TEST
https://doi.org/10.1007/s11749-021-00795-7

ORIGINAL PAPER

Weight smoothing for nonprobability surveys

Ramón Ferri-García[1] · Jean-François Beaumont[2] · Keven Bosa[2] · Joanne Charlebois[2] · Kenneth Chu[2]

Received: 19 March 2021 / Accepted: 21 November 2021

Abstract

Adjustment techniques to mitigate selection bias in nonprobability samples often involve modelling the propensity to participate in the nonprobability sample along with inverse propensity weighting. It is well known that procedures for estimating weights are effective if the covariates selected in the propensity model are related to both the variable of interest and the participation indicator. In most surveys, there are many variables of interest, making weight adjustments difficult to determine as a suitable weight for one variable may be unsuitable for other variables. The standard compromise is to include a large number of covariates in the propensity model but this may increase the variability of the estimates, especially when some covariates are weakly related to the variables of interest. Weight smoothing, developed for probability surveys, could be helpful in these situations. It aims to remove the variability caused by overfit propensity models by replacing the inverse propensity weights with predicted weights obtained using a smoothing model. In this article, we study weight smoothing in the nonprobability survey context, both theoretically and empirically, to understand its effectiveness at improving the efficiency of estimates.

Keywords Nonprobability samples · Propensity score adjustment · Tree-based inverse propensity-weighted estimator · Weight smoothing

Mathematics Subject Classification 62D05

1 Introduction

Probability sampling has been the gold standard for empirical research since its development in the XX$^{\text{th}}$ century based on the work of Neyman (1934) and Horvitz and

✉ Ramón Ferri-García
rferri@ugr.es

[1] Department of Statistics and Operational Research, University of Granada, Granada, Spain

[2] Statistics Canada, Ottawa, Canada

Published online: 20 December 2021

Thompson (1952) among others. For a sample to be considered probabilistic and therefore valid for population inferences, it must be drawn under the assumption that all the individuals in the target population have a known and non-null inclusion probability. If any of these conditions do not apply, we have a nonprobability sample instead. The use of such samples in empirical sciences is widespread nowadays thanks to technological development and social media, which allows pollsters and vendors to use new questionnaire administration methods such as online and smartphone surveys. These surveys are usually administered via opt-in panels or by recruiting volunteers via snowball sampling (see Schonlau and Couper 2017 for an extensive review of methods).

Nonprobability survey methods offer several advantages over the traditional ones: critical reduction in costs and time to accomplish the fieldwork (Bonsjak and Tuten 2003; Greenlaw and Brown-Welty 2009; Díaz de Rada 2012), and larger sample sizes in comparison with traditional methods which are experiencing a decrease in response rates (Kohut et al. 2012). On the other hand, nonprobability sampling induces a selection bias in the estimates, as the participants (or sample individuals) can differ substantially from nonparticipants (Elliott and Valliant 2017).

Several methods are available to reduce selection bias when a probability sample from the same target population is available. Here, we mention Propensity Score Adjustment (PSA), including the tree-based inverse propensity-weighted (TrIPW) estimator proposed by Chu and Beaumont (2019), statistical matching (also referred to as sample matching), as well as doubly robust estimators that combine statistical matching ideas with PSA.

PSA was originally developed to mitigate selection bias in nonrandomized clinical trials (Rosenbaum and Rubin 1983), and it was adapted to the survey nonresponse field shortly after (Little 1986). PSA adapted to the nonprobability survey context as a method to mitigate selection bias was developed by Lee (2006) and Lee and Valliant (2009). With the PSA method, propensities to participate in a nonprobability sample are estimated via classical modelling using a probability sample drawn from the same population. The TrIPW estimator is an extension of the PSA estimator proposed by Chen et al. (2020), where propensities are estimated using a weighted version of the Classification And Regression Trees (CART) methodology (Breiman et al. 1984). The CART algorithm builds a tree that optimizes an homogeneity measure, given a set of covariates, which is then used to estimate propensities.

When the propensity model is properly specified, PSA is able to reduce bias of nonprobability sample estimates at the potential cost of increasing their variability (Lee 2006; Lee and Valliant 2009; Valliant and Dever 2011; Ferri-García and Rueda 2018). The TrIPW estimator shows itself as a more robust estimator under complex relationships between variables, such as nonlinearities (Chu and Beaumont 2019) and the presence of interactions. An alternative is to pool the probability and nonprobability samples, similar to Lee (2006), and to use machine learning algorithms to model propensities (Ferri-García and Rueda 2020).

Statistical matching focuses on another model-based approach whose objective is to predict the unobserved values of the variable of interest in the probability sample. The predictive model is fitted using data from the nonprobability sample. Statistical matching has also been proven to mitigate selection bias in nonprobability samples

(Castro-Martín et al. 2020). The combination of both strategies via doubly robust estimators may outperform both approaches on their own Chen et al. (2020).

Despite the benefits of statistical matching techniques, they could have limitations in surveys that collect multiple variables of interest. In those surveys, which are common in practice, each variable of interest may require a specific model to predict its unobserved values in the probability sample. This could become cumbersome if the number of variables of interest is large and increase the risk of model misspecifications. The use of weighted estimators, such as the PSA and TrIPW estimators, could provide a reasonable solution, as the same vector of weights would be used to obtain estimates for all of the variables of interest. However, research has shown that propensity techniques are more efficient when the covariates used for modelling the propensities are related to the outcome variables, that is, the variables of interest (Hirano and Imbens 2001; Brookhart et al. 2006). In a survey with multiple variables of interest, a suitable set of covariates may vary between variables. The standard compromise is to include a large number of covariates in the propensity model. This may increase the variability of the resulting estimates due to overfitting, especially when the covariates are weakly related to the variables of interest.

In probability surveys, weight smoothing (Beaumont 2008) has been shown to be effective at reducing the variance of survey-weighted estimators by modelling the survey weights conditional on the variables of interest. The variance of survey-weighted estimators can be large when the design variables are unrelated to the variables of interest. To the best of our knowledge, this technique has not been evaluated in a nonprobability survey context, where the inclusion (or participation) probabilities are unknown and estimated. The objective of this study is to examine the adequacy of weight smoothing for nonprobability surveys, both theoretically and empirically, and explore the situations that could enhance its efficiency.

2 Weighting in nonprobability surveys

Let U be a target population of size N from which we want to estimate a population parameter, such as the population mean $\bar{Y} = N^{-1}\sum_{i\in U} y_i$, for a given variable of interest y. To this end, we obtain a nonprobability sample s_v of size n_v from the population U. The participation mechanism may depend on features such as self-selection or device availability (computer, internet access, etc.). In this case, the probability that an individual $i \in U$ is included in s_v is not known a priori.

Let R_i be the indicator variable which measures whether a given individual $i \in U$ has participated or not. We assume that R_i is related to a vector of covariates, $\mathbf{x}_i$ (e.g. demographic variables such as region, age and sex, or education), and that participation is not informative, i.e. R_i does not depend on y_i after conditioning on $\mathbf{x}_i$. We define the inclusion (or participation) probability as

$$\pi_i = P(R_i = 1|\mathbf{x}_i), i \in U.$$

The participation probability is unknown and assumed to be strictly positive. From these assumptions, if $\mathbf{x}_i$ is known for every $i \in U$, the participation probability can be estimated using standard modelling techniques along with maximum likelihood estimation. However, this information is rarely available. An alternative is to use a probability sample s_r of size n_r, drawn from the full population U that measures $\mathbf{x}_i$ for all sample individuals $i \in s_r$. The design weight d_i^r is also available for sample individuals. The covariates $\mathbf{x}_i$ are assumed to be observed in both the probability and nonprobability samples, whereas y_i is observed only in the nonprobability sample.

With PSA methods, a parametric model $\pi_i = m(\mathbf{x}_i, \boldsymbol{\beta})$ is typically postulated, where m is a known function, such as the logistic function $m(\mathbf{x}_i, \boldsymbol{\beta}) = \left\{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})\right\}^{-1}$, and $\boldsymbol{\beta}$ is a vector of unknown model parameters. Assuming the participation indicators R_i, $i \in U$, are mutually independent, Chen et al. (2020) proposed a pseudo-maximum likelihood estimator of π_i, which is computed as $\hat{\pi}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, where the estimator $\hat{\boldsymbol{\beta}}$ maximizes the pseudo-log-likelihood function

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

with respect to $\boldsymbol{\beta}$. The design expectation of the pseudo-log-likelihood function (1) is equal to the standard log-likelihood function, which cannot be used unless $\mathbf{x}_i$ is observed for all population individuals $i \in U$. The pseudo-log-likelihood function proposed by Chen et al. (2020) may be less efficient but does not require $\mathbf{x}_i$ to be observed in the entire population; it only needs $\mathbf{x}_i$ to be observed for all individuals in s_v and s_r.

