d_k), \quad \text{subject to} \quad \sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}. \tag{1.2}$$

The above optimization procedure implies that the new set of weights must provide estimates of the population totals of $\mathbf{x}$ that are exactly equal to the actual totals $\mathbf{X}$, although some recent works consider the relaxation of this condition Kott (2006). Several $G(.,.)$ were defined in Deville and Särndal (1992), where it has also been proven that calibration reweighting provides consistent estimators under several conditions. Calibration weighting reduces bias caused by non-response Särndal et al. (1992) or coverage Folsom and Singh (2000); Kott (2006); Dever et al. (2008) biases.

When applying calibration in the nonprobability sampling context for a nonprobability sample s_v, design weights $d_k \forall k \in s_v$ are not available. In this situation, design weights may be replaced with an unitary vector such that $d_k = 1 \forall k \in s_v$. If linear calibration is applied, this will lead to post-stratification weights Smith (1991). This approach was studied in Bethlehem (2010) for web surveys and demonstrated successful results when the inclusion in the nonprobability sample is not directly related to the target variable.

1.1.2 Superpopulation modeling

Superpopulation modeling can be considered a general case of the calibration framework in where we assume that $\mathbf{y} = (y_1, ..., y_N)$ is a realization of a superpopulation $\mathbf{Y} = (Y_1, ..., Y_N)$ where the following model applies:

$$Y_i = m(\mathbf{x}_i) + e_i, i = 1, ..., N. \tag{1.3}$$

Under this framework, the values of y in the non-sampled population, $\bar{s} = U \setminus s$, can be estimated by regression modeling using a predictive model M, which represents the behavior of the superpopulation model m:

$$\hat{y}_i = E_M(y_i | \mathbf{x}_i). \tag{1.4}$$

Let t_y be the population total of the variable y, $t_y = \sum_{i=1}^{N} y_i$. The predicted values $\hat{y}_i$ can be used to estimate this total using three different approaches:

- The model-based approach Royall (1970), which adds the sums of y for sampled individuals and $\hat{y}$ for non-sampled individuals as follows:

$$\hat{t}_y^{mb} = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i. \tag{1.5}$$

  This estimator is based on strong assumptions. Among them, the ignorability of the sampling design, conditional on relevant covariates. If this condition is not fulfilled, there could be non-negligible bias (Pfeffermann 1993).
- The model-assisted approach Cassel et al. (1976), in which the sum of $\hat{y}$ for all individuals in U (sampled and non-sampled) is then corrected with the errors (differences between y and $\hat{y}$) observed in the sample, and each one is elevated via the vector of weights w, which can be expressed as the vector of the design weights or any adjusted weight:

$$\hat{t}_y^{ma} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} (y_i - \hat{y}_i) w_i. \tag{1.6}$$

  Note that if model M is linear, then this estimator is equivalent to the general regression estimator (GREG), which is a type of calibration estimator developed in Deville and Särndal (1992).
- The model-calibrated approach Wu and Sitter (2001), which is a weighted estimator constructed via calibration of the predicted values:

$$\hat{t}_y^{mc} = \sum_{i \in s} y_i w_i^{CAL} \tag{1.7}$$

  where w_i^{CAL} minimize $\sum_{i \in s} G(w_i^{CAL}, w_i)$, subject to $\sum_{i \in s} w_i^{CAL} \hat{y}_i = \sum_{i \in U} \hat{y}_i$.

These estimators can also be applied when s is a nonprobability sample, which gives successful results Buelens et al. (2018); Ferri-García et al. (2021). However, note that in this case, we still have the issue of having no design weights available, which can be problematic in the case of model-assisted and model-calibrated approaches. As in calibration, the unitary weights ($w_i = 1, \forall i \in s$) or other types of weights could be considered. In addition, these estimators require the use of complete population data for some auxiliary variables, which is rarely available to researchers.

1.2 Adjustments using a probability sample

In other situations, two samples are available. On the one hand, we have a nonprobability sample s_v of size n_v drawn from $U_v \subset U$, which represents the subset of the potentially covered population (e. g. Internet users) and has no design; thus, its inclusion probabilities, π_v, are unknown. On the other hand, we have a probability sample s_r of size n_r, drawn from U with sampling design (S_d, p_d) and design weights d^r, which is available with some common covariates with s_v, $\mathbf{x}$.

Let $R = 0, 1$ be an indicator variable of belonging to s_v, where

$$R_i = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases}, \qquad i \in U. \tag{1.8}$$

We assume the probability (propensity) that the i-th individual will be included in the nonprobability sample, π_{vi}, can be defined as a probability conditional on the available covariates $\mathbf{x}$:

$$\pi_{vi} = Pr(R_i = 1|\mathbf{x}_i, y_i) = Pr(R_i = 1|\mathbf{x}_i), \qquad i \in U. \tag{1.9}$$

This assumption is called the non-informativity assumption and implies that the bias caused by the selection mechanism that governs R can be completely removed if the probability was properly taken into account by the estimators. This is equivalent to Missing At Random (MAR) case following the classification in Little and Rubin (1987).

The strongest assumption is non-informativity, given that self-selection in samples is related to the variable of interest itself. Unfortunately, it is not possible to test this assumption in practice if we only know y for individuals in s_v (see Wu (2022)). Some authors might consider including a large and rich set of covariates that may be related to R or y to ensure the compliance of this assumption (Yang et al. 2020). However, this solution does not guarantee the non-informativity of the sample. Effects of ignoring the sample selection process when fitting models to survey data can have significant effects on the inference process, including bias of point estimators and poor performance in test statistics and confidence intervals. This topic is studied extensively in (Skinner et al. 1989) and (Pfeffermann 1993) for probability sampling.

We also assume that $\pi_{vi} > 0, \forall i \in U$, and that $R_1, ..., R_N$ are independent given $(\mathbf{x}_1, ..., \mathbf{x}_N)$. These assumptions were specified in Chen et al. (2020), and some methods for correcting selection bias can be applied when these assumptions are satisfied.

1.2.1 Propensity Score Adjustment

In a situation where π_v is not known, propensity score adjustment attempts to provide an estimate of its value for each individual in the available samples using model M as follows:

$$\hat{\pi}_{vi} = E_M[R_i^* = 1|\mathbf{x}_i], \qquad i \in s_v \cup s_r \tag{1.10}$$

where $\mathbf{x}$ is a set of covariates available in both s_v and s_r, and R^* is a proxy for R obtained from the pooled sample $s_v \cup s_r$ such that

$$R_i^* = \begin{cases} 1 & i \in s_v \\ 0 & i \in s_r \end{cases}, \qquad i \in s_v \cup s_r. \tag{1.11}$$

It is common in literature to consider a logistic regression model for M. However, some recent approaches involve Machine Learning classification algorithms for M; some comparative studies include Random Forests, Gradient Boosting Machines (GBM), k-nearest neighbors and neural networks, among other approaches

(Castro-Martín et al. 2020; Ferri-García and Rueda 2020). Recently, the XGBoost algorithm Chen et al. (2016) has gained attention owing to its efficacy in estimating propensities (Castro-Martín et al. 2021).

The propensity estimates provided by these predictive methods are used for weighting in the usual Horvitz-Thompson and Hajek estimators, using some formulas:

- Inverse probability weighting: $w_i^{IPW} = 1/\hat{\pi}_{vi}$ as noted in the review by Valliant (2020), or with the slight modification proposed in Schonlau and Couper (2017) to consider that the nonprobability sample individuals are not part of the target population, $w_i^{CIPW} = (1 - \hat{\pi}_{vi})/\hat{\pi}_{vi}$.
- Propensity stratification weighting (also known as stratification matching): the individuals are divided into g strata according to propensity (individuals with similar propensities are classified in the same strata). In the approach proposed in Lee and Valliant (2009), the original weights of the nonprobability sample (if any) are multiplied by the correction factor that takes the design weights of the probability sample into account:

$$w_i^{Strat1} = d_i^v f_c = d_i^v \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_j^v / \sum_{j \in s_v} d_j^v}, \quad i \in s_v, i \in c, \tag{1.12}$$

  where d^v is the vector of weights of the nonprobability sample (normal unitary weights may be used; however, other adjustment weights can be considered), and s_v^c and s_r^c are the individuals in the nonprobability and probability samples, respectively, that belong to the c-th propensity stratum. On the other hand, Valliant and Dever (2011) replaced the propensity of individual by the mean propensity of the stratum they belong to, and then applied the inverse probability weighting formula:

$$w_i^{Strat2} = 1/\overline{\pi}_v(c), \quad \overline{\pi}_v(c) = \frac{\sum_{j \in s_v^c} \hat{\pi}_{vj}}{n_v^c}, i \in s_v, i \in c, \tag{1.13}$$

  where n_v^c is the size of the c-th propensity stratum.

Regarding the use of design weights in PSA modeling, Chen et al. (2020) proposed pseudo-maximum likelihood estimation (PMLE) based on the following pseudo-log-likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i \in s_v} log\left(\frac{m(\mathbf{x}_i, \boldsymbol{\beta})}{1 - m(\mathbf{x}_i, \boldsymbol{\beta})}\right) + \sum_{i \in s_r} d_i log\big(1 - m(\mathbf{x}_i, \boldsymbol{\beta})\big). \tag{1.14}$$

This function is a modification of the original log-likelihood equation for the complete population for the prediction of R, where the second sum is not over $i \in s_r$ but over $i \in U$. The rationale of Chen et al. (2020) is to substitute this sum by an unbiased estimate, which can be obtained with the weighted sum of $log(1 - m(\mathbf{x}_i, \beta))$ of individuals in the probability sample.

If the participation rate in the nonprobability sample is small, i. e. $n_v/N \to 0$, the results of PMLE are approximately equivalent to those obtained by weighted logistic regression, where the regression weights d_i^{pool} are expressed as follows:

$$d_i^{pool} = \begin{cases} 1 & i \in s_v \\ d_i & i \in s_r \end{cases}, \quad i \in s_v \cup s_r. \tag{1.15}$$

Other weighting strategies in the modeling step can also be consulted in Valliant and Dever (2011), although they are unable to provide consistent estimators, as reported in Chen et al. (2020). On the other hand, PSA efficiency largely depends on the covariates used for propensity estimation Lee (2006); Valliant and Dever (2011). The variables included in a propensity score model should be related to the variable of interest y. In a given set of available covariates, some variables might be more strongly related to Y than others. Including unrelated variables could result in covariates sets in which there is no relationship between y or R and some variables of $\mathbf{x}$, and in more complex models which could lead to greater variances and, in the case of machine learning classification algorithms, to worse results. Previous studies have shown that more efficient results can be obtained when the covariates were related to the variables of interest Hirano and Imbens (2001); Brookhart et al. (2006); Ferri-García et al. (2022). However, it can sometimes be difficult to qualitatively assess which variables are related to the variable of interest. For this reason, automatic feature selection techniques can be helpful to select the relevant covariates in terms of propensity estimation. Feature selection before propensity estimation can be advantageous, given its usefulness for removing redundant or irrelevant variables that could increase bias and (especially) the variance of the final estimates Ferri-García and Rueda (2022).

1.2.2 Propensity-Adjusted Probability Prediction

The propensity-adjusted probability prediction approach (PAPP) (Rafei et al. 2020, 2022) is an approach based in the two-phase quasi-randomization method, which was introduced in Elliot (2009) and Elliott et al. (2010), and was further developed in Elliott and Valliant (2017). The proposed method also provides consistent estimators and allows using robust Bayesian inference techniques. Let $\delta = 0, 1$ be an indicator variable for the probability sample, where

$$\delta_i = \begin{cases} 1 & i \in s_r \\ 0 & i \notin s_r \end{cases}, \quad i \in U. \tag{1.16}$$

Under the PAPP approach, assuming no overlap between s_v and s_r, the probability of belonging to the nonprobability sample can be expressed (by developing the Bayes theorem) as a function of the estimated propensity and design weight:

$$\pi_{vi} = P(\delta_i = 1|\mathbf{x}_i)\frac{P(R_i^* = 1|\mathbf{x}_i)}{1 - P(R_i^* = 1|\mathbf{x}_i)}. \tag{1.17}$$

The probability $P(R_i^* = 1|\mathbf{x}_i)$ is not known in advance, meaning that it must be estimated with propensity score adjustments, such as PSA. The probability $P(\delta_i = 1|\mathbf{x}_i)$ depends on the covariates considered for the adjustment. Note that if variables $\mathbf{x}$ correspond to the design variables for sampling the probability sample, the probability $P(\delta_i = 1|\mathbf{x}_i)$ must be known for all individuals in the population, implying that this probability can be plugged-in directly without needing to estimate it. However, in the usual setting, where $\mathbf{x}$ does not fully correspond to the design variables, $P(\delta_i = 1|\mathbf{x}_i)$ can be estimated with the inverse prediction of $d_i, i \in s_r$ using $\mathbf{x}_i$ in a model MP. This prediction can also be used if, for any restriction, we do not have access to the design weights (the most frequent case occurs when sampling weights are released only for sampled units; in addition, for privacy reasons, even the sampling weights are frequently unavailable, in favor of final weights that include post-stratification and non-response correction factors). The final estimates of π_i are:

$$\hat{\pi}_{vi}^{PAPP} = \frac{1}{E_{MP}[d_i|\mathbf{x}_i]} \frac{P(R_i^* = 1|\mathbf{x}_i)}{1 - P(R_i^* = 1|\mathbf{x}_i)}. \tag{1.18}$$

The pseudo-weight to be applied in the estimators is calculated as follows:

$$w_i^{PAPP} = \frac{1}{\hat{\pi}_{vi}^{PAPP}} = E_{MP}[d_i|\mathbf{x}_i] \frac{1 - P(R_i^* = 1|\mathbf{x}_i)}{P(R_i^* = 1|\mathbf{x}_i)}. \tag{1.19}$$

Elliott and Valliant (2017) recommend performing the prediction on $1/d_i$ instead of d_i using a Beta regression given that $1/d_i \in [0, 1]$. In the present study, we consider the prediction of d_i instead, because this prediction allows the use of a wider set of modeling approaches apart from linear regression, trimming the results when necessary (i. e. when the predictions are below 1). Another possibility is to predict $log(1/d_i)$ to avoid issues related to the domain of $1/d_i$; this is done in Rafei et al. (2020), where Bayesian Additive Regression Models (BART) were used both in the prediction of $log(1/d_i)$ and $P(R_i^* = 1|\mathbf{x}_i)$.

