SORT 48 (1) January-June 2024, 1-32 DOI: 10.57645/20.8080.02.15

Kernel Weighting for blending probability and non-probability survey samples

María del Mar Rueda[1,*], Beatriz Cobo[2], Jorge Luis Rueda-Sánchez[1], Ramón Ferri-García[1] and Luis Castro-Martín[2]

Abstract

In this paper we review some methods proposed in the literature for combining a non-probability and a probability sample with the purpose of obtaining an estimator with a smaller bias and standard error than the estimators that can be obtained using only the probability sample. We propose a new methodology based on the kernel weighting method. We discuss the properties of the new estimator when there is only selection bias and when there are both coverage and selection biases. We perform an extensive simulation study to better understand the behaviour of the proposed estimator.

MSC: *62D05.*

Keywords: *Kernel weighting, survey sampling, non-probability sample, coverage bias, selection bias.*

1. Introduction

Probability sampling methods are well established by statistical offices and researchers as one of the primary tools for data collection in surveys. This is because when controlling the sampling design, it is feasible to make valid statistical inference about large finite populations using relative small samples. There exists an extensive literature on methods for probability sampling and design-based inferences for complex surveys.

* *Corresponding author:* mrueda@ugr.es

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

[2] Department of Quantitative Methods for Economics and Business, University of Granada, Spain.
Received: November 2022
Accepted: September 2023

However, the deployment of probability sampling methods has become more challenging, as there has been a notorious decline in response rates (Marken, 2018; Kennedy and Hartig, 2019) with the subsequent increase of the survey costs. In addition, new data sources which have arisen in recent years could be considered as alternatives to survey data. Examples are large volume datasets coming from sources such as passive data or "data lakes", and web surveys that have the potential of providing more timely estimates, as well as offering easier data access and lower data collection costs than traditional probability sampling, leading to larger sample sizes. On the other hand, there are serious issues concerning the use of non-probability survey samples (or volunteer samples) for estimation. Non-probability surveys are those where the inclusion probability is not known and/or not strictly positive for any individual in the population, which is the case for volunteer samples obtained from the Internet or similar means. For this reason, non-probability surveys are often known as voluntary surveys. The primary issue with these data sources is that the selection mechanism, which decides what individuals are eventually included in the dataset, is often unknown and may induce serious coverage and selection biases. Coverage bias can be defined as the bias that arise from the lack of exhaustiveness of the sampling frame from which the sample is drawn, this is, the inability of the sampling frame to include all the members of the target population. Selection bias is a term that comprises different types of errors when drawing the sample, but the most common in the aforementioned data sources is self-selection: the decision of being in the sample or not is taken by the individuals themselves, meaning that the inclusion probabilities are not given by the sampling design but by the participants, and therefore these probabilities remain unknown, constituting a non-probability sample. The generalization of the results under these biases is therefore compromised.

Despite these limitations, non-probability survey designs may be particularly useful in several cases. For example, they can be used in those cases where the target population is a small subpopulation unlikely to meet sample size requirements, or when we are interested in non-demographical strata which cannot be considered in a sampling design. Given the potential of non-probability surveys, statisticians have studied the integration or combination of data from probability and non-probability samples. Some reviews on methods of statistical data integration for finite population inference can be consulted in Buelens, Burger and vanden Brakel (2018), Valliant (2020), Yang and Kim (2020) or Rao (2020). The number of papers that are emerging in recent years in this field is immense. The importance that the topic is taking has motivated the holding of specialized sessions in many statistics and survey congresses as well as a special discussion paper in the Survey Methodology journal (vol 48, n.2). The paper of Wu (2022) ably and usefully summarizes the state of the literature of analysis of non-probability survey data and comments to the article by Bailey (2022), Elliott (2022), Lohr (2022), Meng (2022) and Wang and Kim (2022) deal with aspects of great interest and topicality in this subject.

Different data integration methods, which are based on combining probability and non-probability samples, have been developed in the literature on survey sampling. These integration methods can be divided into three groups depending on the availability of the study variable: available in the non-probability sample only, in the probability sample only, or in both samples.