The population mean $\bar{Y}$ can be estimated by the Hajek estimator

$$\bar{y}_w^H = \hat{N}_v^{-1} \sum_{i \in s_v} w_i y_i, \tag{2}$$

where $\hat{N}_v = \sum_{i \in s_v} w_i$ and $w_i = \hat{\pi}_i^{-1}, i \in s_v$. Chen et al. (2020) proved the consistency of the weighted estimator $\bar{y}_w^H$ under regularity conditions, i.e. they proved that $\bar{y}_w^H - \bar{Y} = O_p\left(n_v^{-1/2}\right)$.

A number of authors (e.g. Lee 2006; Lee and Valliant 2009) have considered estimating π_i using the pooled sample $s = s_r \cup s_v$ along with a weighted logistic regression. If the input weights for the logistic regression are chosen as

$$d_i^{pool} = \begin{cases} 1 & i \in s_v \\ d_i^r & i \in s_r, \end{cases}$$

the resulting pseudo-log-likelihood function can be written as

$$\tilde{l}(\boldsymbol{\beta}) = \sum_{i \in s_v} \log\left(\frac{m(\mathbf{x}_i, \boldsymbol{\beta})}{1 - m(\mathbf{x}_i, \boldsymbol{\beta})}\right) + \sum_{i \in s_r} d_i^r \log\left(1 - m(\mathbf{x}_i, \boldsymbol{\beta})\right) + \sum_{i \in s_v} \log\left(1 - m(\mathbf{x}_i, \boldsymbol{\beta})\right). \tag{3}$$

It can be observed that (3) is equal to the pseudo-log-likelihood function shown in (1), except for the last term $\sum_{i \in s_v} \log\left(1 - m(\mathbf{x}_i, \boldsymbol{\beta})\right)$. Beaumont (2020) pointed out that if the participation probabilities π_i are all small, which could be plausible when the participation rate is small, maximizing (3) is approximately equivalent to maximizing the pseudo-log-likelihood function (1) when the logistic function is used. As a result, using (1) or (3) should yield similar estimated participation probabilities when the participation rate is small. However, note that the pseudo-log-likelihood function of Chen et al. (2020) does not require this condition and should be the preferred choice when it is not satisfied.

This idea of pooling both samples has recently been used along with nonparametric methods of estimating propensities, such as machine learning classification algorithms (e.g. Ferri-García and Rueda 2020). Similar to logistic regression, these methods are expected to be valid only when the participation rate is small and design weights d_i^r are used appropriately.

The literature also accounts for other procedures to calculate weights from propensities. For instance, the original literature on PSA for nonprobability samples (Lee 2006; Lee and Valliant 2009) considered the stratification of propensities into g partitions, usually $g = 5$ following the criteria of Cochran (1968), and the calculation of weights using a correction factor that takes into account the original design weights. Valliant and Dever (2011) considered a similar approach that also involves stratification of propensities. The use of propensity strata may provide some robustness to misspecifications of the logistic model and may reduce the occurrence of extreme weights.

The TrIPW estimator of $\bar{Y}$, developed in Chu and Beaumont (2019), takes the same form as the Hajek estimator (2), but the estimation of the participation probability π_i is based on an adaptation of the Classification And Regression Trees (CART) algorithm (Breiman et al. 1984). This adaptation accounts for the design weights d_i^r, $i \in s_r$, in a way similar to Chen et al. (2020). As a result, it does not require the participation rate to be small. After the tree has been grown, using an objective function that accounts for the design weights, the nonprobability sample s_v is partitioned into G exhaustive and nonoverlapping homogeneous propensity groups (or terminal nodes), $s_{v,g}$, $g = 1, ..., G$. The probability sample is partitioned similarly using the same decision rules into G exhaustive and nonoverlapping groups $s_{r,g}$, $g = 1, ..., G$. The propensity for each individual $i \in s_{v,g}$ is estimated as:

$$\hat{\pi}_i = \frac{n_{v,g}}{\hat{N}_g}, g = 1, ..., G, \tag{4}$$

where $n_{v,g}$ is the number of individuals in the nonprobability sample who fall in propensity group g and $\hat{N}_g = \sum_{i \in s_{r,g}} d_i^r$ is the estimated population size of group g. The estimated probability (4) can be obtained using the Chen et al. (2020) method by defining $\mathbf{x}_i$ as a G-vector indicating to which group individual i belongs. The creation of homogeneous propensity groups using this weighted CART algorithm is expected to provide some robustness to misspecification of the logistic model. This was shown by Chu and Beaumont (2019) in a simulation study.

3 Weight smoothing

The application of weighting methods discussed in Sect. 2 can reduce significantly the participation bias at the possible cost of increasing the variance of the estimates (Lee 2006; Lee and Valliant 2009; Ferri-García and Rueda 2018). This variance is directly tied to the variability of the weights and is amplified when the covariates are weakly associated with the variable of interest. In that case, weighting increases the variance without the benefit of a significant bias reduction. Therefore, it seems reasonable to focus on strategies that reduce the variability of the weights. Beaumont (2008) proposed weight smoothing for probability samples. Our objective is to study this method in the context of nonprobability samples.

Let us assume for a moment that the logistic model holds and that $\boldsymbol{\beta}$ is known. Assuming N is also known, the population mean $\bar{Y}$ can be estimated by $\bar{y}_w^{\boldsymbol{\beta}} = N^{-1}\sum_{i\in s_v} w_i^{\boldsymbol{\beta}} y_i$, where $w_i^{\boldsymbol{\beta}} = [m(\mathbf{x}_i, \boldsymbol{\beta})]^{-1}$. The superscript $\boldsymbol{\beta}$ is used to indicate that $\boldsymbol{\beta}$ is known and to distinguish this case from the one considered throughout this paper, where $\boldsymbol{\beta}$ is estimated. The weighted estimator $\bar{y}_w^{\boldsymbol{\beta}}$ is unbiased in the sense that $E\left(\bar{y}_w^{\boldsymbol{\beta}} \mid \mathbf{X}, \mathbf{Y}\right) = \bar{Y}$, where $\mathbf{X}$ is the N-row matrix formed by the row vectors $\mathbf{x}_i^{\top}$, $i \in U$, and $\mathbf{Y}$ is the N-vector of population y values. Under these assumptions, the nonprobability sample can be viewed as a Poisson sample with known inclusion probabilities $\pi_i = m(\mathbf{x}_i, \boldsymbol{\beta})$, and the original weight smoothing method of Beaumont (2008) can be directly applied to improve the efficiency of $\bar{y}_w^{\boldsymbol{\beta}}$. It consists of replacing the weight $w_i^{\boldsymbol{\beta}}$ with the smoothed weight $\tilde{w}_i^{\boldsymbol{\beta}} = E\left(w_i^{\boldsymbol{\beta}} \mid s_v, \mathbf{Y}\right)$. The basic idea is to extract from the weight $w_i^{\boldsymbol{\beta}}$ its relevant component, i.e. the component that is associated with the variable of interest. Assuming the smoothed weight $\tilde{w}_i^{\boldsymbol{\beta}}$ is known, the population mean $\bar{Y}$ is estimated by $\bar{y}_{\tilde{w}}^{\boldsymbol{\beta}} = N^{-1}\sum_{i\in s_v} \tilde{w}_i^{\boldsymbol{\beta}} y_i$. Beaumont (2008) noted that $E\left(\bar{y}_{\tilde{w}}^{\boldsymbol{\beta}} \mid \mathbf{Y}\right) = \bar{Y}$ and that

$$var\left(\bar{y}_{\tilde{w}}^{\boldsymbol{\beta}} \mid \mathbf{Y}\right) \leq var\left(\bar{y}_w^{\boldsymbol{\beta}} \mid \mathbf{Y}\right). \tag{5}$$

The smoothed weight $\tilde{w}_i^{\boldsymbol{\beta}} = E\left(w_i^{\boldsymbol{\beta}} \mid s_v, \mathbf{Y}\right)$ is generally unknown but can be estimated from sample data by modelling $w_i^{\boldsymbol{\beta}}$ given y_i, $i \in s_v$. Beaumont (2008) showed that the resulting estimator of $\bar{Y}$ remains no less efficient than $\bar{y}_w^{\boldsymbol{\beta}}$ under a linear model. Note that inferences under weight smoothing are conditional on $\mathbf{Y}$ alone. As a result, $\mathbf{x}_i$ is viewed as random, as well as $w_i^{\boldsymbol{\beta}}$, but only the latter needs to be modelled.

In multipurpose surveys, there are multiple variables of interest so that y_i is a vector and $\mathbf{Y}$ is a matrix. The weight smoothing methodology can be extended in a straightforward manner to a vector of variables of interest by modelling the weight $w_i^{\boldsymbol{\beta}}$ conditional on the full vector of variables of interest. However, if the number of y variables is large, it may be expected that together they become strongly predictive of the weight $w_i^{\boldsymbol{\beta}}$, thereby reducing the potential efficiency gains. An alternative would be to determine a specific weight smoothing model for each variable of interest, but this would lead to multiple smoothed weights, which may not be attractive to the ultimate users.