1.2.3 Kernel Weighting

The Kernel Weighting (KW) method (Wang et al. 2020) is similar to PSA in the sense that both methods create pseudo-weights, from estimated propensities for the non-probability sample using a reference probability sample. However, they differ in their approaches to generate these new weights.

Kernel Weighting is based on using the aforementioned propensities to measure the similarity between individuals according to the distributions of the measured auxiliary variables in s_r and s_v. These similarities are used as weights for our estimator by adding a previous step where the distances are smoothed using Kernel functions.

Given a propensity model M, let

$$d(\mathbf{x}_i^{(r)}, \mathbf{x}_j^{(v)}) = \hat{\pi}(\mathbf{x}_i^{(r)}) - \hat{\pi}(\mathbf{x}_j^{(v)}) = E_M[R_i^* = 1|\mathbf{x} = \mathbf{x}_i] - E_M[R_j^* = 1|\mathbf{x} = \mathbf{x}_j], \quad i \in s_r, j \in s_v \tag{1.20}$$

be the distance between the estimated propensity score from $i \in s_r$ and $j \in s_v$. This distance must be bounded between -1 and 1. We smooth these values using a zero-centered Kernel function, which allows for several alternatives (Gaussian, Standard Gaussian, Triangular, etc.). The closer the distance is to 0, the more similar are the individuals in terms of their auxiliary variables, as propensities are estimated using these variables. In addition, the more similar individuals are, the larger proportion KW assigns to the original weight d_i^r in the estimation of the pseudo-weight for individual j. This proportion is known as the kernel weight and it can be expressed as follows (Wang et al. 2020):

$$k_{ij} = \frac{K\{d(\mathbf{x}_i^{(r)}, \mathbf{x}_j^{(v)})/h\}}{\sum_{j \in s_v} K\{d(\mathbf{x}_i^{(r)}, \mathbf{x}_j^{(v)})/h\}} \tag{1.21}$$

where $K(\cdot)$ is a zero-centered Kernel function (Epanechnikov 1969), and h is the corresponding bandwidth. Because k_{ij} is a proportion, the following property applies:

$$\sum_{i \in s_v} k_{ij} = 1, \quad k_{ij} \in [0, 1]. \tag{1.22}$$

The larger the value of k_{ij}, the more similar are the propensities of individuals i and j.

The weights to be used in the final estimator are given by the sum of the weights of the reference sample d^r multiplied by the kernel weight of the nonprobability sample unit as follows:

$$w_j^{KW} = \sum_{i \in s_r} d_i k_{ij}, \quad j \in s_v. \tag{1.23}$$

Estimators derived from KW are consistent as long as the regularity conditions considered in Wang et al. (2020) are met. This estimator is less sensitive to model misspecification than the PSA estimator while avoiding the extreme weights that may appear because of calculating w_i as $1/\hat{\pi}_i$Wang et al. (2020); Kern et al. (2021).

1.2.4 Statistical Matching

Statistical Matching, also known as Mass Imputation, is a model-based method that was introduced for nonprobability sampling estimation in Rivers (2007), and its theoretical properties were further developed in Yang and Kim (2018), Chen et al. (2020) and Kim et al. (2021).

This method predicts the (unknown) values of y for individuals of the reference probability sample rather than predicting propensities for individuals of the nonprobability sample. For the matter, we fit a model *SM* with covariates $\mathbf{x}$ as input, to predict the variable(s) of interest using data from s_v such that

$$\hat{y}_j = E_{SM}[y_j | \mathbf{x}_j, R_j], \qquad j \in s_r \tag{1.24}$$

The predicted values are plugged into Horvitz-Thompson or Hajek estimators as follows:

$$\hat{\bar{Y}}^{Mat}_{HT} = \frac{\sum_{i\in s_r} d_i \hat{y}_i}{N} \qquad \hat{\bar{Y}}^{Mat}_{Hajek} = \frac{\sum_{i\in s_r} d_i \hat{y}_i}{\sum_{i\in s_r} d_i}. \tag{1.25}$$

As noted in Kim et al. (2021), a key assumption must hold for Statistical Matching to work effectively, apart from the ignorability and common support assumptions (which were in place for PSA as well), which is the transportability condition:

$$f(y|\mathbf{x}, R = 1) = f(y|\mathbf{x}). \tag{1.26}$$

This condition requires the nonprobability sample to accurately represent the relationship between $\mathbf{x}$ and y. The ignorability assumption is sufficient for transportability to hold Kim et al. (2021), meaning that, if we assume a MAR selection mechanism, the transportability condition holds; therefore, the model *SM* can be used to predict y in s_r.

Several modeling approaches have been considered in literature for *SM*; Rivers (2007) used a donor imputation model which worked similarly to a k-NN model with $k = 1$, an approach that was generalized to any value of k in Yang and Kim (2018). Chen et al. (2022) applied nonparametric models, including kernel smoothing and GAM, and Castro-Martín et al. (2020) applied various Machine Learning models and compared their performance, also to the results provided by PSA.

An extension of Statistical Matching was proposed in Castro-Martín et al. (2022), which also combines results provided by PSA. In this approach, the predictive model *SM* is a weighted model fitted using data from s_v. Weights used in the model, w^{SM}, are calculated from the propensities estimated in PSA, $\hat{\pi}_{vi}$:

$$w_i^{SM} = 1/\hat{\pi}_{vi}, \quad i \in s_v \tag{1.27}$$

The resulting weighted model, *WSM* is applied to predict the values of y in the probability sample, obtaining a new set of predictions $\hat{y}^W$:

$$\hat{y}_j^W = E_{WSM}[y_j|\mathbf{x}_j, R_j], \qquad j \in s_r \tag{1.28}$$

These predictions can be used as regular Statistical Matching estimators:

$$\hat{\bar{Y}}^{WMat}_{HT} = \frac{\sum_{i\in s_r} d_i \hat{y}_i^W}{N} \qquad \hat{\bar{Y}}^{WMat}_{Hajek} = \frac{\sum_{i\in s_r} d_i \hat{y}_i^W}{\sum_{i\in s_r} d_i} \tag{1.29}$$

The behavior of these estimators was compared in Castro-Martín et al. (2022) with PSA and Doubly Robust estimators (which are to be introduced in the next section), showing better results than the former and very similar results to those of the latter.

1.2.5 Doubly Robust Estimators

Doubly Robust estimators for nonprobability samples (Chen et al. 2020) are based on the class of augmented inverse probability weighted estimators, which were developed in Robins (1994). The Doubly Robust estimator combines the design-based and model-based approaches; let $\hat{y}_i = E_{SM}[y_i|\mathbf{x}_i, R_i]$ be the predicted value for individual i according to model SM, defined as Statistical Matching, and let $\hat{\pi}_{vi} = E_M[R_i^* = 1|\mathbf{x}_i]$ be the estimated propensity of being in the nonprobability sample for individual i according to model M, defined as propensity score matching. The Doubly Robust Estimator of the population mean, $\hat{\bar{Y}}^{DR}$, is defined as follows:

$$\hat{\bar{Y}}^{DR} = \frac{1}{N}\sum_{i\in s_v}\frac{y_i - \hat{y}_i}{\hat{\pi}_{vi}} + \frac{1}{N}\sum_{i\in s_r} d_i\hat{y}_i. \tag{1.30}$$

Alternatively, the value of N can be substituted by the estimators derived from the sums of weights:

$$\hat{\bar{Y}}^{DR} = \frac{1}{\hat{N}_v}\sum_{i\in s_v}\frac{y_i - \hat{y}_i}{\hat{\pi}_{vi}} + \frac{1}{\hat{N}_r}\sum_{i\in s_r} d_i\hat{y}_i, \qquad \hat{N}_v = \sum_{i\in s_v} 1/\hat{\pi}_{vi} \quad \hat{N}_r = \sum_{i\in s_r} d_i. \tag{1.31}$$

As demonstrated in Chen et al. (2020), this estimator is doubly robust in the sense of being a consistent estimator of the population mean if either M or SM is correctly specified. The same work shows that the variance of this estimator also has a closed form if the propensity score model M is a logistic regression model.

2 Materials and methods

2.1 Data

To increase the extent of comparison among the methods proposed in literature, we conducted simulation studies using five different datasets as pseudopopulations. The first dataset was based on the simulation study reported in Wang et al. (2020) and included a pseudopopulation generated from actual population figures for United States counties. Three of the other datasets were publicly available in the UCI Machine Learning Repository Kelly et al. (2024), and were chosen in the basis of being multivariate datasets with more than 1,000 instances and 10 to 100 features (because datasets with more than 100 features are focused on other types of problems such as image processing), while the remaining one was available in the R package *TeachingSampling* (Rojas 2020). In all simulations, unequal probability sampling was performed using Poisson sampling, except for one case where the population was very large, making Poisson sampling to be computationally costly; in that case, systematic method for unequal probability sampling (Madow 1949) was used instead. Both sampling schemes are available in the R package *sampling* (Tillé and Matei 2021). Poisson sampling provides samples with no fixed sample size n_v,

but each sample drawn under this scheme has an expected sample size instead, $E[n_v]$. More details about the sampling designs are discussed in the following subsections. In addition, for each dataset and each iteration we computed the correlation between the indicator variable R and the variable of interest y, $\rho_{R,y}$, and we calculated the mean of $\rho_{R,y}$ and $\rho^2_{R,y}$ across all iterations to check the assumption of non-informativity. This correlation is the "data defect correlation" defined in Meng (2018), which measures the lack of representativeness of the sample. Violations of non-informativity can be associated with larger values of this indicator; for this reason, we will consider that non-informativity is likely to be satisfied if the correlation is low, although a deeper study (involving the modelization of y given $\mathbf{x}$) would be needed. See Wu (2022) and Meng (2018) for more details on the matter. Since in a simple random sampling without replacement $\rho_{R,y} = N^{-1/2}$, the correlations in each simulation will be compared to the inverse of the square root of the population size of the population in order to contextualize the figures.

2.1.1 ACS dataset

In the first dataset, 1000 clusters were simulated up to a size of $\overline{M} = 3000$ units per cluster (constituting a population of size $M = 3,000,000$). The units in each cluster were represented by the following variables: age, gender, income, ethnicity, and rural/urban area; each variable was simulated using data from 1000 random counties from the 2020 American Community Survey. Then, four new variables were generated:

$$Env \sim N(u, 0.5), \qquad u \sim U(0, 0.5), \qquad y \sim B(\mu, 1), \tag{2.1}$$

$$\begin{aligned}\mu = \big[&1 + exp(5 - 0.5 \cdot \text{Age} + I(\text{Gender} = \text{Male}) \\ &- I(\text{Ethnicity} = \text{Hispanic}) - 0.3 \cdot Env - 0.1 \cdot Env \cdot \text{Age})\big]^{-1}\end{aligned} \tag{2.2}$$

$$z = \mu + v, \quad v \sim N(0, 0.085) \tag{2.3}$$

Two other variables were also generated for sampling purposes:

$$q^a_k = exp(0.3 \cdot \text{Age} - 0.4 \cdot \text{Income} + 0.7 \cdot Env + 0.7 \cdot z)^{-1} \tag{2.4}$$

$$q^b_k = exp(0.3 \cdot \text{Age} - 0.4 \cdot \text{Income} + 0.7 \cdot Env + 0.7 \cdot z)^{0.5} \tag{2.5}$$

In each iteration, two samples were drawn from this population as follows:

- Volunteer sample ($E[n_v] = 5000$): two-stage cluster sampling design. In each stage, an unequal probability sampling was performed, with probabilities proportional to q^a_k, using the systematic sampling method.
- Reference sample ($n_r = 1500$):
  - Scenario 1: simple random sampling without replacement.

– Scenario 2: unequal probability sampling design with probabilities proportional to q_k^b using the systematic sampling method.

To construct the different models, all available variables were used as covariates except the variable of interest y and variables q_b and q_a that were used for sample selection. In this dataset, $\bar{\rho}_{R,y} = \sum_{m=1}^{1000} \frac{1}{1000} \rho_{R_m,y} = 0,001752$ and $\bar{\rho}^2_{R,y} = \sum_{m=1}^{1000} \frac{1}{1000} \rho^2_{R_m,y} = 4,04e-06$, (where R_m represents the indicator variable of belonging to s_v in the iteration number m of the simulation). There are very small correlations, so there is a near-zero relationship between the indicator variable R and y, which is compatible with the non-informativity assumption. However, it must be noted that $\overline{\rho}^2_{R,y}$ is 12.1 times bigger than $1/N$ (which is 3.3e−07 here), showing a bias that must be corrected.

2.1.2 Adult dataset

The Adult dataset, although available at the UCI repository, was retrieved from the dataset *AdultUCI* available in the *arules* R package (Hahsler et al. 2022). This dataset was extracted from the 1994 US Census Bureau database by Kohavi and Becker and Kohavi (1996). This dataset contains 48,842 observations from 15 variables, such as age, occupation, marital status, and income. For this analysis, we retained the 30,162 observations with no missing values, and 12 variables, with 1 variable of interest (income) and 11 covariates, removing "fnlwgt" and "native-country" (few units in several categories). After binarizing the qualitative covariates, which is a prerequisite of some machine learning classification algorithms, we also deleted categories with few responses to avoid prediction problems, in particular, "occupation_Armed-Forces", "marital-status_Married-AF-spouse", "education_Preschool", "workclass_Without-pay", "workclass_Never-worked". In each iteration, two samples were drawn from this population as follows:

- Volunteer sample ($E[n_v] = 1000$): unequal probability sampling design using Poisson sampling, where individuals with a large income were twice as likely as individuals with a small income to be included in the sample.
- Reference sample ($n_r = 500$): stratified sampling design, where the strata were the age (17–34, 35–49 and 50–90 years old) and gender. The sample size was allocated uniformly across all strata, meaning that the sample size for stratum h was calculated as $n_{rh} = n_r/L$, where L is the number of strata. Uniform allocation is used here to ensure that the design weights will be unequal (since the population strata sizes are unequal), and therefore ensuring a certain degree of complexity in the sampling scheme.