Many methods consider the first case, where the target variable has been observed in the non-probability sample only. In this situation, the probability sample plays an important role as the reference data, and can be used to increase the efficiency of the estimates through a variety of adjustment approaches to account for the selection bias of non-probability samples. However, other methods were also developed from different perspectives according to the availability of auxiliary information. Calibration (Deville and Särndal, 1992; Ferri-García and Rueda, 2018), propensity score adjustment (PSA) (Lee, 2006; Lee and Valliant, 2009; Castro, Rueda and Ferri-García, 2022), kernel weighting (KW) (Wang et al., 2020), statistical matching (or mass imputation) (Rivers, 2007; Beaumont, 2020), double robust estimation (Kim and Wang, 2019) and superpopulation modeling (Valliant, Dorfman and Royall, 2000; Buelens et al., 2018) are relevant techniques to mitigate selection bias.

When the non-probability (or volunteer) survey contains auxiliary variables but no study variable, Medous et al. (2022) shows how the use of a non-probability database can improve estimates from a probability sample and they define a class of QR predictors (Särndal and Wright, 1984) asymptotically design-unbiased under certain conditions.

In this paper we consider the third situation posed above, where the study variables are measured in both samples. In Section 2 we review the estimation from probability and non-probability samples to introduce the notation and the framework. In Section 3 we revisit some important works in data integration for handling selection bias in our context. In Section 4 we adapt the kernel weighting method introduced in Wang et al. (2020), to data integration. First, we consider a situation where there are no coverage biases (there is a one-to-one correspondence between the target population and the sampling frames), and we propose a KW estimator by a linear combination of biased and unbiased estimators of a population mean. When undercoverage occurs in the non-probability sample (the sampling frame does not include all members of the target population), as is usual in practice, we propose a KW estimator based on dual frame methodology. We derive conditions such that these proposed estimators are asymptotically design-unbiased. In Section 5, we use Monte Carlo simulations to compare the proposed method with several models and show that the kernel weighted estimator is a good compromise for several setups. Finally we conclude and give perspectives in 6.

2. Context and notation

Let U be the target population of size N, $U = \{1, \dots, i, \dots, N\}$. Let s_v be the set of n_v units selected from the frame U_v using a non-probability (volunteer) data collection method. Let s_r be a probability sample of size n_r selected from a frame U_r under the sampling design $d = (S_r, P_r)$, where S_r is a subset of the universal sample space S and P_r is a probability function defined on S_r, with $\pi_i > 0$ the first order inclusion probability for individual i and π_{ij} the second order probabilities for individuals i and j. Let be

$d_i = 1/\pi_i$ the sampling design weight of unit i. We consider a situation in which U_r and U_v coincide with the population under study U. That is, there are no coverage biases in either the probability or the non-probability sample.

Let us denote with y_i the collected value on the unit i for the target variable y and let $\mathbf{x}_i$ be the observed values for individual i for a vector of covariates $\mathbf{x}$. Both y and $\mathbf{x}$ have been measured in both samples.

The target parameter is the population mean, $\overline{Y} = \frac{1}{N}\sum_U y_i$, that can be estimated from the probability sample using the Horvitz-Thompson estimator:

$$\bar{y}_r = \frac{1}{N}\sum_{i \in s_r} d_i y_i, \tag{1}$$

and from the volunteer sample with the naive estimator:

$$\bar{y}_v = \sum_{i \in s_v} \frac{y_i}{n_v}. \tag{2}$$

If we assume a situation in which there are no non-sampling errors (coverage errors, observation errors, non-response...) the estimator $\bar{y}_r$ is unbiased but if the sample size is small it can lead to estimates with large sampling errors.

Let us consider the variable

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \in U - s_v \end{cases}, \quad i = 1, \ldots, N. \tag{3}$$

The estimator $\bar{y}_v$ is biased (Kim and Wang, 2019) and its bias is given by

$$E_v(\bar{y}_v - \overline{Y}_N) = \frac{1}{f_v} E_v\{Cov(I_v, y)\},$$

where $E_v(.)$ denotes the expectation under the selection mechanism model of the non-probability sample and $f_v = n_v/N$. Thus the mean squared error (MSE) is given by the formula

$$MSE(\bar{y}_v) = \frac{1}{f_v^2} E_v\{Corr(I_v, y)^2\} Var(I_v) Var(y).$$

Therefore, a non-probability sampling where $E_v\{Corr(I_v, y)\} \neq 0$ induces a certain selection bias to the results.