With nonprobability samples, the model parameters $\boldsymbol{\beta}$ are unknown but can be estimated using the Chen et al. (2020) pseudo-likelihood method discussed in Sect. 2. This yields the weight $w_i = \left[m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\right]^{-1}$. Assuming N is known, the estimator of the population mean $\bar{Y}$ is $\bar{y}_w = N^{-1}\sum_{i\in s_v} w_i y_i$. We define the smoothed weight $\tilde{w}_i = E\left(w_i \mid s_v, \mathbf{Y}\right)$ along with the smoothed estimator of $\bar{Y}$, $\bar{y}_{\tilde{w}} = N^{-1}\sum_{i\in s_v} \tilde{w}_i y_i$. Similar to (5), it is not difficult to show that

$$E\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) = \bar{Y} + o\left(n_v^{-1/2}\right) \text{ and } var\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) \leq var\left(\bar{y}_w \mid \mathbf{Y}\right). \tag{6}$$

The proof of (6) is given in the appendix. Again, the smoothed weight $\tilde{w}_i = E\left(w_i \mid s_v, \mathbf{Y}\right)$ is generally unknown but can be estimated from sample data by modelling w_i given $y_i, i \in s_v$. For instance, consider the linear model

$$E\left(w_i \mid s_v, \mathbf{Y}\right) = \mathbf{h}_i^\top \boldsymbol{\gamma} \text{ with } var\left(w_i \mid s_v, \mathbf{Y}\right) = \sigma^2, i \in s_v, \tag{7}$$

where the vector of predictors $\mathbf{h}_i$ is a function of the variable(s) of interest y_i, and $\boldsymbol{\gamma}$ and σ^2 are unknown model parameters. The smoothed weight $\tilde{w}_i$ can be estimated by $\hat{w}_i = \mathbf{h}_i^\top \hat{\boldsymbol{\gamma}}$, where

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i\in s_v} \mathbf{h}_i \mathbf{h}_i^\top\right)^{-1} \sum_{i\in s_v} \mathbf{h}_i w_i$$

is the least square estimator of $\boldsymbol{\gamma}$. The smoothed estimator of $\bar{Y}$ becomes $\bar{y}_{\hat{w}} = N^{-1}\sum_{i\in s_v} \hat{w}_i y_i$. After straightforward algebra, it can be shown that

$$\bar{y}_{\hat{w}} = N^{-1}\sum_{i\in s_v} \hat{w}_i y_i = N^{-1}\sum_{i\in s_v} \left(\mathbf{h}_i^\top \hat{\boldsymbol{\gamma}}\right) y_i = N^{-1}\sum_{i\in s_v} w_i \hat{y}_i, \tag{8}$$

where $\hat{y}_i = \mathbf{h}_i^\top \hat{\boldsymbol{\alpha}}$ is a predicted value of y_i with

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{i\in s_v} \mathbf{h}_i \mathbf{h}_i^\top\right)^{-1} \sum_{i\in s_v} \mathbf{h}_i y_i. \tag{9}$$

Therefore, Eq. (8) indicates that smoothing the weight w_i using the predictors $\mathbf{h}_i$ is equivalent to smoothing y_i using the same predictors. It can also be shown that $E\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = \bar{Y} + o\left(n_v^{-1/2}\right)$ and that

$$var\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = var\left(\bar{y}_w \mid \mathbf{Y}\right) - \frac{\sigma^2}{N^2} E\left[\sum_{i\in s_v} \left(y_i - \hat{y}_i\right)^2 \mid \mathbf{Y}\right]. \tag{10}$$

The proof of (10) is given in the appendix. It confirms that the smoothed estimator $\bar{y}_{\hat{w}}$ is never less efficient than $\bar{y}_w$ when the linear model (7) holds. The magnitude of efficiency gains from weight smoothing depends in part on the strength of the relationship between the weight w_i and the predictors $\mathbf{h}_i$. A weak relationship, and thus a large model variance σ^2, will tend to increase efficiency gains. The efficiency gains will also tend to be larger when $\mathbf{h}_i$ is not a strong predictor of y_i. Instead, if $\mathbf{h}_i$ is a perfect predictor of y_i, i.e. there exists a vector $\boldsymbol{\alpha}$ such that $y_i = \mathbf{h}_i^\top \boldsymbol{\alpha}$, then it can be easily shown that $\hat{y}_i = y_i$, $i \in s_v$, and the efficiency gains entirely vanish. Variance reductions may thus vary from one variable of interest to another depending on the strength of their relationship with $\mathbf{h}_i$. On the one hand, overfitting should be avoided as much as possible when choosing the predictors $\mathbf{h}_i$ to maximize variance reductions. Variable selection techniques, such as Least Absolute Shrinkage and Selection Operator (LASSO), can be useful for this purpose. On the other hand, the predictors $\mathbf{h}_i$ should be chosen to ensure the linear model (7) holds, at least its first moment, to avoid introducing bias in the smoothed estimator of the population mean $\bar{Y}$.

The most favourable situation for weight smoothing is when none of the variables of interest is related to the weight w_i so that $\mathbf{h}_i = 1$ is appropriate. Noting that $E\left(w_i^{\boldsymbol{\beta}} \mid s_v, \mathbf{x}_i\right) = w_i^{\boldsymbol{\beta}}$, the most unfavourable situation would be when the variables of interest are strong predictors of the covariates $\mathbf{x}_i$, and thus the weight $w_i^{\boldsymbol{\beta}}$, so that $\tilde{w}_i^{\boldsymbol{\beta}} \approx w_i^{\boldsymbol{\beta}}$ and basically no variance reduction is possible. In particular, this would occur in the extreme and unlikely scenario where all the covariates would also be variables of interest.

An estimator of the variance (10) requires estimating $var\left(\bar{y}_w \mid \mathbf{Y}\right)$. Under regularity conditions given in Chen et al. (2020),

$$var\left(\bar{y}_w \mid \mathbf{Y}\right) = E\left[var\left(\bar{y}_w \mid \mathbf{X}, \mathbf{Y}\right) \mid \mathbf{Y}\right] + o\left(n_v^{-1}\right). \tag{11}$$

The variance (11) can thus be estimated by estimating the conditional variance $var\left(\bar{y}_w \mid \mathbf{X}, \mathbf{Y}\right)$. Chen et al. (2020) proposed linearization and bootstrap estimators of this conditional variance. We denote a consistent estimator of $var\left(\bar{y}_w \mid \mathbf{X}, \mathbf{Y}\right)$ by $v\left(\bar{y}_w\right)$. A plug-in estimator of the variance (10) is thus

$$v\left(\bar{y}_{\hat{w}}\right) = v\left(\bar{y}_w\right) - \frac{(n_v - p)}{N^2}\hat{\sigma}^2 \frac{\sum_{i \in s_v}\left(y_i - \hat{y}_i\right)^2}{n_v - p}, \tag{12}$$

where p is the number of predictors in the vector $\mathbf{h}_i$ and

$$\hat{\sigma}^2 = \frac{\sum_{i \in s_v}\left(w_i - \mathbf{h}_i^\top \hat{\boldsymbol{\gamma}}\right)^2}{n_v - p}.$$

Let us now consider the Hajek estimators $\bar{y}_w^H = \hat{N}_v^{-1} \sum_{i \in s_v} w_i y_i$ and $\bar{y}_{\hat{w}}^H = \left(\sum_{i \in s_v} \hat{w}_i\right)^{-1} \sum_{i \in s_v} \hat{w}_i y_i$. Using a first-order Taylor linearization and assuming an intercept is included in the vector of predictors $\mathbf{h}_i$, it can be shown that $var\left(\bar{y}^H_{\hat{w}}\right)$ can be approximated as

$$var\left(\bar{y}^H_{\hat{w}} \mid \mathbf{Y}\right) \approx var\left(\bar{y}^H_{w} \mid \mathbf{Y}\right) - \frac{\sigma^2}{N^2} E\left[\sum_{i \in s_v} \left(y_i - \hat{y}_i\right)^2 \mid \mathbf{Y}\right]. \qquad (13)$$

As a result, the variance reduction for the Hajek estimator $\bar{y}^H_{\hat{w}}$ is asymptotically the same as the variance reduction for $\bar{y}_{\hat{w}}$. Similar to (12), an estimator of the variance (13) can be obtained as

$$v\left(\bar{y}^H_{\hat{w}}\right) = v\left(\bar{y}^H_{w}\right) - \frac{(n_v - p)}{\hat{N}^2_v} \hat{\sigma}^2 \frac{\sum_{i \in s_v} \left(y_i - \hat{y}_i\right)^2}{n_v - p},$$

where $v\left(\bar{y}^H_{w}\right)$ is a consistent estimator of the conditional variance $var\left(\bar{y}^H_{w} \mid \mathbf{X}, \mathbf{Y}\right)$, such as the linearization and bootstrap estimators proposed in Chen et al. (2020).