In this dataset, $\bar{\rho}_{R,y} = 0.06396638$ and $\bar{\rho}^2_{R,y} = 0.004135575$. The values obtained show very small correlations, so there is a near-zero relationship between R and y. However, $\overline{\rho}^2_{R,y}$ is 124.74 times bigger than $1/N$ (which is 3.3e−05 here), showing a large bias that could be caused by a violation of the non-informativity assumption.

The reason behind this large bias could be the fact that the volunteer sample was drawn according to a criteria where the variable of interest (income) was involved.

2.1.3 Bank dataset

The Bank dataset, available in the UCI repository Moro et al. (2012), was obtained from direct marketing campaigns of a Portuguese institution. Each one of the 45,211 individuals in the dataset represents a person who answered at least one phone call offering a bank term deposit. 16 variables are available for the analysis, including one variable (y) that measures whether the person has subscribed the term deposit or not, which was considered the variable of interest. The remaining 15 variables were used and were related to sociodemographic factors (age, occupation, education, marital status) and financial information, as well as details from the campaign (previous contacts, duration of the call, etcetera). In each iteration, two samples were drawn from this population:

- Volunteer sample ($E[n_v] = 1000$): unequal probability sampling design using Poisson sampling with probabilities proportional to the size of the variable "campaign", which measures the number of contacts performed prior to this campaign and for this client. It is a positively skewed variable (Fisher's coefficient of skewness: 4.898) with a mean of 2.764 calls, a median of 2 calls and a maximum of 63 calls.
- Reference sample ($n_r = 500$): stratified sampling design, where the strata represent age (18–34, 35–49 and 50–95 years old) and education level (primary, secondary, tertiary and unknown). The sample size was allocated uniformly across all strata, meaning that the sample size for stratum h was calculated as $n_{rh} = n_r/L$, where L is the number of strata.

In this dataset, $\bar{\rho}_{R_m,y} = -0.01222959$ and $\bar{\rho}^2_{R_m,y} = 0.000166763$. The latter figure is 7.54 times larger than $1/N$ (2.2e−05 here).

2.1.4 BigLucy dataset

The BigLucy dataset was retrieved from the *BigLucy* file available in the *TeachingSampling* R package (Rojas 2020). As stated in the package documentation, contains "some financial variables of 85,396 industrial companies in a city during a particular fiscal year". However, the actual file contains 85,296 observations of 11 variables (7 variables if we discard the variables "ID", "Ubication", "zone" and "segments", being identifiers of each company). For this analysis, we considered the income of the company as the variable of interest, and other 6 variables (company size, number of employees, amount of income tax paid by the company, use of SPAM, ISO certification and age of the company) as covariates. In each iteration, two samples were drawn from this population as follows:

- Volunteer sample ($E[n_v] = 1000$): unequal probability sampling design using Poisson sampling, where the probability of individual i in the pseudopopulation was calculated according to the following formula:

$$\pi_{vi} = \frac{z_i \cdot n_v}{\sum_{j=1}^{N} z_j}, \qquad z_i = \frac{exp(\text{Taxes}_i^2/400 + 5 \cdot (\text{SPAM} = \text{Yes})_i)}{1 + exp(\text{Taxes}_i^2/400 + 5 \cdot (\text{SPAM} = \text{Yes})_i)} \tag{2.6}$$

- Reference sample ($n_r = 10,000$): stratified sampling design in which the stratification variable was the county of the company (100 counties in total). The sample size was allocated uniformly across all strata, with a size of $n_{rh} = 100$ for each stratum h.

In this dataset, $\bar{\rho}_{R_m,y} = 0.006976786$ and $\bar{\rho}^2_{R_m,y} = 6.12293e-05$, so there is a near-zero relationship between R and y. $\bar{\rho}^2_{R,y}$ is 5.22 times larger than $1/N$ (1.17e−05 here).

2.1.5 Diabetes dataset

The Diabetes dataset, which is available in the UCI repository (Clore et al. 2014), contains hospital records of patients diagnosed with diabetes. The objective of the original research was to predict the early readmission of patients according to their characteristics. The original file contains 101,766 instances; however, each instance does not correspond to a person, but to a record, meaning that a given person can appear several times in the dataset. To overcome this problem, we retained only the first appearance of a given person (thanks to the identifier variables in the dataset), resulting in a final size of 68,629 instances after removing cases with missing values. In total 42 variables were available for the analysis, including one variable measuring whether the person has been readmitted or not, which was the variable of interest. The remaining 41 variables were related to sociodemographic (age, gender, race), medical and clinical information, of which we omitted the variables: "weight", "payer_code", "medical_specialty", "encounter_id patient_nbr", "diag1", "diag2" and "diag3". After binarizing the qualitative covariates, which is a prerequisite of some machine learning classification algorithms, we also deleted categories with few responses and variables with very unfrequent categories to avoid prediction problems: gender == "Unknown/Invalid", "admission_source_id", "discharge_disposition_id,admission_type_id", "metformin.pioglitazone", "metformin.rosiglitazone", "glyburide.metformin", "glipizide.metformin", "glimepiride.pioglitazone", "citoglipton", "examide", "troglitazone", "miglitol", "repaglinide", "nateglinide", "chlorpropamide", "acetohexamide", "tolbutamide", "acarbose", "tolazamide", "rosiglitazone", and "pioglitazone". In each iteration, two samples were drawn from this population as follows:

- Volunteer sample ($E[n_v] = 1000$): unequal probability sampling design using Poisson sampling, where the probability of individual i in the pseudopopulation was calculated according to the following formula:

$$\pi_{vi} = \frac{z_i \cdot n_v}{\sum_{j=1}^{N} z_j}, \qquad z_i = \frac{exp(\frac{x_{1i}-\bar{x}_1}{\sigma_{x_1}} + \frac{x_{2i}-\bar{x}_2}{\sigma_{x_2}})}{1 + exp(\frac{x_{1i}-\bar{x}_1}{\sigma_{x_1}} + \frac{x_{2i}-\bar{x}_2}{\sigma_{x_2}})}, \tag{2.7}$$

where x_1 is a variable measuring the number of days spent in hospital, with mean $\bar{x}_1$ and standard deviation σ_{x_1}, and x_2 is a variable measuring the number of diagnoses entered to the computer system of the hospital, with mean $\bar{x}_2$ and standard deviation σ_{x_2}.

- Reference sample ($n_r = 500$): stratified sampling design in which the strata were the age (0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, and 90–99 years old) and gender. The sample size was allocated uniformly across all strata, meaning that the sample size of stratum h was calculated as $n_{rh} = n_r/L$, where L is the number of strata.

In this dataset, $\bar{\rho}_{R_m,y} = 0.006472247$ and $\bar{\rho}^2_{R_m,y} = 5.61038e-05$, thus we also assumes non-informativity. However, $\overline{\rho}^2_{R,y}$ is 3.85 times larger than $1/N$ (1.46e−05 here).

2.2 Sampling and estimation

For each dataset, the experiments were repeated across 1, 000 iterations, estimating each mean value of the variable of interest (y in the ACS dataset, *income* = Large in Adult dataset, y = Yes in the Bank dataset, *income* in the Adult dataset, and *readmitted* = Yes in the Diabetes dataset). The methods tested in each simulation were the following:

- Linear calibration.
- Superpopulation modeling: model-based, model-assisted, and model-calibrated estimators, using linear regression, Ridge regression, and XGBoost.
- Propensity score adjustment with all 4 types of weight calculation, using logistic regression and XGBoost for propensity estimation.
- PAPP with the following algorithmic combinations:
  - Poisson regression for predicting d and logistic regression for predicting propensities.
  - XGBoost for predicting both d and propensities.
- Kernel Weighting (KW) using propensities obtained in PSA (with all types of weight calculation and the three predictive algorithms).
- Statistical Matching using linear regression and XGBoost to predict the target variable.
- Doubly Robust estimator with the following algorithmic combinations:
  - Linear regression for predicting the target variable and logistic regression for predicting the propensities.
  - XGBoost for prediction of target variable and propensities.

- PSA + Matching method proposed in Castro-Martín et al. (2022) with the same algorithmic combinations as in the Doubly Robust estimator.

When propensities had to be estimated, two approaches were tested: unweighted predictive algorithms (unweighted logistic regression/XGBoost), and weighted predictive algorithms (weighted logistic regression/XGBoost). In the weighted case, the vector of weights applied to both predictive algorithms were those of Eq 1.15. In addition, PSA, PAPP and KW were tested with and without variable selection prior to propensity estimation using the CFS algorithm (which showed the best performance across all algorithms tested in Ferri-García and Rueda (2022)) to select variables that were more correlated with the variable of interest in each dataset. Each of these combinations were used when combining PSA and Matching, but not in the case of the Doubly Robust estimator where we followed the original form proposed by Chen et al. (2020); therefore, we used weighted predictive algorithms with no variable selection.

For each approach, two measures were calculated for each dataset and scenario: percentage of relative bias (*%RB*),

$$\%RB = \frac{\left| \frac{\sum_{m=1}^{1000} \hat{\overline{Y}}_m}{1000} - \overline{Y} \right|}{\overline{Y}} \times 100, \tag{2.8}$$

and the Efficiency (*Eff*) of the estimates when compared to the unweighted estimator in terms of Mean Square Error (*MSE*)

$$Eff_{\text{Method k}} = \frac{MSE_{\text{Method k}}}{MSE_{\text{Unweighted}}}, \qquad MSE = \frac{\sum_{m=1}^{1000} (\hat{\overline{Y}}_m - \overline{Y})^2}{1000}, \tag{2.9}$$

where $\hat{\overline{Y}}_m$ represents the estimated population mean using a given method in iteration number m of the simulation.

3 Results

Complete results from the simulations can be found in Appendix, Tables 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19. For a better understanding of the vast number of results, we include some graphics and multivariate analyses in this section to describe the behavior of each approaches tested. Figure 1 shows boxplots of efficiency indicator *Eff* for each adjustment method separately for cases in which linear models (linear regression, logistic regression, or Poisson regression in PAPP) or XGBoost are used for prediction. The numbers in those boxplots represent the results of each method with all variations (variable selection/no variable selection, weighted/not weighted algorithms) in all datasets. Boxplots show that the methods that usually performed better than the unweighted case were those based on superpopulation models that used the complete census of the covariates: model-based,

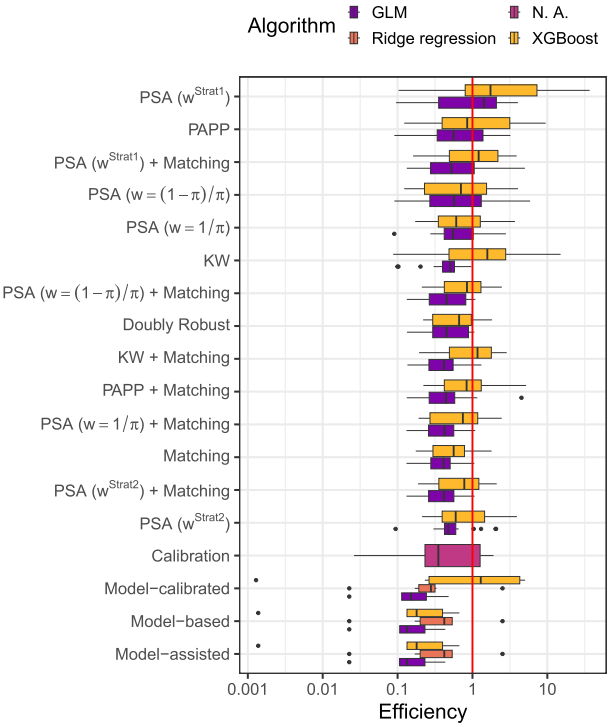

Fig. 1 Boxplots of *Eff* obtained using each adjustment method

model-adjusted, and model-calibrated estimators. These methods were followed by linear calibration, although this method's performance showed a larger heterogeneity.

The remaining methods also use a reference probability sample. It can be observed that Statistical Matching or methods that combine propensity estimation and Statistical Matching (including PSA + Matching and the Doubly Robust estimator) provided better results, with the exception of PSA using the stratification proposed by Valliant and Dever (2011). The performance of all these methods was very similar in terms of the median *Eff*. Followed by these methods, we obtained the performance of Kernel Weighting. This technique involves further smoothing (using kernel functions or simple averaging) of propensity predictions, leading to more stable behaviors. The remaining of the methods evaluated showed a behavior that was not far from Statistical Matching (or its combination with it), but were more unstable and, in terms of the median, were closer to the unadjusted case. Finally, the superiority of linear models over XGBoost can be pointed out; except for a few cases (PSA with no propensity stratification and perhaps PAPP and the Doubly Robust estimator), all adjustment methods provided smaller values for *Eff* when a linear model was used instead of XGBoost.