In the next section we will consider how we can estimate the mean population by using a data integration estimator that combine information for these two independent surveys.

3. Methodology in data integration for handling selection bias

3.1. Some previous works

Starting with the work of Elliott and Haviland (2007), these authors consider the problem of combining the two samples by means of a linear combination of the biased and unbiased estimators of the population mean:

$$\bar{y}_{com} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_v.$$

The best estimator, in terms of efficiency, of this combination when the magnitude of the bias is known is given by:

$$\hat{y}_{EH} = \frac{\bar{y}_v \frac{\sigma_r}{n_r} + \bar{y}_r(B^2 + \frac{\sigma_v}{n_v})}{B^2 + \frac{\sigma_r}{n_r} + \frac{\sigma_v}{n_v}}, \quad (4)$$

being $\bar{y}_v$ and $\bar{y}_r$ the sample means, with variances $\frac{\sigma_v}{n_v}$ and $\frac{\sigma_r}{n_r}$ and B the bias of $\bar{y}_v$.
In practice, the bias and variances have to be estimated using the information available from both samples. The bias can be estimated as the difference between the sample means of both samples. In addition, the authors calculate the maximal contribution of the non-probability sample in terms of effective sample size, the role of the non-probability sample size in approaching this limit and the roles of both sample sizes in estimating bias with enough precision. They show that a large probability sample size (1000–10000) is needed for reasonably precise estimates of the remaining bias in initially bias-adjusted non-probability sample estimators.
Other important work is due to Disogra et al. (2011). Their proposal, based on calibration weighting, considers that auxiliary variables needed for calibration weighting must reliably differentiate between the probability sample and the non-probability sample. This calibration method has four steps:

1. Authors do a post-stratification raking calibration of s_r, using a set of demographic and geographical variables.

2. They combine the weighted s_r with the unweighted s_v. The combined sample is then weighted according to the probability sample's benchmarks from the previous step.

3. They compare the answers from early-adopter questions between the probability sample from step 1 to the answers from the blended sample from step 2.

4. They select some minimum number of early adopter questions to include in the raking due in Step 2.

Therefore, this procedure requires a good selection of early adopter questions that are included in the two surveys and that we believe will help to differentiate the samples.
Recently, Kim and Tam (2021) developed two estimators using combined data from probability sampling and non-probability sampling based on the total decomposition:

$$Y = Y_v + Y_c,$$

where $Y_v = \sum_{i \in s_v} y_i = \sum_{i \in U} I_{vi}\ y_i$ and $Y_c = \sum_{i \in U - s_v} y_i = \sum_{i \in U}(1 - I_{vi})y_i$. Since y is measured for all units of non-probability sampling, Y_v is known. Therefore, we only have to estimate Y_c. Authors proposed a first estimator where Y_c is estimated using the expansion estimator based on the probability sample

$$\bar{y}_{DI} = \frac{1}{N}(Y_v + \sum_{i \in s_r} d_i(1 - I_{vi})y_i).$$

In Poisson sampling, the variance of $\bar{y}_{DI}$ is smaller or equal to the variance of $\bar{y}_r$ if a condition on the study variable for simple random sampling without replacement holds. When N is known, Kim and Tam (2021) propose to improve the previous estimator using the following one:

$$\bar{y}_{PDI} = \frac{1}{N}\left(Y_v + (N - n_v)\frac{\sum_{i \in s_r} d_i(1 - I_{vi})y_i}{\sum_{i \in s_r} d_i(1 - I_{vi})}\right).$$

Authors prove that the variance of $\bar{y}_{PDI}$ is smaller than the variance of $\bar{y}_r$ for simple random sampling. They also discuss how to improve the efficiency of this data integration estimator by using ratio and calibration estimation.
Other works in this matter are briefly introduced below.