The variance expressions (10) and (13) are valid only for the linear smoothing model (7). In practice, a linear model may not always hold even after accounting for interactions and/or polynomial effects. Nonlinear smoothing models could also be considered. For variance estimation under nonlinear models, Beaumont (2008) proposed two bootstrap methods that could be adapted to the context of nonprobability surveys. In our simulation studies, described in the next sections, we evaluate the prediction algorithm XGBoost as an alternative to the linear model (7) for the estimation of smoothed weights. The development of theoretical properties of XGBoost for weight smoothing is beyond the scope of this paper.

4 Data

4.1 Artificial data

We created a population of size $N = 500,000$ with 10 covariates $(x_1, ..., x_{10})$, 10 variables of interest $(y_1, ..., y_{10})$ and a variable π_i indicating the propensity of each individual to participate in a volunteer sample. The covariates were generated from Bernoulli and Normal distributions as follows:

$$\begin{aligned} &x_1, x_4, x_7 \sim Be(0.5),\\ &x_3, x_6, x_9 \sim Be(0.2) \text{ and}\\ &x_2, x_5, x_8, x_{10} \sim N(0, 1). \end{aligned}$$

Nonprobability samples of size n_v were selected without replacement with probabilities proportional to

$$\pi^*_i = \frac{exp(-0.5 + 2.5x_7 + \sqrt{2\pi}x_8 - (11/3)x_9)}{1 + exp(-0.5 + 2.5x_7 + \sqrt{2\pi}x_8 - (11/3)x_9)}, \quad i = 1, 2, ..., 500, 000,$$

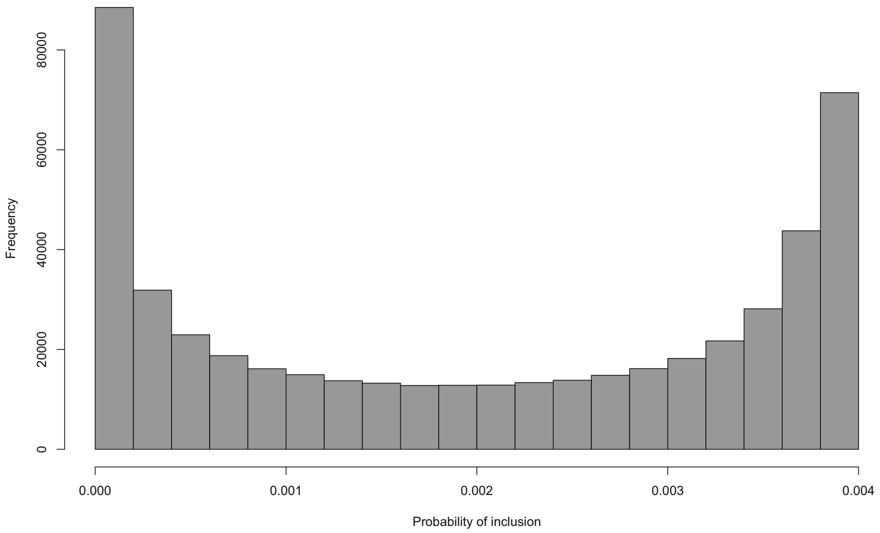

Fig. 1 Histogram of the population propensities

using the sample() function in R. Therefore, the participation probabilities π_i depend on the values of x_7, x_8 and x_9 and are approximately equal to

$$\pi_i \approx n_v \frac{\pi_i^*}{\sum_{i \in U} \pi_i^*}.$$

This participation mechanism was intended to create weights with high variability, a situation where the advantages of weight smoothing might be more visible. The histogram of propensities $\pi_i, i \in U$, is provided in Fig. 1; the mean propensity is 0.002, with a standard deviation of 0.00147, and thus a coefficient of variation of 0.7351. The first and third quartiles are 0.00044 and 0.00354, respectively.

The variables of interest were created to have different relationships with the covariates and the propensities according to two scenarios:

Sc. 1. No relationship between any variable in $(y_1, ..., y_{10})$ and π
Sc. 2. Relationship between every variable in $(y_1, ..., y_{10})$ and π

Scenario 1 is favourable to weight smoothing, whereas Scenario 2 is unfavourable. In practice, we may expect to have a hybrid between these two scenarios, where some but not all covariates that explain π_i are unrelated to the variables of interest.

The generation of the variables of interest was performed according to the following formulas:

$$y_1 \sim B\left(\frac{exp(-1 + 3x_1 + x_2 + x_3 + \mathbf{1}_{\text{Sc. }2}5\pi)}{1 + exp(-1 + 3x_1 + x_2 + x_3 + \mathbf{1}_{\text{Sc. }2}5\pi)}\right),$$
$$y_2 \sim N(0, 1) - 1 + 3x_1 + x_2 + x_3 + \mathbf{1}_{\text{Sc. }2}5,$$

Table 1 Population Pearson's correlation coefficients between π and $(y_1, ..., y_{10})$ in Scenarios 1 and 2

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
Scenario 1	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.0	0.00	0.0
Scenario 2	0.39	0.66	−0.43	0.18	0.6	0.16	−0.18	−0.6	−0.12	0.6

$$y_3 \sim B\left(\frac{exp(-1 + x_1 + x_2 + 3x_3 - \mathbf{1}_{\text{Sc. }2}5\pi)}{1 + exp(-1 + x_1 + x_2 + 3x_3 - \mathbf{1}_{\text{Sc. }2}5\pi)}\right),$$
$$y_4 \sim B\left(\frac{exp(\mathbf{1}_{\text{Sc. }2}\pi)}{1 + exp(\mathbf{1}_{\text{Sc. }2}\pi)}\right),$$
$$y_5 \sim N(0, 1) + 2 + \mathbf{1}_{\text{Sc. }2}2\pi,$$
$$y_6 \sim B\left(\frac{exp(-1.5 + \mathbf{1}_{\text{Sc. }2}\pi)}{1 + exp(-1.5 + \mathbf{1}_{\text{Sc. }2}\pi)}\right),$$
$$y_7 \sim B\left(\frac{exp(-\mathbf{1}_{\text{Sc. }2}\pi)}{1 + exp(-\mathbf{1}_{\text{Sc. }2}\pi)}\right),$$
$$y_8 \sim N(0, 1) - 2 - \mathbf{1}_{\text{Sc. }2}2\pi,$$
$$y_9 \sim B\left(\frac{exp(-1.5 - \mathbf{1}_{\text{Sc. }2}\pi)}{1 + exp(-1.5 - \mathbf{1}_{\text{Sc. }2}\pi)}\right) \text{ and}$$
$$y_{10} \sim N(2, 1) + \mathbf{1}_{\text{Sc. }2}2\pi,$$

where $\mathbf{1}_{\text{Sc. }2}$ is an indicator variable which takes the value 1 if the simulation is conducted under Scenario 2 and 0, otherwise. It can be observed that we have 6 Bernoulli and 4 Gaussian variables among $(y_1, ..., y_{10})$ whose parameters depend on the scenario. The vector of population means for Scenario 1 is (0.60, 0.70, 0.50, 0.50, 2.00, 0.18, 0.50, -2.00, 0.18, 0.00), while the vector of population means for Scenario 2 is (0.84, 3.25, 0.21, 0.62, 3.02, 0.28, 0.38, -3.02, 0.12, 1.02). Table 1 contains Pearson's correlation coefficients between the propensities π and each variable of interest for both scenarios. We see that the correlation is nonexistent in Scenario 1 and notable for all the variables in Scenario 2, with different levels of strength caused by the limitations of using this measure on binary variables. Table 2 presents the results of t tests of the equality between the means of π for $y_k = 0$ and $y_k = 1$, for $k = 1, 3, 4, 6, 7, 9$, in Scenario 2.

4.2 Real data

The dataset used to experiment in a real life situation comes from the 2012 edition of the Spanish Life Conditions Survey (National Institute of Statistics 2012). This is an annual survey measuring several aspects of life conditions, such as health status, degree of deprivation and employment conditions, in the Spanish adult population. The survey includes specific modules in each edition; in 2012, the module consisted of a battery of questions regarding household conditions. The sampling design follows a stratified cluster scheme, where the primary units are the households and the secondary units are their members. The total sample size in 2012 was $n = 33,579$.