Figure 2 shows the boxplots of the efficiency indicator *Eff* only for those methods that involve variable selection and the possibility of using weighted algorithms. The

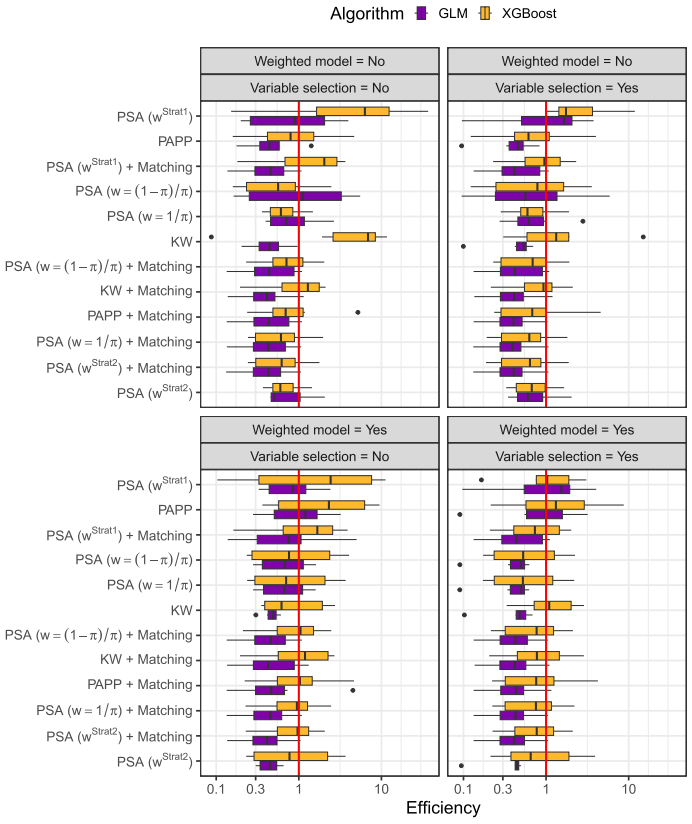

Fig. 2 Boxplots of *Eff* for methods with variable selection and weighted models

effect of selecting variables or using weights is different for each method; in general, we can see that using weighted predictive models improved the efficiency measure except for PAPP and combining Statistical Matching with PSA when the approach from Lee and Valliant (2009) was used for weighting (w^{Strat1}). The improvement was more noticeable in the rest of the cases in which PSA was used on its own. Regarding variable selection, it also generally improved the estimates of PAPP, PSA without stratification and KW; however, these improvements did not happen when these methods were used in combination with Statistical Matching. Finally, the superiority of linear models over XGBoost does not seem to be altered by using weighted models or variable selection, although XGBoost was able to improve its performance in some cases in which Statistical Matching was not involved (without reaching the levels of linear models).

To verify the effect of each methodology, for each adjustment method assessed in Fig. 2 we conducted an analysis based on raw differences in the percentage of relative bias (%RB) and Efficiency (*Eff*) of a given method applied in a given simulation and dataset when any of the three variations considered (GLM vs XGBoost,

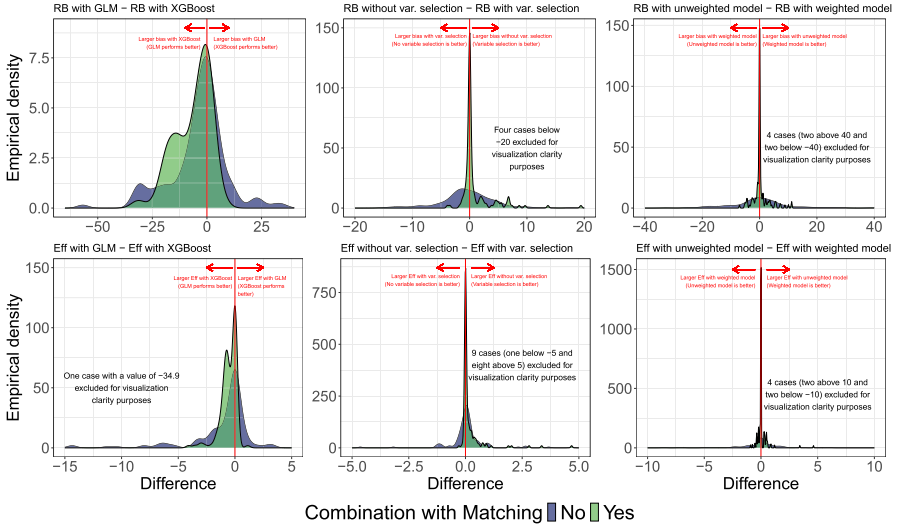

Fig. 3 Density plots of *RB* and *Eff* comparing predictive algorithm, variable selection and weighted models

no variable selection vs variable selection, unweighted vs weighted models) was changed. By doing so, we directly compared each approach with its counterfactual. These differences were also analyzed separately for adjustment methods that involve Statistical Matching versus those that do not. The differences are represented in density plots in Fig. 3. We also computed the percentage of times where the "novel" approach (XGBoost, variable selection and weighted models) was better than the "classic" approach (GLM, no variable selection and unweighted models). The percentages for each case are presented in Table 1.

Table 1 Percentages of *RB* and *Eff* comparing predictive algorithms, variable selection and weighted models

	%RB		*Eff*	
	No SM involved	SM involved	No SM involved	SM involved
XGBoost better	38.2%	31.2%	31.9%	9.7%
No difference	2.1%	1.4%	11.1%	26.4%
GLM better	59.7%	67.4%	56.9%	63.9%
Variable selection better	45.1%	44.4%	52.1%	34.7%
No difference	0%	29.9%	14.6%	61.8%
No variable selection better	54.9%	25.7%	33.3%	3.5%
Weighted models better	47.9%	38.2%	45.8%	20.1%
No difference	1.4%	22.2%	12.5%	57.6%
Unweighted models better	50.7%	39.6%	41.7%	22.2%

Some interesting patterns can be observed in the density plots. First, it is clear that the use of GLM provides better results most of the times, both in terms of bias and MSE, with this difference being greater when in combination with Statistical Matching. In 36.1% of the cases analyzed (38.9% when Statistical Matching was not involved vs 37.5% when it was) the difference in the efficiency indicator *Eff* was greater than 0, which indicates that XGBoost performed better in the exactly same situation. Second, the effects of variable selection are less clear; however, in most of the cases analyzed, it had a negative impact on the bias of the estimates but had a positive impact on the efficiency of the estimates, and generally worked better when Statistical Matching was not involved. Finally, regarding weighted algorithms, they had a null or almost null impact most of the times when applied in combination with Statistical Matching (50% of the times the difference in the *Eff* indicator was between −0.0224 and 0.0211). Otherwise, the effect was rather heterogeneous, but the cases where using weighted algorithms for propensity prediction positively affected the estimates (in terms of efficiency) outnumbered the cases in which it caused a negative effect.

Finally, Tables 2 and 3 provide a fair comparison of adjustment methods, where only the best result for %RB and *Eff* can be, respectively, consulted for each adjustment method and simulation, along with information about the combination of approaches that provided the best results. We can see that in the ACS dataset under the Scenario 1 (when the reference sample is drawn using simple random sampling) the methods that combine propensity estimation and Statistical Matching provided the smallest bias, although the MSE of the estimates was smallest when Statistical Matching was not used. Under Scenario 2 (when the reference sample is drawn using an unequal probability sampling scheme), the best performance is provided by superpopulation modeling methods that use the entire population census for a set of covariates. The best predictive algorithm was provided by linear models in all cases under Scenario 1, while XGBoost provided the best results under Scenario 2 when propensity estimation alone was used. It is also interesting to identify the approaches that provided the best efficiency: under Scenario 1, variable selection was involved in every best combination, and in most of them in Scenario 2. On the other hand, there is no consensus on weighted or unweighted algorithms under Scenario 1, while it is clear that almost every best combination in Scenario 2 used weighted algorithms. This is remarkable because the sampling design of the reference sample in Scenario 2 is much more complex than Scenario 1 and indicates that design weights should be considered in the modelization step, specially when informative.

The simulation using the Adult dataset yielded similar results to that using the ACS dataset in Scenario 1. Again, some methods based on propensity estimation seemed to outperform those based on superpopulation modeling using complete census data. In particular, this is remarkable in the case of Kernel Weighting and PSA with propensity stratification using the approach of Lee and Valliant (2009); in these cases, the percentages of relative bias were 9.3% and 3.8% respectively, while the remaining methods provided percentages above 20%. In this dataset, we can also see that almost all of the best combinations do not use variable selection, and that XGBoost provided the best results for almost all methods, probably revealing nonlinear relationships between the variables involved.

Table 2 Smallest percentage of relative bias (%RB) provided by each adjustment method across all approaches tested

Adjustment method	Value	Approach	Value	Approach	Value	Approach
	ACS dataset (Scenario 1)		*ACS dataset (Scenario 2)*		*Adult dataset*	
Calibration	21.10%		21.1%		34.5%	
Model-based	0.11%	GLM	0.1%	GLM	29.4%	XGB
Model-assisted	0.11%	GLM	0.1%	GLM	29.4%	XGB
Model-calibrated	0.29%	GLM	0.3%	GLM	30.3%	GLM
Statistical Matching	0.12%	GLM	1.4%	GLM	29.4%	XGB
Doubly Robust	0.13%	GLM	1.4%	GLM	29%	XGB
PSA ($w = 1/\pi$)	0.56%	W. M., V. S., GLM	0.4%	W. M., No V. S., XGB	29.2%	W. M., No V. S., XGB
PSA ($w = (1-\pi)/\pi$)	0.59%	W. M., V. S., GLM	0.3%	W. M., No V. S., XGB	23.3%	U. M., No V. S., GLM
PSA (w^{Strat1})	1.14%	U. M., V. S., GLM	1.6%	W. M., V. S., XGB	3.8%	W. M., No V. S., XGB
PSA (w^{Strat2})	1.25%	W. M., V. S., GLM	1.6%	W. M., No V. S., XGB	28.4%	W. M., No V. S., XGB
PAPP	0.59%	W. M., V. S., GLM	0.2%	U. M., V. S., GLM	34%	W. M., No V. S., XGB
KW	1.21%	U. M., V. S., GLM	1.4%	W. M., No V. S., XGB	9.3%	U. M., No V. S., XGB
PSA (w $= 1/\pi$) + Matching	0.04%	W. M., No V. S., GLM	1.3%	W. M., V. S., GLM	27.5%	W. M., V. S., XGB
PSA (w $= (1-\pi)/\pi$) + Matching	0.03%	U. M., No V. S., GLM	1.3%	W. M., V. S., GLM	26.4%	W. M., No V. S., XGB
PSA (w^{Strat1}) + Matching	0.02%	U. M., No V. S., GLM	1.4%	W. M., No V. S., GLM	21.4%	W. M., No V. S., XGB
PSA (w^{Strat2}) + Matching	0.03%	W. M., V. S., GLM	1.3%	W. M., V. S., GLM	27.5%	W. M., V. S., XGB
KW + Matching	0.01%	U. M., V. S., GLM	1.1%	W. M., V. S., GLM	24.9%	U. M., No V. S., XGB
PAPP + Matching	0.03%	U. M., No V. S., GLM	0.5%	W. M., No V. S., GLM	27.1%	W. M., No V. S., XGB
	Bank dataset		*BigLucy dataset*		*Diabetes dataset*	
Calibration	6.9%		0.41%		1.47%	
Model-based	3.7%	GLM	0.05%	XGB	0.47%	Ridge reg.

Table 2 (continued)

Adjustment method	Value	Approach	Value	Approach	Value	Approach
Model-assisted	3.7%	GLM	0.05%	XGB	0.47%	Ridge reg.
Model-calibrated	3.4%	XGB	0.02%	XGB	0.47%	Ridge reg.
Statistical Matching	3.8%	GLM	0.02%	XGB	0.59%	XGB
Doubly Robust	3.4%	GLM	0.01%	XGB	0.21%	GLM
PSA ($w = 1/\pi$)	2.2%	W. M., No V. S., GLM	0.31%	W. M., V. S., GLM	0.37%	U. M., No V. S., GLM
PSA ($w = (1-\pi)/\pi$)	1.7%	U. M., No V. S., GLM	0.27%	W. M., V. S., GLM	0.41%	U. M., V. S., GLM
PSA (w^{Strat1})	2.2%	W. M., No V. S., GLM	1.2%	W. M., V. S., XGB	0.64%	W. M., No V. S., GLM
PSA (w^{Strat2})	7.9%	W. M., No V. S., GLM	0.63%	W. M., V. S., GLM	0.25%	W. M., No V. S., GLM
PAPP	3.9%	U. M., No V. S., GLM	0.19%	U. M., V. S., GLM	0.26%	U. M., V. S., XGB
KW	9.3%	U. M., No V. S., GLM	0.27%	W. M., No V. S., GLM	1.06%	W. M., No V. S., XGB
PSA (w $= 1/\pi$) + Matching	2.7%	W. M., No V. S., GLM	0.01%	W. M., No V. S., XGB	0.76%	W. M., No V. S., GLM
PSA (w $= (1-\pi)/\pi$) + Matching	2.4%	U. M., No V. S., GLM	0.02%	W. M., V. S., XGB	0.52%	U. M., No V. S., GLM
PSA (w^{Strat1}) + Matching	2.2%	U. M., V. S., GLM	0.02%	W. M., V. S., XGB	0.05%	W. M., No V. S., GLM
PSA (w^{Strat2}) + Matching	2.8%	W. M., V. S., GLM	0%	W. M., No V. S., XGB	0.62%	W. M., No V. S., GLM
KW + Matching	1.5%	U. M., V. S., GLM	0.01%	U. M., No V. S., XGB	0.61%	U. M., No V. S., GLM
PAPP + Matching	2.7%	W. M., No V. S., GLM	0%	W. M., V. S., XGB	0.56%	W. M., No V. S., GLM

U. M., unweighted models were used in propensity estimation; W. M., weighted models were used in propensity estimation; V. S., variable selection; GLM, logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 3 Smallest value of efficiency (*Eff*) provided by each adjustment method across all tested approaches