Fahimi et al. (2015) improve the blended calibration estimator provided by Disogra et al. (2011). Elliot (2009) develop pseudo-weights to create a representative sample using data from the non-probability sample under model assumptions that can be partially tested. With this approach, probability and non-probability samples can be blended, and the resulting sample can be treated as a probability sample with these new pseudo-weights. Dever (2018) proposes a hybrid estimation method based on the combined data file containing probability-based and nonprobability sample cases in a similar way as dual-frame estimation. For this hybrid estimation method , both samples cover the same portion of the population, referred to as common support. The common support assumption is a necessary first step and the authors propose sample matching as the method to evaluate this common support assumption. It is very difficult to make this assumption when we work with web surveys (or social media) and with probabilistic surveys based on population records, as the coverage differences between these samples may be very large and the method cannot be applied. On the other hand, the authors do not solve the problem of the determination of the lambda factor that glues the samples into one data file for population inferences. Wiśniowski et al. (2020) consider a Bayesian approach for integrating a small probability sample with a non-probability sample. They show that considering informative priors based on non-probability data can reduce the variance and mean squared error of the coefficients of a linear model.

Recently, Xi et al. (2022) do an extensive simultation study for comparing various weighting strategies where probability and non-probability samples are combined with weight normalization and raking adjustment. They apply these methods to a teen smoking behaviour survey. Nekrašaitė-Liegė, Čiginas and Krapavickaitė (2022) consider the case of estimating proportions when a non-probabilistic sample and scraped data are available. Some important works (Robbins, Ghosh-Dastidar and Ramchand, 2021;

Rueda et al., 2022) have appeared in which probability and non-probability samples are combined based on the propensity score adjustment technique. In the next section, we explain this technique and how it has been used by these authors.

3.2. Some estimators based on propensity score adjustment

The key concept in a non-probability survey sample is the selection mechanism. This mechanism is usually unknown and requires a suitable prediction model for the inclusion indicator variable. In this context, propensity scores, π_{vi}, can be defined as the probability of the i-th individual of being included in the sample, $P(I_{vi}=1)$, given the characteristics of the unit.

Let $\mathbf{x}$ be a vector of covariates measured in s_v and also in s_r. We make the following assumption:

Assumption 1 (strong ignorability condition): the indicator variable I_v and the study variable y are conditionally independent given x; i.e. $P(I_v=1|\mathbf{x},y)=P(I_v=1|\mathbf{x})$.

We assume that the selection mechanism of s_v verifies Assumption 1 and follows the model:

$$\pi_{vi}=P(I_{vi}=1|\mathbf{x}_i)=p_i(\mathbf{x})=m(\gamma,\mathbf{x}_i),\ \ i=1,...,N, \tag{5}$$

where $m(\cdot)$ is a given function with second continuous derivatives with respect to γ.

We aim to estimate propensity scores using data from pooling both samples. The maximum likelihood estimator of π_{vi} is $m(\hat{\gamma},\mathbf{x}_i)$ where $\hat{\gamma}$ maximizes the pseudo-likelihood (Chen, Li and Wu, 2020):

$$\tilde{l}(\gamma)=\sum_{s_v} log\frac{m(\gamma,\mathbf{x}_i)}{1-m(\gamma,\mathbf{x}_i)}+\sum_{s_r}\frac{1}{\pi_i}log(1-m(\gamma,\mathbf{x}_i)). \tag{6}$$

The estimated propensities $\hat{\pi}_{vi}=m(\hat{\gamma},\mathbf{x}_i)$ are thus used to readjust the propensity bias of the volunteer sample.