Table 2 Results of t tests of the equality between the means of π for binary classes of y_1, y_3, y_4, y_6, y_7 and y_9 in Scenario 2

k	Mean of π $(y_k = 0)$	Mean of π $(y_k = 1)$	t	p-value
1	0.0007	0.0023	−401.35	<2.2e-16
3	0.0023	0.0008	397.20	<2.2e-16
4	0.0017	0.0022	−128.61	<2.2e-16
6	0.0019	0.0024	−116.84	<2.2e-16
7	0.0022	0.0017	128.70	<2.2e-16
9	0.0021	0.0015	88.17	<2.2e-16

For its use as a pseudopopulation, the sample dataset was filtered to rule out those individuals and variables with high amounts of missing data. This reduced the dataset to $n = 28,210$ and 146 variables, from which 61 were selected for the simulations. The sample was subsequently bootstrapped in order to increase its size to $1,000,000$. Finally, all the individuals who selected any of the refusal options ("Does not know" or "Does not answer") were also ruled out of the analysis to avoid further problems with rare classes. The final pseudopopulation size for the experiments was $N = 990,838$.

For the experiments, we chose HS090 (Owning a computer at home) as the volunteering variable, given that its behaviour would be very similar to a variable measuring access to internet (see Ferri-García and Rueda 2020 for further details on this matter). The extraction of the nonprobability sample s_v was done under two different mechanisms:

- Simple Random Sampling Without Replacement (SRSWOR) from the population who has a computer at home.
- Unequal probability sampling without replacement from the population who has a computer at home, where the probabilities are calculated as

$$\frac{(\text{Year of birth} - 1925)^4}{(1996 - 1925)^4}.$$

Regarding covariates, two different sets were considered:

- A set of nine demographic variables, namely region, urbanization level, number of members of the household and consumption units (weighted mean of the number of members of the household following OECD criteria, where adults have more weight than teenagers and teenagers have more weight than children), sex, marital status, country of birth, nationality, and whether the individual is currently a student or not.
- A set of eight variables related to economic and material deprivation, namely capacity of the household to make ends meet, minimum income required by the household to make ends meet, whether the household has the capacity to go on holiday, have a meat or fish meal at least every two days, and deal with unforeseen expenses, household under the poverty threshold, person under the poverty threshold, and household in a situation of severe material deprivation.

Ten variables were used as variables of interest, the last two being randomly generated:

- y_1 = Household expenses are a heavy burden (ordinal scale, 1-3)
- y_2 = Household has a car (dichotomous)
- y_3 = Self-reported health (5-point Likert scale)
- y_4 = Disability in the previous 6 months (ordinal scale, 1-3)
- y_5 = Number of months working part-time (integer, 0-12)
- y_6 = Household expenses in EUR (continuous)
- y_7 = Household with noise problems (dichotomous)
- y_8 = Household with heating system (dichotomous)
- y_9 = Simulated random Be(0.5) variable
- y_{10} = Simulated random N(2, 1) variable

5 Experimental design and metrics

The settings of the experiment were kept as equal as possible for all simulation scenarios. Each simulation was run 500 times, drawing probability and nonprobability samples of equal sizes ($n_r = n_v = 1,000$) using the nonprobability sampling designs described in the previous section to select s_v. The probability sample s_r was selected using SRSWOR in all scenarios so that $d_i^r = \frac{N}{n_r}$.

Two approaches were applied to estimate nonprobability sample propensities π_i: weighted logistic regression with main effects only, and weighted CART described in Sect. 2 (see Eq. 4) with a fixed minimum cell size of 50 and minimum cell impurity of 0.0001. For weighted logistic regression, the model parameters $\boldsymbol{\beta}$ were estimated by maximizing the pseudo-log-likelihood function (3), which should be approximately equivalent to maximizing the pseudo-log-likelihood function of Chen et al. (2020), given the participation rates are small in our simulation scenarios (0.002 for the artificial data simulation and 0.001 for the real data simulation).

Two methods were considered for the estimation of the smoothed weights $E\left(w_i \mid s_v, \mathbf{Y}\right)$, where the weight $w_i = 1/\hat{\pi}_i$ is obtained using either weighted logistic regression or weighted CART:

- XGBoost algorithm (XGB) using the *xgboost* package in R (Chen and Guestrin 2016).
- Least Absolute Shrinkage and Selection Operator (LASSO) regression using the *glmnet* package in R (Friedman et al. 2010).

All 10 variables of interest (described in the previous section) were considered as potential predictors of w_i for both XGBoost and LASSO. The XGBoost algorithm was trained with default hyperparameters: a L1 regularization term of 0.1, a L2 regularization term of 0.0001 and a learning rate of 0.3. The number of rounds was fixed at 50. In the case of LASSO, w_i was modelled using a linear model with main effects only (no interaction). The optimal shrinkage parameter was obtained with a tenfold cross-validation procedure in each run of the simulation.

Relative measures of Monte Carlo bias and Monte Carlo Mean Square Error (MSE) were calculated to allow for the comparison between seven estimators of the population mean $\bar{Y}$. These seven estimators can be divided into three categories: the naive unweighted (Unw) estimator,

$$\bar{y}^{\mathrm{Unw}} = \frac{\sum_{i \in s_v} y_i}{n_v},$$

the nonsmoothed (NS) Hajek estimator,

$$\bar{y}_w^H = \frac{\sum_{i \in s_v} w_i y_i}{\sum_{i \in s_v} w_i}$$

and the smoothed Hajek estimator

$$\bar{y}_{\hat{w}}^H = \frac{\sum_{i \in s_v} \hat{w}_i y_i}{\sum_{i \in s_v} \hat{w}_i},$$

where $\hat{w}_i$ is the smoothed weight obtained using either XGB or LASSO. The unweighted estimator $\bar{y}^{\mathrm{Unw}}$ is used as a reference for the comparisons. It can be obtained from the smoothed Hajek estimator by replacing $\hat{w}_i$ with the average of w_i over the nonprobability sample individuals. It is the most extreme form of smoothing that can be obtained from the linear model (7) with $\mathbf{h}_i = 1$. It is well known that the unweighted estimator may result in significant biases. There are two versions of the NS estimator $\bar{y}_w^H$, depending on whether $\hat{\pi}_i$ was obtained using a weighted logistic regression, yielding the PSA estimator of $\bar{Y}$, or weighted CART, yielding the TrIPW estimator of $\bar{Y}$. There are two smoothed estimators (XGB and LASSO), and each of them has two versions depending on the estimation method for π_i. There are thus four different smoothed estimators.

Let $\bar{y}^*$ be any of the seven estimators described above. The Monte Carlo bias, standard deviation and MSE of $\bar{y}^*$ are defined as

$$\mathrm{Bias}(\bar{y}^*) = \frac{1}{500} \sum_{j=1}^{500} \bar{y}_j^* - \bar{Y},$$

$$\mathrm{StdDev}(\bar{y}^*) = \sqrt{\frac{1}{499} \sum_{j=1}^{500} \left(\bar{y}_j^* - \frac{1}{500} \sum_{l=1}^{500} \bar{y}_l^* \right)^2}$$

and

$$\mathrm{MSE}(\bar{y}^*) = \mathrm{StdDev}^2(\bar{y}^*) + \mathrm{Bias}^2(\bar{y}^*),$$

respectively, where $\bar{y}_j^*$ is the j^{th} simulation replicate of $\bar{y}^*$, computed from the j^{th} replicates of s_v and s_r. From these quantities, we computed the Monte Carlo absolute relative bias and Monte Carlo relative MSE defined as

$$\mathrm{RelBias}(\bar{y}^*) = \left| \frac{\mathrm{Bias}(\bar{y}^*)}{\bar{Y}} \right|$$

and

$$\text{RelMSE}(\bar{y}^*) = \frac{\text{MSE}(\bar{y}^*)}{\text{MSE}(\bar{y}^{\text{Unw}})},$$

respectively.