Adjustment method	Value	Approach	Value	Approach	Value	Approach
	ACS dataset (Scenario 1)		*ACS dataset (Scenario 2)*		*Adult dataset*	
Calibration	1.909		1.909		0.334	
Model-based	0.107	GLM	0.107	GLM	0.242	XGB
Model-assisted	0.107	GLM	0.107	GLM	0.242	XGB
Model-calibrated	0.113	GLM	0.113	GLM	0.257	GLM
Statistical matching	0.132	GLM	1.060	GLM	0.251	XGB
Doubly Robust	0.133	GLM	1.060	GLM	0.248	XGB
PSA ($w = 1/\pi$)	0.090	W. M., V. S., GLM	0.172	W. M., V. S., XGB	0.266	W. M., No V. S., XGB
PSA ($w = (1-\pi)/\pi$)	0.091	W. M., V. S., GLM	0.173	W. M., V. S., XGB	0.162	U. M., No V. S., GLM
PSA (w^{Strat1})	0.096	U. M., V. S., GLM	0.165	W. M., V. S., XGB	0.103	W. M., No V. S., XGB
PSA (w^{Strat2})	0.094	W. M., V. S., GLM	0.213	W. M., V. S., XGB	0.256	W. M., No V. S., XGB
PAPP	0.091	W. M., V. S., GLM	0.834	U. M., V. S., GLM	0.338	U. M., No V. S., GLM
KW	0.100	U. M., V. S., GLM	0.351	W. M., No V. S., XGB	0.087	U. M., No V. S., XGB
PSA (w $= 1/\pi$) + Matching	0.131	U. M., V. S., GLM	1.051	W. M., V. S., GLM	0.223	W. M., V. S., XGB
PSA (w $= (1-\pi)/\pi$) + Matching	0.133	W. M., V. S., GLM	1.051	W. M., V. S., GLM	0.211	W. M., No V. S., XGB
PSA (w^{Strat1}) + Matching	0.133	W. M., V. S., GLM	1.055	W. M., V. S., GLM	0.161	W. M., No V. S., XGB
PSA (w^{Strat2}) + Matching	0.131	U. M., V. S., GLM	1.049	W. M., No. V. S., GLM	0.225	W. M., V. S., XGB
KW + Matching	0.135	U. M., V. S., GLM	1.055	W. M., No. V. S., GLM	0.194	U. M., No V. S., XGB
PAPP + Matching	0.133	W. M., V. S., GLM	1.038	U. M., V. S., GLM	0.221	W. M., No V. S., XGB
	Bank dataset		*BigLucy dataset*		*Diabetes dataset*	
Calibration	0.206		0.026		0.368	
Model-based	0.167	GLM	0.001	XGB	0.434	GLM

Table 3 (continued)

Adjustment method	Value	Approach	Value	Approach	Value	Approach
Model-assisted	0.167	GLM	0.001	XGB	0.434	GLM
Model-calibrated	0.173	Ridge reg.	0.001	XGB	0.480	GLM
Statistical Matching	0.340	GLM	0.491	XGB	0.499	GLM
Doubly Robust	0.373	GLM	0.491	XGB	0.990	XGB
PSA ($w = 1/\pi$)	0.341	W. M., V. S., GLM	0.412	U. M., No V. S., XGB	0.480	U. M., V. S., XGB
PSA ($w = (1-\pi)/\pi$)	0.243	U. M., V. S., GLM	0.363	U. M., No V. S., XGB	0.552	W. M., V. S., GLM
PSA (w^{Strat1})	0.351	U. M., V. S., GLM	0.750	W. M., V. S., XGB	0.820	W. M., No V. S., GLM
PSA (w^{Strat2})	0.420	W. M., V. S., GLM	0.430	W. M., V. S., GLM	0.403	U. M., V. S., XGB
PAPP	0.330	U. M., V. S., GLM	0.551	U. M., V. S., GLM	0.453	U. M., V. S., GLM
KW	0.322	U. M., No V. S., GLM	0.465	W. M., No V. S., GLM	0.359	U. M., No V. S., GLM
PSA (w $= 1/\pi$) + Matching	0.339	U. M., V. S., GLM	0.490	W. M., V. S., XGB	0.456	U. M., V. S., GLM
PSA (w $= (1-\pi)/\pi$) + Matching	0.339	U. M., V. S., GLM	0.491	W. M., V. S., XGB	0.618	W. M., V. S., GLM
PSA (w^{Strat1}) + Matching	0.337	W. M., V. S., GLM	0.491	U. M., V. S., XGB	0.701	U. M., No V. S., GLM
PSA (w^{Strat2}) + Matching	0.337	U. M., V. S., GLM	0.491	U. M., V. S., XGB	0.488	U. M., V. S., GLM
KW + Matching	0.331	W. M., V. S., GLM	0.492	U. M., V. S., XGB	0.490	U. M., No V. S., GLM
PAPP + Matching	0.341	U. M., V. S., GLM	0.492	U. M., V. S., XGB	0.480	U. M., V. S., GLM

U. M., unweighted models were used in propensity estimation; W. M., weighted models were used in propensity estimation; V. S., variable selection; GLM, logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

The simulations using the Bank and BigLucy datasets presented very similar results. We can see again that the less biased results among the best performances were provided by methods that involve the use of a reference sample, notably PSA, and any propensity estimation method as long as it was used to provide weights for Statistical Matching models. However, this is not true for efficiency - the smallest values of *Eff* across the best results were provided by superpopulation modeling methods. In general, linear models appear to work better on the Bank dataset, while XGBoost obtained some of the best results on the BigLucy dataset. Interestingly, in the Bank dataset simulation, variable selection was involved in almost all of the best results when we looked at efficiency, but was not involved in almost none of the best results when the relative bias was considered. This should not be surprising because variable selection techniques should reduce the complexity of models, which would have an effect on the variance of the estimates (and therefore on their MSE); they could also have an effect on bias if the model is more accurate with the set of selected variables, but this is not as usual as for the reduction in variance due to the parsimony principle explained above. A similar thing happened with the choice of weighted or unweighted models in both simulations: most of the best results in terms of bias involved weighted models (especially on the BigLucy dataset simulation), but if we look at the best efficiency results, they mostly involved unweighted models.

Finally, in the Diabetes dataset simulation, we observed a similar pattern, although in this case it seems that not selecting variables is preferable to selecting them. Again, we can see how methods based on propensity estimation can lead to smaller relative bias percentages than that of superpopulation modeling estimators, but when comparing their efficiency the former methods were not as competitive although some of them (namely, Kernel Weighting and PSA with the propensity stratification proposed in Valliant and Dever (2011)) had similar or even lower figures than the latter methods. It is also remarkable that linear calibration could outperform almost all the methods in terms of efficiency (although not in terms of bias). Weighted models were also involved in some of the best results, with no consensus across all methods.

3.1 Robustness with respect to possible violations of the non-informativity assumption

In order to see the behavior of the estimators when the propensity depends on the variable under study, we ran a new simulation for the ACS dataset. This is exactly the same as the previous one, but we changed the inclusion probability for the sample s_v, now depending on the variable of interest, simulating a case where the non-information assumption breaks down. The volunteer sample is selected with probabilities proportional to

$$q_k^a = \frac{\exp(0.1 \cdot Income + 5y)}{1 + \exp(0.1 \cdot Income + 5y)}.$$

Furthermore, we did not consider in the predictive model those variables that are part of y, namely $Gender = Male$, $Ethnicity = Hispanic$, Env and Age, thus simulating a Missing Not At Random (MNAR) situation, where survey participation depends exclusively on y, whereas before we had a Missing At Random (MAR) situation. We compute the percentage of relative bias and the MSE in the new situation for PSA, PSA+Matching in all of their variants, and Doubly Robust estimators, always using GLM as the predictive model. We compare them with the original scenario to see the relevance of this assumption when using these methods. The results are broken down in Tables 4, 5, 6 and 7.

We can see in all the tables that there is a large difference between the MNAR scenario and the original MAR scenario, regardless of the estimator, the sample design of s_r or the alternatives in the computation of propensity scores. Both bias and MSE increased significantly in the new scenario, where the sample selection s_v depended on the variable of interest, i.e., in the case of violation of the non-informativity assumption we obtained much more imprecise and erroneous estimates. These results show that the effectiveness of estimation methods to reduce bias are strongly dependent on the assumption of non-informativity. These results are consistent with Wu's view Wu (2022) that this assumption is the most crucial assumption for the validity of estimators based on propensity scores. On the other hand, *Eff* indicator shows that no method is able to consistently increase the efficiency of the estimates (relative to the unweighted case) under a MNAR selection mechanism. In addition, when a method achieves an increase in efficiency relative to the unweighted case, this increase is rather modest.

4 Discussion and conclusions

Due to the increasing attention that nonprobability surveys have received in the last few years, a wide variety of methods have been developed to handle selection biases that estimates from these surveys usually have. These methods can be divided between pseudodesign-based methods that aim to predict the (unknown) inclusion probabilities of each individual in the sample to obtain a new set of weights, and model-based methods, which aim to model the behavior of the variable of interest using data from the nonprobability sample and use it to predict the values of that variable in non-sampled individuals.

Different approaches based on this divide have been considered. In our literature review, we outlined propensity score adjustment, Kernel Weighting and probability-adjusted propensity prediction in the pseudodesign-based case, and superpopulation modeling in the model-based case (Model-assisted, Model-based and Model-calibrated estimators, and Statistical Matching). Calibration adjustments can also be considered a special case of Model-assisted estimators. We also identified combinations of these methods, such as the Doubly Robust estimator or Statistical Matching using PSA-weighed algorithms as proposed by Castro-Martín et al. (2022). Although some comparison studies can be found between several of these methods, this study attempts to fill the gap caused by the lack of a full comparison between all of the main approaches for selection bias mitigation in nonprobability

Table 4 Percentage of relative bias and MSE for each adjustment method with weighted models and variable selection

Weighted Model=Yes / Var. Selection=Yes												
Situation	MAR (original)						MNAR (new)					
Prob. Sample	MAS			Unequal prob.			MAS			Unequal prob.		
Estimators	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*
Unweighted	14.60	1.8E-04		14.60	0.00018		50.69	0.0016		50.69	0.0016	
PSA ($w = 1/\pi$)	0.56	1.6E-05	0.09	0.72	0.00011	0.62	50.69	0.0016	1.00	45.12	0.0020	1.20
PSA ($w = (1-\pi)/\pi$)	0.59	1.6E-05	0.09	0.75	0.00011	0.63	50.69	0.0016	1.00	45.12	0.0020	1.20
PSA (w^{Strat1})	1.22	1.7E-05	0.10	22.13	0.00037	2.08	49.92	0.0016	0.97	52.43	0.0019	1.17
PSA (w^{Strat2})	1.25	1.7E-05	0.09	0.83	0.00008	0.46	50.69	0.0016	1.00	45.17	0.0020	1.21
PSA ($w = 1/\pi$)+Matching	0.05	2.3E-05	0.13	1.27	0.00019	1.05	49.37	0.0015	0.95	55.61	0.0116	7.12
PSA ($w = (1-\pi)/\pi$)+Matching	0.05	2.3E-05	0.13	1.27	0.00019	1.05	49.39	0.0015	0.95	106.44	0.0266	16.42
PSA (w^{Strat1})+Matching	0.04	2.3E-05	0.13	1.48	0.00019	1.06	51.90	0.0017	1.05	52.49	0.0019	1.15
PSA (w^{Strat2})+Matching	0.03	2.3E-05	0.13	1.31	0.00019	1.05	49.37	0.0015	0.95	56.44	0.0125	7.73

Table 5 Percentage of relative bias and MSE for each adjustment method with weighted models and without variable selection

Weighted Model=Yes / Var. Selection=No												
Situation	MAR (original)						MNAR (new)					
Prob. Sample	MAS			Unequal prob.			MAS			Unequal prob.		
Estimators	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*
Unweighted	14.60	1.8E-04		14.60	0.00018		50.69	0.0016		50.69	0.0016	
PSA ($w = 1/\pi$)	5.24	4.9E-05	0.28	5.83	0.00010	0.59	52.65	0.0018	1.14	52.40	0.0019	1.15
PSA ($w = (1-\pi)/\pi$)	5.23	4.9E-05	0.28	5.82	0.00010	0.59	52.65	0.0018	1.14	52.40	0.0019	1.15
PSA (w^{Strat1})	6.63	5.8E-05	0.33	17.11	0.00024	1.36	51.57	0.0018	1.09	49.04	0.0015	0.95
PSA (w^{Strat2})	6.34	5.3E-05	0.30	6.03	0.00008	0.42	53.33	0.0019	1.16	51.61	0.0018	1.09
PSA ($w = 1/\pi$)+Matching	0.04	2.4E-05	0.13	1.35	0.00019	1.09	50.36	0.0031	1.90	42.84	0.0032	1.96
PSA ($w = (1-\pi)/\pi$)+Matching	0.04	2.4E-05	0.13	1.35	0.00019	1.09	85.18	0.0107	6.57	82.88	0.0099	6.11
PSA (w^{Strat1})+Matching	0.05	2.4E-05	0.14	1.41	0.00019	1.07	50.70	0.0016	1.00	51.98	0.0018	1.13
PSA (w^{Strat2})+Matching	0.04	2.4E-05	0.14	1.32	0.00019	1.05	57.69	0.0049	3.01	45.21	0.0027	1.66
Doubly Robust	0.13	2.4E-05	0.13	1.40	0.00019	1.06	49.34	0.0015	0.95	50.39	0.0017	1.07

Table 6 Percentage of relative bias and MSE for each adjustment method with variable selection and unweighted models

Weighted Model=No / Var. Selection=Yes												
Situation	MAR (original)						MNAR (new)					
Prob. Sample	MAS			Unequal prob.			MAS			Unequal prob.		
Estimators	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*
Unweighted	14.60	1.8E-04		14.60	0.00018		50.69	0.0016		50.69	0.0016	
PSA ($w = 1/\pi$)	6.68	4.8E-05	0.27	27.25	0.00049	2.80	50.69	0.0016	1.00	50.70	0.0016	1.00
PSA ($w = (1-\pi)/\pi$)	1.25	1.7E-05	0.10	39.92	0.00104	5.89	50.69	0.0016	1.00	50.70	0.0016	1.00
PSA (w^{Strat1})	1.14	1.7E-05	0.10	31.78	0.00067	3.76	49.92	0.0016	0.97	49.92	0.0016	0.97
PSA (w^{Strat2})	7.84	6.1E-05	0.35	22.88	0.00036	2.04	50.69	0.0016	1.00	50.70	0.0016	1.00
PSA ($w = 1/\pi$)+Matching	0.05	2.3E-05	0.13	1.42	0.00019	1.07	49.38	0.0015	0.95	50.42	0.0017	1.07
PSA ($w = (1-\pi)/\pi$)+Matching	0.05	2.3E-05	0.13	1.47	0.00019	1.09	49.38	0.0015	0.95	50.42	0.0017	1.07
PSA (w^{Strat1})+Matching	0.04	2.3E-05	0.13	1.54	0.00019	1.08	51.90	0.0017	1.05	52.92	0.0019	1.17
PSA (w^{Strat2})+Matching	0.05	2.3E-05	0.13	1.49	0.00019	1.07	49.38	0.0015	0.95	50.42	0.0017	1.07