Based on these propensities, Robbins et al. (2021) define several estimators integrating the two samples. A first estimator is calculated weighting estimators from each sample:

$$\bar{y}_{RDR1}=\alpha_1\bar{y}_r+(1-\alpha_1)\bar{y}_v, \tag{7}$$

where $\bar{y}_v=\frac{1}{N}\sum_{s_v} y_i/q_i$ with $q_i=\frac{\pi_i*\hat{\pi}_{vi}}{1-\hat{\pi}_{vi}}$ and $\alpha_1=\frac{(\sum_{s_r}\pi_i^{-1})(\sum_{s_v}\hat{\pi}_{vi}^{-2})}{(\sum_{s_r}\pi_i^{-1})(\sum_{s_v}\hat{\pi}_{vi}^{-2})+(\sum_{s_r}\pi_i^{-2})(\sum_{s_v}\hat{\pi}_{vi}^{-1})}$.

For the second estimator, the authors calculate the values $p_i=\pi_i/(1-\hat{\pi}_{vi})$ for all individuals in the joined $s=s_v\cup s_r$ and obtain a simple Horvitz-Thompson type estimator with these new weights:

$$\bar{y}_{RDR2}=\frac{1}{N}\sum_{i\in s} y_i/p_i, \tag{8}$$

Let $\mathbf{x}$ be a set of auxiliary variables, related to y, whose population totals are known. Two calibration estimators are also proposed:

$\overline{y}_{RDR3} = \frac{1}{N}(\sum_{s_v} y_i * w_{1i} + \sum_{s_r} y_i * w_{2i})$ where w_{1i} and w_{2i} are as close as possible to $1/p_i$ fulfilling $T_x = \sum_{\in s_v} w_{1i}\mathbf{x}_i = \sum_{\in s_r} w_{2i}\mathbf{x}_i$ and the estimator:
$\overline{y}_{RDR4} = \alpha_2 \overline{y}_r + (1-\alpha_2)\overline{y}_v$ being $\alpha_2 = \frac{(\sum_r w_{i1})(\sum_v w_{i2}^2)}{(\sum_r w_{i1})(\sum_v w_{i2}^2)+(\sum_r w_{i1}^2)(\sum_v w_{i2})}$.

Rueda et al. (2022) propose the combined estimator:

$$\overline{y}_{CPSA} = \alpha_0 \overline{y}_r + (1-\alpha_0)\overline{y}_{IPW}, \tag{9}$$

being $\overline{y}_{IPW} = \frac{1}{N}\sum_{s_v} y_i/\hat{\pi}_{vi}$, and $\alpha_0 = \frac{\hat{V}_2}{\hat{V}_1+\hat{V}_2}$ where $\hat{V}_1$ and $\hat{V}_2$ are estimators of the variance of $\overline{y}_r$ and the MSE of $\overline{y}_{IPW}$ respectively. They also propose alternative methods that combine propensity score adjustment and calibration using machine learning predictive algorithms.

Burakauskaitė and Čiginas (2022) consider a few ways on non-probability integration by combining generalized difference estimator and post-stratified calibration estimator with the inverse probability weighted estimating for estimating proportions in the survey on population by religion, native language and ethnicity in Lithuania.

The above methods can reduce bias by using propensity scores to estimate participation rates of non-probability sample units. However, they are sensitive to propensity model misspecifications and can largely increase the variance of the estimators due to extreme weights. A possible way to reduce the effect of extreme weights is the kernel weighting (KW) method Wang et al. (2020) that uses propensity scores as a measure of similarity, and therefore is less sensitive to model misspecification while avoiding the extreme weights that may be produced in propensity score estimation. In the next section we introduce the KW approach to create pseudo-weights for the non-probability sample and propose a new method of integration based on this KW estimator.

4. Estimators based on kernel weighting

The KW method was developed by Wang et al. (2020), and is a method similar to the PSA since both consist of creating pseudo-weights for the non-probability sample using auxiliary variables of a reference probability sample. However, what differentiates them is the way in which these new weights are generated, although as in PSA we will use the estimated propensities to participate in the survey. As it occurred in that case, these propensities can be estimated in different ways, even though the most commonly used one is by means of logistic regression models which may entail several disadvantages for large populations in comparison to modern prediction methods such as machine learning (ML) algorithms. The ML methodology does not require strong parametric model assumptions and therefore is robust to model misspecification. Recently, ML algorithms have been considered in the literature for the treatment of nonprobability samples (see e.g. Ferri-García and Rueda (2020), Buelens et al. (2018), Kern, Li and Wang (2021), Chu and Beaumont (2019), Castro et al. (2021). Their findings showed that ML methods have the potential to remove selection bias in nonprobability samples to a greater extent than logistic regression in some scenarios.