6 Results

6.1 Artificial data simulation

The relative bias of estimators can be consulted in Table 3 for both scenarios. As expected, all the estimators in Scenario 1, where there is no relationship between any of the 10 variables of interest and the participation probability π, show very low bias. This is not the case of Scenario 2 for which each variable of interest is related to π. As expected, the unweighted estimator is the most biased. Both nonsmoothed estimators (PSA and TrIPW) were effective at reducing the bias of the unweighted estimator. The TrIPW estimator achieved reductions of more than half of the original bias for almost every variable. The PSA estimator reduced bias to a lesser extent. The magnitude

Table 3 Relative bias ($RelBias$) for each variable, estimator and artificial data scenario

Sc.	Obj.	Unw	PSA			TrIPW		
			NS	XGB	LASSO	NS	XGB	LASSO
1	y_1	0.0009	0.0012	0.0012	0.0003	0.0007	0.0006	0.0018
	y_2	0.0033	0.0070	0.0069	0.0025	0.0007	0.0007	0.0024
	y_3	0.0017	0.0005	0.0006	0.0010	0.0013	0.0014	0.0021
	y_4	0.0009	0.0002	0.0001	0.0007	0.0011	0.0011	0.0001
	y_5	0.0004	0.0005	0.0005	0.0005	0.0022	0.0022	0.0013
	y_6	0.0025	0.0008	0.0009	0.0025	0.0045	0.0042	0.0051
	y_7	0.0014	0.0020	0.0020	0.0015	0.0013	0.0012	0.0016
	y_8	0.0008	0.0004	0.0004	0.0006	0.0000	0.0001	0.0001
	y_9	0.0013	0.0007	0.0008	0.0017	0.0033	0.0030	0.0004
	y_{10}	0.0009	0.0016	0.0016	0.0011	0.0012	0.0012	0.0016
2	y_1	0.1259	0.0952	0.0954	0.0957	0.0704	0.0706	0.0724
	y_2	0.4240	0.2795	0.2796	0.2808	0.1798	0.1798	0.1866
	y_3	0.5970	0.4471	0.4486	0.4494	0.3312	0.3332	0.3421
	y_4	0.1024	0.0699	0.0702	0.0710	0.0468	0.0470	0.0512
	y_5	0.1824	0.1212	0.1213	0.1219	0.0777	0.0777	0.0810
	y_6	0.1946	0.1259	0.1267	0.1283	0.0820	0.0827	0.0910
	y_7	0.1676	0.1117	0.1124	0.1137	0.0706	0.0709	0.0781
	y_8	0.1825	0.1211	0.1211	0.1217	0.0783	0.0783	0.0816
	y_9	0.2350	0.1607	0.1618	0.1634	0.1086	0.1090	0.1177
	y_{10}	0.1824	0.1218	0.1219	0.1225	0.0778	0.0778	0.0811

of bias remains moderate, except for variables y_2 and y_3. Given participation is not informative and the logistic model is correctly specified, albeit with the inclusion of too many covariates, the bias of the PSA estimator is most likely explained by the presence of very small participation probabilities so that a non-negligible proportion of individuals never get selected in any of the 500 simulation replicates (around 47%). Monte Carlo bias occurs if the population mean of those who are never selected is different from the overall population mean. This bias would be expected to decrease if the number of simulation replicates could be significantly increased so that a smaller proportion of individuals never get selected.

In both scenarios, the application of weight smoothing did not produce significant changes in bias. Weight smoothing is intended to reduce variance, not bias. It is thus not surprising to observe that it did not reduce bias, but it is reassuring to see that it did not significantly increase it either.

The relative MSE or efficiency of estimators for each scenario can be seen in Table 4. Values below 1 indicate that the estimator performed better than the unweighted estimator. Scenario 1 is favourable to weight smoothing, and the unweighted estimator is the most efficient since it corresponds to the most extreme form of smoothing. As expected, the nonsmoothed estimators (PSA and TrIPW) were both less efficient than the unweighted estimator due to the variability of weights. The TrIPW estimator was less efficient than the PSA estimator with an MSE around twice that of the unweighted estimator. On the one hand, smoothing using LASSO was very effective at improving efficiency. The LASSO smoothed estimators were almost as efficient as the unweighted estimator with a relative MSE close to 1. On the other hand, smoothing using XGBoost produced only marginal efficiency improvements. It appears that variable selection is useful for variance reduction as pointed out in Sect. 3 for the linear model (7). The following hybrid approach might provide better results than LASSO or XGBoost alone: first, select predictors from the variables of interest using LASSO and then smooth using XGBoost and the selected predictors in the first step.

In Scenario 2, the unweighted estimator is the least efficient due to its bias. Both nonsmoothed estimators improve efficiency by reducing bias, the MSE reduction being more pronounced for the TrIPW estimator. Scenario 2 is unfavourable to weight smoothing as all variables of interest are related to the participation probability. The smoothed weights $\hat{w}_i$ are thus expected to be in the neighbourhood of the original propensity weights w_i. As a result, none of the two smoothing methods produced any significant change in MSE, neither positive nor negative.

Overall, considering both scenarios, the TrIPW estimator combined with LASSO smoothing seems to offer the best compromise in terms of both bias and variance. In practice, a scenario in between these two extreme ones could be expected, where some variables of interest would be related to the propensity weight w_i and others would not. In that case, propensity weighting would contribute to bias reduction for variables of interest related to the propensity weight and LASSO smoothing would reduce variance for other variables provided the predictors are not too strongly related to the propensity weight.

Table 4 Relative MSE ($RelMSE$) for each variable, estimator and artificial data scenario

Sc.	Obj.	PSA			TrIPW		
		NS	XGB	LASSO	NS	XGB	LASSO
1	y_1	1.2063	1.1918	0.9906	2.0446	2.0346	1.1913
	y_2	1.3785	1.3806	1.0173	2.1675	2.1673	1.1885
	y_3	1.3097	1.2909	1.0714	2.1058	2.0749	1.1705
	y_4	1.1612	1.1516	1.0255	1.8567	1.8366	1.0877
	y_5	1.1731	1.1701	1.0287	1.8120	1.8107	1.0946
	y_6	1.2651	1.2483	1.0670	1.7955	1.7812	1.1350
	y_7	1.1379	1.1219	1.0358	2.0513	2.0291	1.1812
	y_8	1.1272	1.1261	1.0124	1.9452	1.9491	1.0734
	y_9	1.1819	1.1701	1.0329	1.7407	1.7265	1.0724
	y_{10}	1.1882	1.1901	1.0175	1.9817	1.9799	1.0969
2	y_1	0.5802	0.5822	0.5856	0.3407	0.3417	0.3586
	y_2	0.4377	0.4380	0.4417	0.1895	0.1895	0.2034
	y_3	0.5667	0.5703	0.5722	0.3280	0.3318	0.3488
	y_4	0.5039	0.5080	0.5178	0.3047	0.3060	0.3405
	y_5	0.4455	0.4457	0.4501	0.1922	0.1922	0.2080
	y_6	0.4713	0.4762	0.4865	0.3067	0.3089	0.3413
	y_7	0.4853	0.4903	0.4994	0.2839	0.2853	0.3177
	y_8	0.4440	0.4443	0.4486	0.1935	0.1935	0.2094
	y_9	0.5554	0.5602	0.5688	0.4516	0.4509	0.4673
	y_{10}	0.4494	0.4495	0.4540	0.1945	0.1945	0.2105

6.2 Real data simulation

The relative bias of estimators for the two different set of covariates can be observed in Table 5 for the real data simulation when SRSWOR is used to draw s_v from the subpopulation having a computer at home. The unweighted estimator shows a small-to-moderate bias in all cases, except for variable y_5 where the relative bias is slightly above 20%. The nonsmoothed estimators show relative biases similar to those of the unweighted estimators, albeit slightly reduced. This indicates that covariates are weakly associated with both the variables of interest and having a computer at home. As expected, the smoothed estimators did not reduce further the bias but did not increase it either.

The relative bias of estimators for the two sets of covariates when unequal probability sampling is used to select s_v from the subpopulation having a computer at home can be seen in Table 6. The unweighted estimator shows a small-to-moderate bias, except for variables y_3, y_5 and y_6. Again, the nonsmoothed estimators were ineffective at reducing bias in general. Indeed, for variable y_5, the relative bias of the nonsmoothed estimators was significantly larger than the unweighted estimator. As expected, the smoothed estimators did not reduce the bias of the nonsmoothed estimator but did not increase it either.

Table 5 Relative bias ($RelBias$) for each variable, estimator and set of covariates when using SRSWOR to draw s_v from the subpopulation having a computer at home

Cov.	Obj.	Unw	PSA			TrIPW		
			NS	XGB	LASSO	NS	XGB	LASSO
Dem.	y_1	0.012	0.016	0.016	0.015	0.022	0.021	0.022
	y_2	0.118	0.113	0.113	0.114	0.111	0.111	0.111
	y_3	0.085	0.072	0.072	0.073	0.071	0.071	0.071
	y_4	0.036	0.031	0.031	0.032	0.032	0.031	0.032
	y_5	0.209	0.188	0.188	0.193	0.174	0.185	0.174
	y_6	0.105	0.092	0.092	0.094	0.090	0.094	0.090
	y_7	0.051	0.038	0.039	0.044	0.034	0.042	0.035
	y_8	0.084	0.084	0.084	0.084	0.088	0.085	0.088
	y_9	0.002	0.002	0.002	0.002	0.000	0.000	0.000
	y_{10}	0.001	0.001	0.001	0.001	0.000	0.000	0.000
Dep.	y_1	0.012	0.004	0.004	0.004	0.002	0.003	0.002
	y_2	0.119	0.108	0.108	0.109	0.096	0.098	0.096
	y_3	0.085	0.080	0.080	0.080	0.075	0.075	0.075
	y_4	0.036	0.034	0.034	0.034	0.033	0.032	0.033
	y_5	0.211	0.224	0.224	0.220	0.228	0.232	0.228
	y_6	0.106	0.080	0.080	0.081	0.066	0.071	0.066
	y_7	0.058	0.071	0.071	0.070	0.080	0.083	0.080
	y_8	0.082	0.047	0.047	0.049	0.028	0.030	0.028
	y_9	0.000	0.001	0.001	0.000	0.001	0.001	0.001
	y_{10}	0.001	0.000	0.000	0.000	0.002	0.002	0.002

The relative MSE of estimators for the two sets of covariates under SRSWOR can be seen in Table 7. The nonsmoothed estimators are moderately more efficient than the unweighted estimator in general, which may partly be explained by their slightly smaller biases. However, for the artificial variables y_9 and y_{10}, the relative MSE of the nonsmoothed estimators is larger than 1, albeit only marginally. This indicates that the propensity weights w_i do not exhibit a large variability. As a result, weight smoothing cannot achieve large variance reductions.