Table 7 Percentage of relative bias and MSE for each adjustment method without variable selection and unweighted models

Weighted Model=No / Var. Selection=No												
Situation	MAR (original)						MNAR (new)					
Prob. Sample	MAS			Unequal prob.			MAS			Unequal prob.		
Estimators	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*	%RB	MSE	*Eff*
Unweighted	14.60	1.8E-04		14.60	0.00018		50.69	0.0016		50.69	0.0016	
PSA ($w = 1/\pi$)	8.51	7.0E-05	0.40	26.41	0.00047	2.66	46.57	0.0019	1.15	45.84	0.0014	0.87
PSA ($w = (1-\pi)/\pi$)	2.34	3.1E-05	0.18	38.31	0.00098	5.53	46.57	0.0019	1.15	45.84	0.0014	0.87
PSA (w^{Strat1})	4.58	3.9E-05	0.22	32.67	0.00070	3.98	50.94	0.0018	1.08	51.64	0.0017	1.06
PSA (w^{Strat2})	9.38	8.1E-05	0.46	23.00	0.00037	2.07	47.34	0.0019	1.18	46.50	0.0015	0.90
PSA ($w = 1/\pi$)+Matching	0.08	2.4E-05	0.13	1.55	0.00019	1.07	68.08	0.0141	8.71	60.28	0.0073	4.51
PSA ($w = (1-\pi)/\pi$)+Matching	0.03	2.4E-05	0.13	1.64	0.00019	1.09	108.38	0.0168	10.35	135.46	0.0249	15.35
PSA (w^{Strat1})+Matching	0.02	2.4E-05	0.14	1.58	0.00019	1.08	50.44	0.0016	0.99	51.42	0.0018	1.11
PSA (w^{Strat2})+Matching	0.05	2.3E-05	0.13	1.50	0.00019	1.06	37.16	0.0072	4.44	40.83	0.0038	2.37

samples. In addition, we have introduced some original methodologies, such as Statistical Matching with weighted algorithms where the weights are provided by Kernel Weighting and by PAPP instead of PSA. Finally, we also attempted to compare not only adjustment methods, but also alternative approaches in these methods; for example, we compared four different approaches for weight calculation in PSA and have also compared how propensity estimation works when a variable selection method is applied. We also attempted to check the properties of weighted algorithms when estimating propensities, which might be the most efficient and consistent approach according to the theoretical results (Chen et al. 2020). Finally, we have done five different simulation studies using both synthetical data and real-life data, and considered various levels of complexity in sampling design for probability and nonprobability samples and different types of variables of interest. We consider that including more datasets and simulations has increased the value of this comparison study because it makes its results more generalizable to a wider variety of situations in survey methodology and also considers how different situations may require different approaches to obtain the best possible results.

The findings of our study reveal some remarkable patterns:

- Adjustments based on superpopulation modeling that use the whole population census for a set of covariates (Model-based, Model-assisted and Model-calibrated estimators) provide overall, the best or almost the best results in all situations considered in terms of efficiency, although they might not be as good at bias reduction. The superiority of these methods have already been documented in the literature Ferri-García et al. (2021), but they require observing all individuals in the population for a set of common covariates with nonprobability sample. This makes it difficult to apply them with a sufficient number of variables in real situations.
- Other superpopulation modeling approaches (namely Statistical Matching) also provided good results, both alone and in combination with propensity estimation through Doubly Robust estimators or propensity-weighted predictive algorithms. However, the latter approach seemed to work better than the Doubly Robust estimators in the six simulations evaluated, a circumstance that has already been observed in the literature Castro-Martín et al. (2021). The superiority of Doubly Robust adjustments, along with model-based estimators, was also observed in Valliant (2020).
- Regarding pseudodesign-based methods, their performance was vastly associated with the dataset in which they were used, but methods such as Kernel Weighting showed good results in all situations. The success of combining these methods with Statistical Matching was not consistent although it can be observed that this combination produced less biased estimates but with higher MSE.
- The use of weighted models in propensity estimation was particularly positive for cases in which the design effect of the reference sample could be larger (such as the ACS dataset under Scenario 2 or the Diabetes dataset, where the number of strata in the sampling design of the reference sample was relatively large). When propensities were applied as weights for the Statistical Matching model, this correction had a mostly a null effect on the final estimates. We consider that

this could be explained by the fact that Statistical Matching models focus on predicting the variable of interest; therefore, the prediction does not require to be elevated to the population size, only to reflect the relationships between the target variable and covariates.

- As expected, the effect of variable selection was more noticeable when propensity estimation methods were not combined with Statistical Matching. The overall effect of variable selection on efficiency was more positive than its effect on bias reduction. As noted in Ferri-García and Rueda (2022), the main objective of variable selection algorithms is reducing the complexity of the models and the variance of the estimates (something observed in that work). However, these methods can also help to reduce bias if they can transform a misspecified model into a properly specified one. As stated in the literature Mercer et al. (2018); Boyd et al. (2024), finding the appropriate auxiliary covariates may be the most important step in nonprobability sampling adjustment, even more important than the method chosen for adjustment. Although some differences between methods were observed in our results, their main driver might be the covariates used.

Considering the findings described above, we conclude that the best strategy when dealing with selection bias in a nonprobability sample is to use estimators based on superpopulation modeling using the whole population census on a set of covariates, if available, with a predictive model that accurately reflects the associations between variables in the dataset; cross-validation techniques commonly used in data science can be helpful for the assessment of this step. If this assessment cannot be performed, we recommend using linear models in prediction.

If a complete census of a set of covariates is not available, we recommend the use of different approaches depending on the datasets and the properties we are attempting to optimize. If we focus on unbiasedness, the combination of propensity estimation model with Statistical Matching (through weighted Statistical Matching models) might be the best option. However, if this combination is not possible (which can be the case of many multipurpose surveys with multiple variables of interest), we recommend using variable selection techniques to fit the propensity estimation model. It is also important to consider that in a multipurpose survey in which some variables of interest have uniform bias while others have a random bias (where the selection mechanism is not related to any observed variable), a possible solution would be to use weight smoothing Beaumont (2008); Ferri-García et al. (2022). Weight smoothing is based on substituting the final vector of weights with the predicted values of those weights according to a model in which the independent variables are the variables of interest. It can be theoretically proven that this approach mitigates the increase in variance that might be caused by misspecification of the propensity model for some variables (i.e., considering irrelevant variables, which is the case for the variables where the bias is completely random), and the empirical results from Ferri-García et al. (2022) show that this property holds when propensity estimation is used in nonprobability samples.

On the other hand, if we focus on efficiency or MSE, propensity estimation modeling on its own seems to be the best option, especially if the propensities are further smoothed using Kernel Weighting. In addition, if the probability sample used

as a reference has a complex sampling design, we recommend using weighted models for the prediction of propensities; if it has a simple design, we also recommend weighted models because they are theoretically consistent, but the decision does not make a significant difference. Finally, the use of linear models is generally recommended if the use of other nonlinear or nonparametric models cannot be assessed.

The estimation procedures for nonprobability survey samples are based on several assumptions and the effectiveness of the estimators depends on satisfying these assumptions. How to test the assumptions in practical applications of the methods is a question that cannot be fully answered. Wu (2022) presented a method using comparisons of marginal distributions and conditional models that can be useful for building confidence in the ignorability assumption. It is also therefore important that, in the study design phase before data collection, researchers and practitioners pay attention to potential factors and features of units that may be related to sample participation. In this article, we have seen the large errors that occur when the noninformativity assumption is not satisfied.

Our work has several limitations. First, it is worth noting that adjustment methods are continuously being developed; therefore, there are some promising techniques that were not included in this comparison. Second, we considered five different datasets; however, this number should be ideally larger, especially if statistical inference is used to study the effects of applying a given approach. The datasets can be considered random effects, as there are infinite datasets that could appear in real-world situations; however, we can only test the proposed methods in a handful of them; therefore, multilevel models could be used to study the effects of each algorithm, method and modification, and a large pool of datasets can be very valuable for the estimation of the effects and their interactions.

It should be noted that each dataset and problem require a different approach to obtain the best possible performance; however, we do not have any information about the exact approach that should be used in real situations (although the present study provides some hints and recommendations about the best choice in some situations). Further studies should focus on research of metrics and measures that allow researchers to estimate the amount of bias present in the estimation from a given sample and what fraction of that bias can be removed using a given method. This research line would be valuable for practitioners.

Appendix A: Results of relative bias and efficiency of the simulation studies

Table 8 Percentage of relative bias of each adjustment method in the ACS dataset simulation when Scenario 1 is applied (simple random sampling for the reference sample)

Adjustment method								
Unweighted	14.60%							
Calibration	21.10%							
	GLM	XGBoost	Ridge reg.					
Model-based	0.11%	2.98%	2.80%					
Model-assisted	0.11%	2.98%	2.80%					
Model-calibrated	0.29%	33.93%	2.57%					
Statistical Matching	0.12%	2.94%						
Doubly Robust	0.13%	3.34%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	8.51%	7.62%	6.68%	6.60%	5.24%	7.70%	0.56%	5.36%
PSA ($w = (1-\pi)/\pi$)	2.34%	2.35%	1.25%	2.77%	5.23%	7.77%	0.59%	5.40%
PSA (w^{Strat1})	4.58%	13.96%	1.14%	19.49%	6.63%	12.05%	1.22%	13.58%
PSA (w^{Strat2})	9.38%	7.78%	7.84%	7.39%	6.34%	7.79%	1.25%	9.97%
PAPP	2.34%	2.35%	1.25%	2.77%	5.23%	7.77%	0.59%	5.40%
KW	3.18%	16.66%	1.21%	19.07%	5.85%	8.00%	1.22%	13.88%
PSA (w $= 1/\pi$) + Matching	0.08%	5.32%	0.05%	3.49%	0.04%	12.42%	0.05%	5.68%
PSA (w $= (1-\pi)/\pi$) + Matching	0.03%	9.50%	0.05%	4.85%	0.04%	12.51%	0.05%	5.66%
PSA (w^{Strat1}) + Matching	0.02%	19.82%	0.04%	12.95%	0.05%	17.35%	0.04%	7.67%
PSA (w^{Strat2}) + Matching	0.05%	5.35%	0.05%	3.44%	0.04%	12.71%	0.03%	7.66%
KW + Matching	0.05%	17.67%	0.01%	12.12%	0.02%	12.65%	0.02%	8.39%
PAPP + Matching	0.03%	9.50%	0.05%	4.85%	0.04%	12.51%	0.05%	5.66%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 9 Efficiency of each adjustment method in the ACS dataset simulation when Scenario 1 is applied (simple random sampling for the reference sample)

Adjustment method								
Calibration	1.909							
	GLM	XGBoost	Ridge reg.					
Model-based	0.107	0.134	0.537					
Model-assisted	0.107	0.134	0.537					
Model-calibrated	0.113	4.284	0.279					
Statistical Matching	0.132	0.175						
Doubly Robust	0.133	0.219						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	0.397	0.358	0.274	0.283	0.279	0.395	0.090	0.212
PSA ($w = (1-\pi)/\pi$)	0.176	0.159	0.095	0.123	0.279	0.400	0.091	0.214
PSA (w^{Strat1})	0.221	1.096	0.096	1.657	0.327	0.799	0.097	0.800
PSA (w^{Strat2})	0.457	0.368	0.347	0.329	0.301	0.403	0.094	0.486
PAPP	0.176	0.159	0.095	0.123	0.279	0.400	0.091	0.214
KW	0.203	1.906	0.100	1.768	0.302	0.465	0.103	0.844
PSA (w = $1/\pi$) + Matching	0.134	0.253	0.131	0.191	0.134	0.737	0.133	0.277
PSA (w = $(1-\pi)/\pi$) + Matching	0.134	0.483	0.133	0.237	0.134	0.742	0.133	0.278
PSA (w^{Strat1}) + Matching	0.137	1.616	0.133	0.812	0.138	1.308	0.133	0.381
PSA (w^{Strat2}) + Matching	0.132	0.256	0.131	0.188	0.136	0.775	0.132	0.397
KW + Matching	0.138	1.318	0.135	0.753	0.135	0.813	0.137	0.436
PAPP + Matching	0.134	0.483	0.133	0.237	0.134	0.742	0.133	0.278

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 10 Percentage of relative bias of each adjustment method in the ACS dataset simulation when Scenario 2 is applied (unequal probability sampling for the reference sample)