The relative MSE of estimators for the two sets of covariates under unequal probability sampling can be seen in Table 8. The nonsmoothed estimators show mitigated results, sometimes being more efficient than the unweighted estimator (typically when propensity weighting reduces bias) and sometimes not. For variable y_5, the inefficiency of the nonsmoothed estimators is due to their increased bias as noted above. Weight smoothing did not achieve large efficiency gains, except for a few cases where the nonsmoothed estimator was less efficient than the unweighted estimator. This limited efficiency improvement might be explained by a somewhat strong relationship between the variables of interest and the covariates, or a small variability of the propensity weights w_i, so that the smoothed weights may not exhibit substantial deviations from w_i.

Table 6 Relative bias ($RelBias$) for each variable, estimator and set of covariates, when using unequal probability sampling to draw s_v from the subpopulation having a computer at home

Cov.	Obj.	Unw	PSA			TrIPW		
			NS	XGB	LASSO	NS	XGB	LASSO
Dem.	y_1	0.018	0.014	0.014	0.016	0.010	0.012	0.010
	y_2	0.121	0.128	0.128	0.125	0.129	0.123	0.129
	y_3	0.243	0.195	0.196	0.200	0.172	0.175	0.172
	y_4	0.080	0.074	0.074	0.074	0.071	0.070	0.071
	y_5	0.282	0.560	0.554	0.499	0.597	0.594	0.588
	y_6	0.177	0.203	0.203	0.194	0.249	0.256	0.249
	y_7	0.093	0.077	0.079	0.090	0.091	0.110	0.092
	y_8	0.038	0.048	0.048	0.050	0.062	0.059	0.063
	y_9	0.000	0.001	0.001	0.000	0.004	0.004	0.004
	y_{10}	0.001	0.002	0.002	0.001	0.000	0.000	0.001
Dep.	y_1	0.019	0.010	0.010	0.011	0.002	0.004	0.003
	y_2	0.121	0.115	0.115	0.116	0.106	0.107	0.106
	y_3	0.243	0.241	0.241	0.242	0.240	0.240	0.240
	y_4	0.080	0.080	0.080	0.080	0.080	0.079	0.080
	y_5	0.286	0.327	0.326	0.315	0.365	0.365	0.364
	y_6	0.177	0.145	0.145	0.147	0.130	0.135	0.130
	y_7	0.092	0.092	0.092	0.092	0.110	0.111	0.110
	y_8	0.038	0.018	0.018	0.020	0.005	0.005	0.005
	y_9	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	y_{10}	0.002	0.002	0.002	0.002	0.001	0.001	0.001

7 Discussion

Weight smoothing was introduced by Beaumont (2008) for reducing the variance of estimates from probability samples. It consists of modelling the survey weight conditional on the variables of interest and then replacing the weight with its predicted value. This paper extends this idea to nonprobability samples, where the weight is itself estimated from a propensity model. Our assumption was that it could improve efficiency when there are covariates in the propensity model that are weakly associated with the variables of interest.

First, we have shown theoretically that the smoothed estimator is never less efficient that its nonsmoothed version under a linear model for the propensity weight w_i. The magnitude of the efficiency gains depends on the strength of the variables of interest for predicting the propensity weight. It also depends on how powerful the predictors of the propensity weight are for predicting each of the variables of interest. Then, we designed two simulation studies, based on artificial and real data, to evaluate the properties of weight smoothing. The results showed that weight smoothing may contribute to reduce the MSE, particularly when nonsmoothed estimators (obtained using weighted logistic regression or weighted CART) are less efficient than the simple unweighted estimator. For instance, this would occur when the variables of interest are weakly related to the

Table 7 Relative MSE ($RelMSE$) for each variable, estimator and set of covariates when using SRSWOR to draw s_v from the subpopulation having a computer at home

Cov.	Obj.	PSA			TrIPW		
		NS	XGB	LASSO	NS	XGB	LASSO
Dem.	y_1	1.397	1.392	1.277	2.271	2.123	2.239
	y_2	0.917	0.918	0.930	0.870	0.874	0.873
	y_3	0.722	0.723	0.750	0.704	0.713	0.705
	y_4	0.740	0.741	0.769	0.755	0.748	0.756
	y_5	0.866	0.865	0.889	0.847	0.894	0.846
	y_6	0.763	0.763	0.797	0.727	0.778	0.728
	y_7	0.861	0.867	0.907	0.910	0.952	0.909
	y_8	0.997	0.996	0.996	1.122	1.067	1.122
	y_9	1.034	1.033	1.008	1.123	1.091	1.117
	y_{10}	1.051	1.050	1.020	1.106	1.100	1.107
Dep.	y_1	0.459	0.460	0.477	0.692	0.668	0.688
	y_2	0.829	0.831	0.841	0.667	0.688	0.671
	y_3	0.885	0.885	0.900	0.797	0.797	0.798
	y_4	0.899	0.898	0.904	0.817	0.794	0.816
	y_5	1.096	1.095	1.060	1.229	1.251	1.229
	y_6	0.572	0.573	0.596	0.417	0.473	0.418
	y_7	1.228	1.228	1.207	1.573	1.570	1.569
	y_8	0.409	0.411	0.437	0.243	0.251	0.245
	y_9	1.020	1.017	1.015	1.122	1.072	1.115
	y_{10}	1.000	0.999	0.991	1.165	1.143	1.166

covariates used in the propensity model. When the nonsmoothed estimators reduced bias and were more efficient than the unweighted estimator, weight smoothing did not yield significant efficiency gains in our simulation scenarios although it would remain theoretically possible.

In the real data simulation, there were some remarkable exceptions to the behaviour described above. In a few cases, the nonsmoothed estimators were largely inefficient, but weight smoothing could not improve the results. In those cases, propensity weighting contributed to increasing the bias, rather than reducing it. This may occur when the propensity model is misspecified (Lee 2006; Ferri-García and Rueda 2020). Therefore, the resulting augmentation of the MSE was due to an increase in bias, not variance. This explains why weight smoothing could not improve efficiency in those cases as it is not designed to reduce bias.

Regarding weight smoothing methods, LASSO regression presented better results overall than XGBoost in terms of MSE reduction. LASSO regression involves variable selection, which can be particularly relevant when some variables of interest are weakly related to the propensity weight. As shown theoretically for a linear model, no efficiency gain can be achieved for the estimation of the population mean of a variable when this variable is included in the smoothing model. Therefore, a hybrid method that would first select important variables using LASSO and then apply XGBoost to

Table 8 Relative MSE ($RelMSE$) for each variable, estimator and set of covariates, when using unequal probability sampling to draw s_v from the subpopulation having a computer at home

Cov.	Obj.	PSA			TrIPW		
		NS	XGB	LASSO	NS	XGB	LASSO
Dem.	y_1	0.797	0.797	0.930	0.817	0.796	0.817
	y_2	1.124	1.117	1.073	1.144	1.049	1.138
	y_3	0.650	0.654	0.679	0.506	0.529	0.508
	y_4	0.852	0.853	0.848	0.802	0.779	0.801
	y_5	3.488	3.414	2.847	4.697	4.597	4.571
	y_6	1.318	1.320	1.206	1.955	2.065	1.957
	y_7	0.983	0.996	1.031	1.646	1.721	1.655
	y_8	1.522	1.547	1.556	2.701	2.408	2.766
	y_9	1.464	1.444	1.172	2.261	1.915	2.244
	y_{10}	1.454	1.456	1.190	1.990	1.830	1.990
Dep.	y_1	0.468	0.473	0.504	0.386	0.383	0.386
	y_2	0.907	0.909	0.920	0.760	0.777	0.764
	y_3	0.990	0.990	0.994	0.977	0.976	0.977
	y_4	0.996	0.996	0.996	0.992	0.982	0.992
	y_5	1.265	1.261	1.188	1.588	1.581	1.580
	y_6	0.673	0.674	0.696	0.544	0.589	0.544
	y_7	1.001	1.004	0.999	1.322	1.311	1.324
	y_8	0.529	0.531	0.575	0.467	0.446	0.465
	y_9	1.002	1.000	1.003	1.230	1.160	1.227
	y_{10}	1.040	1.039	1.021	1.200	1.173	1.199

predict the propensity weight using the variables selected in the first step might be more effective than LASSO or XGBoost alone. This could be investigated in future research.