Adjustment method								
Unweighted	14.6%							
Calibration	21.1%							
	GLM	XGBoost	Ridge reg.					
Model-based	0.1%	3.0%	2.8%					
Model-assisted	0.1%	3.0%	2.8%					
Model-calibrated	0.3%	33.9%	2.6%					
Statistical Matching	1.4%	4.3%						
Doubly Robust	1.4%	4.4%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	26.4%	19.1%	27.2%	22.0%	5.8%	0.4%	0.7%	1.2%
PSA ($w = (1-\pi)/\pi$)	38.3%	25.4%	39.9%	30.8%	5.8%	0.3%	0.7%	1.2%
PSA (w^{Strat1})	32.7%	42.9%	31.8%	56.1%	17.1%	2.4%	22.1%	1.6%
PSA (w^{Strat2})	23.0%	18.9%	22.9%	20.4%	6.0%	1.6%	0.8%	2.2%
PAPP	7.6%	2.5%	0.2%	1.8%	11.3%	17.2%	20.3%	19.1%
KW	11.5%	45.3%	3.3%	60.1%	6.4%	1.4%	2.1%	13.2%
PSA (w = $1/\pi$) + Matching	1.5%	4.3%	1.4%	4.2%	1.4%	15.6%	1.3%	9.0%
PSA (w = $(1-\pi)/\pi$) + Matching	1.6%	5.4%	1.5%	4.8%	1.4%	15.5%	1.3%	9.2%
PSA (w^{Strat1}) + Matching	1.6%	12.1%	1.5%	10.0%	1.4%	12.7%	1.5%	7.7%
PSA (w^{Strat2}) + Matching	1.5%	4.6%	1.5%	4.4%	1.3%	13.4%	1.3%	9.6%
KW + Matching	1.3%	10.9%	1.1%	8.9%	1.2%	16.3%	1.1%	15.9%
PAPP + Matching	1.3%	18.0%	1.3%	12.5%	0.5%	23.8%	1.3%	15.1%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 11 Efficiency of each adjustment method in the ACS dataset simulation when Scenario 2 is applied (unequal probability sampling for the reference sample)

Adjustment method								
Calibration	1.909							
	GLM	XGBoost	Ridge reg.					
Model-based	0.107	0.134	0.537					
Model-assisted	0.107	0.134	0.537					
Model-calibrated	0.113	4.284	0.279					
Statistical Matching	1.060	1.801						
Doubly Robust	1.060	1.829						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	2.661	1.481	2.795	1.898	0.585	0.237	0.624	0.172
PSA ($w = (1-\pi)/\pi$)	5.532	2.476	5.893	3.577	0.586	0.238	0.626	0.173
PSA (w^{Strat1})	3.981	7.232	3.760	11.870	1.359	0.244	2.084	0.165
PSA (w^{Strat2})	2.065	1.444	2.039	1.655	0.425	0.231	0.459	0.213
PAPP	1.414	4.685	0.834	4.009	3.211	9.493	3.188	8.727
KW	0.967	8.897	0.698	15.018	0.607	0.351	0.685	1.411
PSA (w = $1/\pi$) + Matching	1.069	1.969	1.072	1.814	1.088	2.461	1.051	2.208
PSA (w = $(1-\pi)/\pi$) + Matching	1.092	2.031	1.090	1.915	1.088	2.469	1.051	2.111
PSA (w^{Strat1}) + Matching	1.075	2.559	1.079	2.315	1.067	2.145	1.055	2.013
PSA (w^{Strat2}) + Matching	1.061	1.776	1.067	1.882	1.049	2.065	1.053	2.108
KW + Matching	1.147	2.108	1.198	2.106	1.055	2.701	1.100	2.867
PAPP + Matching	1.093	5.195	1.038	4.581	4.518	4.654	1.159	4.239

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 12 Percentage of relative bias of each adjustment method in the Adult dataset simulation

Adjustment method								
Unweighted	60%							
Calibration	34.5%							
	GLM	XGBoost	Ridge reg.					
Model-based	30.4%	29.4%	34.4%					
Model-assisted	30.4%	29.4%	34.4%					
Model-calibrated	30.3%	37.4%	34.4%					
Statistical Matching	29.9%	29.4%						
Doubly Robust	30.3%	29%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	41.6%	47%	44.5%	43.7%	31.8%	29.2%	40.5%	29.8%
PSA ($w = (1-\pi)/\pi$)	23.3%	25.9%	29%	24%	30.9%	26.8%	39.8%	28.3%
PSA (w^{Strat1})	26%	16.2%	65.4%	42.5%	34.8%	3.8%	66.3%	43.7%
PSA (w^{Strat2})	43.2%	46.6%	46.1%	44.4%	32.9%	28.4%	41.5%	30.1%
PAPP	34.4%	35.4%	40.3%	35.2%	35.5%	34%	48.4%	37.7%
KW	44%	9.3%	45.4%	20.2%	38%	34.6%	40%	31.3%
PSA (w = $1/\pi$) + Matching	29.9%	28.8%	29.8%	28.8%	29.6%	27.6%	29.8%	27.5%
PSA (w = $(1-\pi)/\pi$) + Matching	30.5%	27.9%	30.1%	27.9%	29.8%	26.4%	29.8%	26.8%
PSA (w^{Strat1}) + Matching	30.4%	23.6%	30.6%	27.1%	29.9%	21.4%	30.4%	25.2%
PSA (w^{Strat2}) + Matching	29.9%	28.8%	29.9%	28.8%	29.8%	27.6%	29.8%	27.5%
KW + Matching	30.3%	24.9%	30.1%	26.7%	29.9%	25%	29.9%	25.9%
PAPP + Matching	30.2%	28.4%	29.8%	28.4%	30.4%	27.1%	30%	27.3%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 13 Efficiency of each adjustment method in the Adult dataset simulation

Adjustment method								
Calibration	0.334							
	GLM	XGBoost	Ridge reg.					
Model-based	0.259	0.242	0.330					
Model-assisted	0.259	0.242	0.330					
Model-calibrated	0.257	0.389	0.330					
Statistical Matching	0.260	0.251						
Doubly Robust	0.273	0.248						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	0.485	0.615	0.553	0.534	0.322	0.266	0.469	0.326
PSA ($w = (1-\pi)/\pi$)	0.162	0.201	0.245	0.188	0.308	0.234	0.456	0.313
PSA (w^{Strat1})	0.198	0.151	1.507	1.358	0.355	0.103	1.517	1.333
PSA (w^{Strat2})	0.522	0.604	0.595	0.554	0.315	0.256	0.496	0.345
PAPP	0.338	0.360	0.465	0.376	0.394	0.362	0.686	0.523
KW	0.544	0.087	0.580	0.303	0.412	0.363	0.459	0.334
PSA (w = $1/\pi$) + Matching	0.260	0.242	0.259	0.242	0.257	0.226	0.258	0.223
PSA (w = $(1-\pi)/\pi$) + Matching	0.269	0.230	0.263	0.229	0.266	0.211	0.258	0.215
PSA (w^{Strat1}) + Matching	0.269	0.180	0.280	0.228	0.261	0.161	0.278	0.209
PSA (w^{Strat2}) + Matching	0.261	0.242	0.259	0.242	0.258	0.228	0.258	0.225
KW + Matching	0.266	0.194	0.262	0.214	0.261	0.194	0.260	0.205
PAPP + Matching	0.265	0.237	0.258	0.236	0.269	0.221	0.262	0.222

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 14 Percentage of relative bias of each adjustment method in Bank dataset simulation

Adjustment method								
Unweighted	22.4%							
Calibration	6.9%							
	GLM	XGBoost	Ridge reg.					
Model-based	3.7%	16.7%	4.9%					
Model-assisted	3.7%	16.7%	4.9%					
Model-calibrated	5.4%	3.4%	4.9%					
Statistical Matching	3.8%	16.9%						
Doubly Robust	3.4%	17.3%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	10.4%	21.5%	13.1%	22.5%	2.2%	32.9%	3.8%	33.4%
PSA ($w = (1-\pi)/\pi$)	1.7%	18.8%	3.8%	22.7%	1.7%	34.5%	3.3%	33.7%
PSA (w^{Strat1})	3.4%	32.2%	9.9%	18.6%	2.2%	34.6%	10.8%	32.3%
PSA (w^{Strat2})	13.4%	21.6%	16.7%	22.8%	7.9%	33.1%	11.8%	35.4%
PAPP	3.9%	19.7%	5.8%	24.4%	5.2%	36.9%	3.9%	33.6%
KW	9.3%	31.7%	12.5%	14%	13.3%	34.5%	12.2%	38.5%
PSA (w $= 1/\pi$) + Matching	3.4%	18.3%	4%	17.5%	2.7%	22%	3.7%	21%
PSA (w $= (1-\pi)/\pi$) + Matching	2.4%	21%	3.7%	18.6%	2.5%	24.5%	3.7%	21.5%
PSA (w^{Strat1}) + Matching	2.5%	34.1%	2.2%	20.4%	6.7%	39.7%	2.2%	20.3%
PSA (w^{Strat2}) + Matching	3.4%	18.4%	3.3%	17.6%	2.8%	21.7%	2.8%	22.1%
KW + Matching	3.1%	21.1%	1.5%	20.5%	3%	32.3%	1.8%	24.9%
PAPP + Matching	3.1%	21.5%	3.8%	19.1%	2.7%	24.1%	3.6%	22.1%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 15 Efficiency of each adjustment method in Bank dataset simulation

Adjustment method								
Calibration	0.206							
	GLM	XGBoost	Ridge reg.					
Model-based	0.167	0.669	0.173					
Model-assisted	0.167	0.669	0.173					
Model-calibrated	0.202	0.231	0.173					
Statistical Matching	0.340	0.836						
Doubly Robust	0.373	0.896						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	0.445	0.940	0.427	1.017	1.257	2.457	0.341	2.193
PSA ($w = (1-\pi)/\pi$)	0.729	0.922	0.243	1.089	1.285	2.786	0.346	2.241
PSA (w^{Strat1})	0.414	14.818	0.351	0.991	2.414	7.671	0.386	2.115
PSA (w^{Strat2})	0.481	0.957	0.611	1.036	0.481	2.559	0.420	2.446
PAPP	0.337	0.973	0.330	1.285	1.056	3.241	0.549	2.324
KW	0.322	7.208	0.511	0.993	0.538	2.561	0.499	2.875
PSA (w $= 1/\pi$) + Matching	0.361	0.935	0.339	0.881	0.394	1.218	0.348	1.157
PSA (w $= (1-\pi)/\pi$) + Matching	0.370	1.155	0.339	0.961	0.394	1.481	0.347	1.203
PSA (w^{Strat1}) + Matching	0.386	3.012	0.337	1.120	5.017	3.894	0.337	1.090
PSA (w^{Strat2}) + Matching	0.354	0.947	0.337	0.877	0.350	1.199	0.342	1.241
KW + Matching	0.346	1.255	0.340	1.148	0.353	2.464	0.331	1.534
PAPP + Matching	0.360	1.185	0.341	1.007	0.411	1.460	0.356	1.272

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 16 Percentage of relative bias of each adjustment method in the BigLucy dataset simulation

Adjustment method								
Unweighted	4.03%							
Calibration	0.41%							
	GLM	XGBoost	Ridge reg.					
Model-based	0.36%	0.05%	0.36%					
Model-assisted	0.36%	0.05%	0.36%					
Model-calibrated	0.36%	0.02%	0.36%					
Statistical Matching	0.43%	0.02%						
Doubly Robust	0.35%	0.01%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	4.69%	2.33%	4.25%	3.36%	1.62%	7.46%	0.31%	4.49%
PSA ($w = (1-\pi)/\pi$)	5.36%	0.59%	4.48%	2.67%	1.59%	7.86%	0.27%	4.59%
PSA (w^{Strat1})	6.02%	2.39%	4.85%	5.04%	3.28%	10.37%	7.94%	1.2%
PSA (w^{Strat2})	4.86%	2.51%	4.3%	3.75%	1.4%	7.37%	0.63%	8.04%
PAPP	0.87%	5.03%	0.19%	2.15%	2.22%	11.52%	3.5%	6.83%
KW	1.77%	1.74%	1.47%	0.85%	0.27%	6.28%	1.79%	5.68%
PSA (w $= 1/\pi$) + Matching	0.35%	0.02%	0.34%	0.02%	0.28%	0.01%	0.27%	0.02%
PSA (w $= (1-\pi)/\pi$) + Matching	0.31%	0.02%	0.3%	0.03%	0.28%	0.07%	0.27%	0.02%
PSA (w^{Strat1}) + Matching	0.33%	0.34%	1.16%	0.02%	0.27%	0.45%	1.65%	0.02%
PSA (w^{Strat2}) + Matching	0.36%	0.02%	0.33%	0.02%	0.15%	0%	0.15%	0.02%
KW + Matching	0.31%	0.01%	0.27%	0.01%	0.38%	0.38%	0.33%	0.03%
PAPP + Matching	0.32%	0.02%	0.26%	0.02%	0.35%	0.08%	0.3%	0%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 17 Efficiency of each adjustment method in the BigLucy dataset simulation

Adjustment method								
Calibration	0.026							
	GLM	XGBoost	Ridge reg.					
Model-based	0.023	0.001	0.024					
Model-assisted	0.023	0.001	0.024					
Model-calibrated	0.023	0.001	0.024					
Statistical Matching	0.505	0.491						
Doubly Robust	0.544	0.491						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	1.221	0.412	1.025	0.664	0.767	3.688	0.532	1.358
PSA ($w = (1-\pi)/\pi$)	1.728	0.363	1.289	0.566	0.777	4.056	0.543	1.414
PSA (w^{Strat1})	2.093	5.498	1.805	1.830	0.866	7.384	4.046	0.750
PSA (w^{Strat2})	1.302	0.454	1.040	0.813	0.654	3.673	0.430	3.922
PAPP	0.582	1.758	0.551	0.671	1.372	7.799	1.627	3.122
KW	0.574	6.566	0.560	0.494	0.465	2.746	0.592	2.263
PSA (w $= 1/\pi$) + Matching	0.509	0.491	0.510	0.492	0.522	0.492	0.524	0.490
PSA (w $= (1-\pi)/\pi$) + Matching	0.518	0.492	0.519	0.493	0.521	0.495	0.523	0.491
PSA (w^{Strat1}) + Matching	0.539	0.509	0.509	0.491	0.542	0.511	0.565	0.492
PSA (w^{Strat2}) + Matching	0.520	0.492	0.520	0.491	0.564	0.491	0.566	0.492
KW + Matching	0.521	0.496	0.542	0.492	0.512	0.499	0.534	0.492
PAPP + Matching	0.518	0.494	0.524	0.492	0.519	0.498	0.526	0.493