The nonsmoothed TrIPW estimator (weighted CART) appeared more effective at reducing bias than the nonsmoothed PSA estimator (weighted logistic regression) in a majority of cases. However, the TrIPW estimator was sometimes less efficient than the PSA estimator. In those cases, weight smoothing was effective at reducing the MSE to a level similar to the MSE of the PSA estimator. This suggests that a reasonable weighting strategy to use as a default choice when adjusting for selection bias in nonprobability surveys could be to use weighted CART to obtain the propensity weight w_i followed by weight smoothing.

The present empirical study has some limitations that should be noted. First, we did not consider creating homogeneous propensity strata after logistic regression. This is quite common in the context of survey nonresponse and has the advantage of reducing the occurrence of extreme propensity weights as well as providing some robustness to model misspecifications. Second, only two prediction algorithms were proposed for weight smoothing. There is currently a wide range of algorithms in the machine learning literature. Further studies could explore other prediction algorithms for weight smoothing or consider other strategies for hyperparameter tuning, which could lead to better results. Finally, the data used for our simulations cover a limited range of situations; for instance, the artificial data simulation only considered a U-shaped distribution for the participation probabilities, and the real data simulation presented a situation where the selection bias was not extremely large. Other more realistic scenarios should be considered in future research on this topic. In addition, the number of variables of interest in the simulations was fixed at 10. Further research could consider scenarios where the number of variables of interest is significantly larger, as it is likely to have an impact on the properties of weight smoothing estimators.

Funding This study was partially supported by Ministerio de Ciencia e Innovación, Spain [grant number PID2019-106861RB-I00 / AEI / 10.13039/501100011033], the IMAG-Maria de Maeztu grant [grant number CEX2020-001105-M / AEI / 10.13039/501100011033] and, in terms of the first author, a FPU grant from the Ministerio de Ciencia, Innovación y Universidades, Spain. Funding for open access charge: Universidad de Granada / CBUA. Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Availability of data and material Datasets used for the study are available upon request from the authors.

Code availability Code used for the simulations is available upon request from the authors.

A Proof of result (6)

Under regularity conditions given in Chen et al. (2020),

$$E\left(\bar{y}_w \mid \mathbf{X}, \mathbf{Y}\right) = \bar{Y} + o\left(n_v^{-1/2}\right).$$

As a result,

$$E\left(\bar{y}_w \mid \mathbf{Y}\right) = E\left[E\left(\bar{y}_w \mid \mathbf{X}, \mathbf{Y}\right) \mid \mathbf{Y}\right] = \bar{Y} + o\left(n_v^{-1/2}\right). \tag{14}$$

In addition, we also have that

$$E\left(\bar{y}_w \mid \mathbf{Y}\right) = E\left[E\left(\bar{y}_w \mid s_v, \mathbf{Y}\right) \mid \mathbf{Y}\right] = E\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right). \tag{15}$$

Combining (14) and (15), we obtain the first part of the result:

$$E\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) = \bar{Y} + o\left(n_v^{-1/2}\right).$$

To obtain the second part of the result, it suffices to observe that

$$var\left(\bar{y}_w \mid \mathbf{Y}\right) = var\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) + E\left[var\left(\bar{y}_w \mid s_v, \mathbf{Y}\right) \mid \mathbf{Y}\right]. \tag{16}$$

The result $var\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) \leq var\left(\bar{y}_w \mid \mathbf{Y}\right)$ is proven by noting that the second term on the right-hand side of (16) cannot be negative.

B Proof of result (10)

Under the linear model (7), and from result (6) and Eq. (15), we have

$$E\left(\bar{y}_w \mid \mathbf{Y}\right) = E\left(\bar{y}_{\tilde{w}} \mid \mathbf{Y}\right) = E\left[N^{-1}\sum_{i\in s_v}\left(\mathbf{h}_i^{\top}\boldsymbol{\gamma}\right)y_i \mid \mathbf{Y}\right] = \bar{Y} + o\left(n_v^{-1/2}\right).$$

Under the linear model (7), we also have that

$$E\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = E\left[E\left(\bar{y}_{\hat{w}} \mid s_v, \mathbf{Y}\right) \mid \mathbf{Y}\right] = E\left[N^{-1}\sum_{i\in s_v}\left(\mathbf{h}_i^{\top}\boldsymbol{\gamma}\right)y_i \mid \mathbf{Y}\right].$$

Combining the last two equations, we obtain: $E\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = \bar{Y} + o\left(n_v^{-1/2}\right)$.

Assuming w_i given s_v and $\mathbf{Y}$ are mutually independent (at least asymptotically), it is also straightforward to show that

$$var\left(\bar{y}_w \mid \mathbf{Y}\right) = var\left[N^{-1}\sum_{i\in s_v}\left(\mathbf{h}_i^{\top}\boldsymbol{\gamma}\right)y_i \mid \mathbf{Y}\right] + \frac{\sigma^2}{N^2}E\left(\sum_{i\in s_v} y_i^2 \mid \mathbf{Y}\right). \tag{17}$$

and that

$$var\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = var\left[N^{-1}\sum_{i\in s_v}\left(\mathbf{h}_i^{\top}\boldsymbol{\gamma}\right)y_i \mid \mathbf{Y}\right] + \frac{\sigma^2}{N^2}E\left(\sum_{i\in s_v} y_i\mathbf{h}_i^{\top}\hat{\alpha} \mid \mathbf{Y}\right), \tag{18}$$

where $\hat{\alpha}$ is given in (9). Combining (17) and (18), noting that $\hat{y}_i = \mathbf{h}_i^{\top}\hat{\alpha}$ and rearranging the terms yield:

$$var\left(\bar{y}_{\hat{w}} \mid \mathbf{Y}\right) = var\left(\bar{y}_w \mid \mathbf{Y}\right) - \frac{\sigma^2}{N^2}E\left[\sum_{i\in s_v}\left(y_i - \hat{y}_i\right)^2 \mid \mathbf{Y}\right].$$

References

Beaumont JF (2008) A new approach to weighting and inference in sample surveys. Biometrika 95(3):539–553

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

Bosnjak M, Tuten TL (2003) Prepaid and promised incentives in web surveys: an experiment. Soc Sci Comput Rev 21(2):208–217

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Florida

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006) Variable selection for propensity score models. Am J Epidemiol 163(12):1149–1156

Castro-Martín L, Rueda M, Ferri-García R (2020) Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. Mathematics 8(6):879. https://doi.org/10.3390/math8060879

Chen T, Guestrin C (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016. San Francisco, United States: Association for Computing Machinery (pp. 785-794)

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

Chu K, Beaumont J-F (2019) The Use Of Classification Trees To Reduce Selection Bias For A Non-Probability Sample With Help From A Probability Sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting. Statistical Society of Canada, Calgary, Canada Available at: https://ssc.ca/sites/default/files/imce/survey_methods_4_-_the_use_of_classification_trees_to_reduce_selection_bias_for_a_non-probability_sample_with_help_from_a_probability_sample_chu_beaucmont-2019.pdf (accessed December 2020)

Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24(2):295–313

Díaz de Rada V (2012) Ventajas e inconvenientes de la encuesta por Internet. Papers: revista de sociologia 97(1):193–223

Elliott MR, Valliant R (2017) Inference for nonprobability samples. Stat Sci 32(2):249–264

Ferri-García R, Rueda M (2018) Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat Oper Res T 42(2):159–182

Ferri-García R, Rueda M (2020) Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PloS One 15(4):e0231500

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Greenlaw C, Brown-Welty S (2009) A comparison of web-based and paper-based survey methods: testing assumptions of survey mode and response cost. Evaluation Rev 33(5):464–480

Hirano K, Imbens GW (2001) Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcomes Res Method 2(3–4):259–278

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47(260):663–685

Kohut A, Keeter S, Doherty C, Dimock M, Christian L (2012) Assessing the representativeness of public opinion surveys. Pew Research Center, Washington, DC

Kuhn M (2018) caret: Classification and Regression Training. R package version 6.0-81. Accessed December 2020. https://CRAN.R-project.org/package=caret

Lee S (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J Off Stat 22(2):329–349

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

Little RJ (1986) Survey nonresponse adjustments for estimates of means. Int Stat Rev 54(2):139–157

National Institute of Statistics (2012) Life Conditions Survey. Microdata. Accessed December 2020. https://ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=resultados&idp=1254735976608#!tabs-1254736195153

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

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

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

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

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