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 18 Percentage of relative bias of each adjustment method in the Diabetes dataset simulation

Adjustment method								
Unweighted	6.86%							
Calibration	1.47%							
	GLM	XGBoost	Ridge reg.					
Model-based	1.51%	0.71%	0.47%					
Model-assisted	1.51%	0.71%	0.47%					
Model-calibrated	1.34%	16.09%	0.47%					
Statistical Matching	1.48%	0.59%						
Doubly Robust	0.21%	0.69%						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	0.37%	4.24%	3.36%	1.18%	0.92%	0.84%	2.48%	3.1%
PSA ($w = (1-\pi)/\pi$)	6.7%	2.52%	0.41%	5.67%	0.84%	1.49%	2.42%	3.33%
PSA (w^{Strat1})	6.89%	18.26%	1.52%	7.52%	0.64%	7.24%	0.82%	3.45%
PSA (w^{Strat2})	0.48%	3.88%	2.18%	1.04%	0.25%	0.94%	2.23%	3.55%
PAPP	0.97%	2.94%	2.25%	0.26%	3.63%	1.3%	7.02%	1.01%
KW	1.35%	15.32%	1.91%	5.67%	3.09%	1.06%	2.03%	2.09%
PSA (w = $1/\pi$) + Matching	0.88%	1.51%	0.79%	1.99%	0.76%	3.67%	0.76%	3.89%
PSA (w = $(1-\pi)/\pi$) + Matching	0.52%	2.98%	0.73%	3.46%	0.68%	4.2%	0.73%	4.25%
PSA (w^{Strat1}) + Matching	0.24%	8.25%	0.09%	5.48%	0.05%	5.62%	0.21%	4.83%
PSA (w^{Strat2}) + Matching	0.66%	1.61%	0.76%	2.17%	0.62%	3.75%	0.81%	4.23%
KW + Matching	0.61%	5.54%	0.85%	3.97%	0.91%	4.03%	1.07%	4.09%
PAPP + Matching	0.56%	2.18%	0.92%	2.63%	0.56%	3.08%	1.04%	3.35%

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Table 19 Efficiency of each adjustment method in the Diabetes dataset simulation

Adjustment method								
Calibration	0.368							
	GLM	XGBoost	Ridge reg.					
Model-based	0.434	0.471	6.931					
Model-assisted	0.434	0.471	6.931					
Model-calibrated	0.480	5.047	6.931					
Statistical Matching	0.499	0.641						
Doubly Robust	1.040	0.990						
	Weighted model = No				Weighted model = Yes			
	Var. selection = No		Var. selection = Yes		Var. selection = No		Var. selection = Yes	
	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost	GLM	XGBoost
PSA ($w = 1/\pi$)	1.020	0.599	0.692	0.480	1.599	1.245	0.551	0.841
PSA ($w = (1-\pi)/\pi$)	4.035	0.868	1.380	1.867	1.615	1.448	0.552	0.892
PSA (w^{Strat1})	1.925	36.794	2.149	4.590	0.820	11.217	1.546	3.070
PSA (w^{Strat2})	0.453	0.590	0.412	0.403	0.560	1.476	0.469	0.873
PAPP	0.568	0.649	0.453	0.560	1.772	1.650	1.426	0.744
KW	0.359	11.676	0.418	1.919	0.511	0.820	0.427	0.680
PSA (w = $1/\pi$) + Matching	0.759	0.752	0.456	0.798	0.657	1.293	0.535	1.162
PSA (w = $(1-\pi)/\pi$) + Matching	1.041	1.021	1.085	1.028	0.751	1.514	0.618	1.241
PSA (w^{Strat1}) + Matching	0.701	3.660	0.991	1.628	1.054	2.727	1.111	1.592
PSA (w^{Strat2}) + Matching	0.632	0.781	0.488	0.829	0.488	1.356	0.495	1.209
KW + Matching	0.490	1.947	0.495	1.188	1.317	1.737	0.581	1.219
PAPP + Matching	0.860	0.972	0.480	0.940	0.729	1.431	0.537	1.168

GLM: logistic regression if the method involves propensity estimation (PSA, PAPP, KW), and linear regression if the method involves prediction (Model-based, Model-assisted, Model-calibrated, Statistical Matching, Doubly Robust); w^{Strat1}: weighting in PSA using Lee and Valliant (2009) propensity stratification approach; w^{Strat2}: weighting in PSA using Valliant and Dever (2011) propensity stratification approach

Acknowledgements This research was partially supported by a grant from the Ministry of Science and Innovation (PID2019-106861RB-I00, PDC2022-133293-I00, Spain), Strategic Action in Health (DTS23/00032, Spain), from IMAG-María de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033, and from the Own Plan Research and Transfer, University of Granada (PPJIA2023-030, Spain).

Funding Funding for open access publishing: Universidad de Granada/CBUA.

Data availability Data are available from the authors upon reasonable request.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Use of AI tools The authors declare that they have not used Artificial 14 Intelligence (AI) tools in the creation of this article.

References

Beaumont, J.-F.: A new approach to weighting and inference in sample surveys. Biometrika **95**(3), 539–553 (2008)

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

Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20(1996)

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

Boyd, R.J., Stewart, G.B., Pescott, O.L.: Descriptive inference using large, unrepresentative nonprobability samples: an introduction for ecologists. Ecology **105**(2), 4214 (2024)

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T.: Variable selection for propensity score models. Am. J. Epidemiol. **163**(12), 1149–1156 (2006)

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

Cassel, C.M., Särndal, C.E., Wretman, J.H.: Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika **63**(3), 615–620 (1976)

Castro-Martín, L., Rueda, M.D.M., Ferri-García, R., Hernando-Tamayo, C.: On the use of gradient boosting methods to improve the estimation with data obtained with self-selection procedures. Mathematics **9**(23), 2991 (2021)

Castro-Martín, L., Rueda, Md.M., Ferri-García, R.: Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. Mathematics **8**(6), 879 (2020)

Castro-Martín, L., Mar Rueda, M., Ferri-García, R.: Combining statistical matching and propensity score adjustment for inference from non-probability surveys. J. Comput. Appl. Math. **404**, 113414 (2022)

Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

Chen, J., Sitter, R., Wu, C.: Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika **89**(1), 230–237 (2002)
Chen, Y., Li, P., Wu, C.: Doubly robust inference with nonprobability survey samples. J. Am. Stat. Assoc. **115**(532), 2011–2021 (2020)
Chen, S., Yang, S., Kim, J.K.: Nonparametric mass imputation for data integration. J. Surv. Stat. Methodol. **10**(1), 1–24 (2022)
Clore, J., Cios, K., DeShazo, J., Strack, B.: Diabetes 130-US hospitals for years 1999–2008. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5230J (2014)
Dever, J.A., Rafferty, A., Valliant, R.: Internet surveys: can statistical adjustments eliminate coverage bias? Surv. Res. Methods **2**(2), 47–60 (2008)
Deville, J.-C., Särndal, C.-E.: Calibration estimators in survey sampling. J. Am. Stat. Assoc. **87**(418), 376–382 (1992)
Elliot, M.R.: Combining data from probability and non-probability samples using pseudo-weights. Surv. Practice **2**(6) (2009)
Elliott, M.R., Valliant, R.: Inference for nonprobability samples. Stat. Sci. **32**(2), 249–264 (2017)
Elliott, M.R., Resler, A., Flannagan, C.A., Rupp, J.D.: Appropriate analysis of ciren data: using nass-cds to reduce bias in estimation of injury risk factors in passenger vehicle crashes. Accident Anal. & Prevention **42**(2), 530–539 (2010)
Epanechnikov, V.A.: Non-parametric estimation of a multivariate probability density. Theory Probability & Appl. **14**(1), 153–158 (1969)
Ferri-García, R., Rueda, Md.M.: Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PLoS One **15**(4), 0231500 (2020)
Ferri-García, R., Rueda, Md.M.: Variable selection in propensity score adjustment to mitigate selection bias in online surveys. Stat. Pap. **63**(6), 1829–1881 (2022)
Ferri-García, R., Castro-Martín, L., Mar Rueda, M.: Evaluating machine learning methods for estimation in online surveys with superpopulation modeling. Math. Comput. Simul. **186**, 19–28 (2021)
Ferri-García, R., Beaumont, J.-F., Bosa, K., Charlebois, J., Chu, K.: Weight smoothing for nonprobability surveys. TEST **31**(3), 619–643 (2022)
Folsom, R.E., Singh, A.C.: The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In: Proceedings of the American Statistical Association, Survey Research Methods Section, vol. 598603 (2000)
Hahsler, M., Buchta, C., Gruen, B., Hornik, K.: Arules: mining association rules and frequent itemsets. (2022). R package version 1.7-5. https://CRAN.R-project.org/package=arules
Haziza, D., Beaumont, J.-F.: Construction of weights in surveys: a review. Stat. Sci. **32**(2), 206–226 (2017)
Hirano, K., Imbens, G.W.: Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv. Outcomes Res. Method. **2**, 259–278 (2001)
Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**(260), 663–685 (1952)
Kelly, M., Longjohn, R., Nottingham, K.: The UCI machine learning repository. https://archive.ics.uci.edu, Last accessed on 2024 Jan 10
Kern, C., Li, Y., Wang, L.: Boosted kernel weighting-using statistical learning to improve inference from nonprobability samples. J. Surv. Stat. Methodol. **9**(5), 1088–1113 (2021)
Kim, J.K., Park, S., Chen, Y., Wu, C.: Combining non-probability and probability survey samples through mass imputation. J. R. Stat. Soc. Ser. A Stat. Soc. **184**(3), 941–963 (2021)
Kott, P.S.: Using calibration weighting to adjust for nonresponse and coverage errors. Surv. Methodol. **32**(2), 133 (2006)
Lee, S.: Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J. Off. Stat. **22**(2), 329–349 (2006)
Lee, S., Valliant, R.: Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Soc. Methods & Res. **37**(3), 319–343 (2009)
Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
Madow, W.G.: On the theory of systematic sampling, ii. Ann. Math. Stat. **20**(3), 333–354 (1949)
Meng, X.-L.: 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 (2018)

Mercer, A., Lau, A., Kennedy, C.: For weighting online opt-in samples, what matters most? Pew Research Center (2018). https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/

Moro, S., Rita, P., Cortez, P.: Bank marketing. UCI machine learning repository. https://doi.org/10.24432/C5K306(2012)

Neyman, J.: On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. R. Stat. Soc. Ser. A Stat. Soc. **97**(4), 558–606 (1934)

Pfeffermann, D.: The role of sampling weights when modeling survey data. Int. Stat. Rev. **61**, 317–337 (1993)

Rafei, A., Elliott, M.R., Flannagan, C.A.: Robust and efficient bayesian inference for non-probability samples. arXiv preprint arXiv:2203.14355 (2022)

Rafei, A., Flannagan, C.A., Elliott, M.R.: Big data for finite population inference: applying quasi-random approaches to naturalistic driving data using bayesian additive regression trees. J. Surv. Stat. Methodol. **8**(1), 148–180 (2020)

Rivers, D.: Sampling for web surveys. In: joint statistical meetings, vol. 4. American Statistical Association Alexandria, VA (2007)

Robins, J.M.: Correcting for non-compliance in randomized trials using structural nested mean models. Commun. Stat.-Theory Methods **23**(8), 2379–2412 (1994)

Rojas, H.A.G.: TeachingSampling: selection of samples and parameter estimation in finite population. (2020). R package version 4.1.1. https://CRAN.R-project.org/package=TeachingSampling

Royall, R.M.: On finite population sampling theory under certain linear regression models. Biometrika **57**(2), 377–387 (1970)

Royall, R.M., Herson, J.: Robust estimation in finite populations i. J. Am. Stat. Assoc. **68**(344), 880–889 (1973)

Särndal, C.-E., Swensson, B., Wretman, J.: Model assisted survey sampling. Springer Series in Statistics (1992)

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

Skinner, C.J., Holt, D., Smith, T.M.F.: Analysis of Complex Surveys. New York, ??? (1989)

Smith, T.M.: Post-stratification. J. Royal Stat. Soc. Ser. D: Stat. **40**(3), 315–323 (1991)

Tillé, Y., Matei, A.: Sampling: survey sampling. (2021). R package version 2.9. https://CRAN.R-project.org/package=sampling

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

Valliant, R., Dever, J.A.: Estimating propensity adjustments for volunteer web surveys. Soc. Methods & Res. **40**(1), 105–137 (2011)

Wang, L., Graubard, B.I., Katki, H.A., Li, Y.: Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. J. R. Stat. Soc. Ser. A Stat. Soc. **183**(3), 1293–1311 (2020)

Wu, C.: Statistical inference with non-probability survey samples. Surv. Methodol. **48**(2), 283–311 (2022)

Wu, C., Sitter, R.R.: A model-calibration approach to using complete auxiliary information from survey data. J. Am. Stat. Assoc. **96**(453), 185–193 (2001)

Yang, S., Kim, J.K.: Integration of survey data and big observational data for finite population inference using mass imputation. arXiv preprint arXiv:1807.02817 (2018)

Yang, S., Kim, J.K., Song, R.: Doubly robust inference when combining probability and non-probability samples with high dimensional data. J. R. Stat. Soc. Ser. B Stat Methodol. **82**(2), 445–465 (2020)

Authors and Affiliations

Jorge Luis Rueda-Sánchez[1] · Ramón Ferri-García[1] · María del Mar Rueda[1] · Beatriz Cobo[2]

✉ María del Mar Rueda
mrueda@ugr.es

Jorge Luis Rueda-Sánchez
jorgerueda@ugr.es

Ramón Ferri-García
rferri@ugr.es

Beatriz Cobo
beacr@ugr.es

[1] Department of Statistics and Operations Research, University of Granada, Avenida Fuentenueva, s/n, Granada 18017, Granada, Spain

[2] Department of Quantitative Methods for Economics and Business, University of Granada, Campus Universitario de Cartuja, Granada 18071, Granada, Spain