The KW is based on using these propensities to measure the similarity between individuals based on the distributions of the auxiliary variables of the reference sample s_r and the non-probability sample s_v. These similarities will be used as weights for our estimator, after smoothing the distances using kernel functions.

The estimated propensity score for $k \in s_v \cup s_r$ is obtained as

$$\hat{\pi}_k = E_M[\hat{I}_{vk} = 1|\mathbf{x}_k],$$

where M will be one of the mentioned machine learning models to estimate this propensity and

$$\hat{I}_{vk} = \begin{cases} 1 \;\; for \;\; k \in s_v \\ 0 \;\; for \;\; k \in s_r \end{cases}, \quad k \in s_v \cup s_r.$$

Once we have these estimated propensities, we will calculate the distance between the two individuals belonging to the different samples. We define this distance as:

$$d_{ij} = \hat{\pi}_i - \hat{\pi}_j, \;\; i \in s_v, \;\; j \in s_r.$$

This distance between individuals will have a value between -1 and 1. We seek to smooth these values, which is why we use a kernel function centered at zero. There are many alternative kernel functions that can be used (normal function, standard normal, triangular, etc.), see Servy et al. (2006). The closer this distance is to zero, the more similar the individuals are with respect to their auxiliary variables (propensities are estimated depending on the values of the auxiliary variables). Moreover, the more similar the individuals are, the greater the proportion that the KW will assign to the original weight of the reference sample d_{kj} to the i unit of the volunteer sample. This proportion is called the kernel weight, whose expression is as follows:

$$k_{ij} = \frac{K\{d_{ij}/h\}}{\sum_{i \in s_v} K\{d_{ij}/h\}}, \;\; i \in s_v, \;\; j \in s_r,$$

where $K\{\cdot\}$ is a zero-centred kernel function Epanechnikov (1969), and h is the bandwidth corresponding to that kernel function. In addition:

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

The larger the value of the kernel weight k_{ij} is, the more similar the propensities will be among individuals $i \in s_v$ and $j \in s_r$.

Once we have the kernel weights, the pseudo-weights KW can be calculated, w_i^{KW} for $i \in s_v$ which are the sum of the weights of the reference sample d_j, where $j \in s_r$, weighted by the kernel weights k_{ij} for the unit $i \in s_v$:

$$w_i^{KW} = \sum_{j \in s_r} d_j \, k_{ij}, \;\; i \in s_v, \;\; j \in s_r.$$

Therefore a KW estimator for the population mean is:

$$\bar{y}_{KW} = \frac{1}{N} \sum_{i \in s_v} w_i^{KW} y_i,$$

where $\sum_{i \in s_v} w_i^{KW} = \sum_{j \in s_r} d_j$, because of $\sum_{i \in s_v} k_{ij} = 1$.

The KW estimator is consistent if certain regularity conditions are met (see Appendix 1). Furthermore Kern et al. (2021) improve the KW method by pairing it with machine learning, in particular, they considered conditional random forests, model-based recursive partitioning, gradient tree boosting and model-based boosting for estimating the propensities and constructing pseudo-weights. Kernel smoothing is also used by Chen, Yang and Kim (2022) in the case when the study variable of interest is measured only in the non-probability sample. These authors consider mass imputation for the probability sample using the non-probability data as the training set for imputation.

Next, we proceed to present the new proposed method based on KW in two different situations: firstly, if there is no coverage bias for the sample of volunteers, and secondly, when such bias exists.

4.1. Blending the samples with kernel weighting

First, we consider the situation where there is no coverage bias (U_r and U_v are equivalent to the population under study U). In this situation we propose a class of estimators based on both samples:

$$\bar{y}_C = \alpha \bar{y}_r + (1-\alpha)\bar{y}_{KW}, \tag{10}$$

where α is a nonnegative constant such that $0 \leq \alpha \leq 1$.
We study the asymptotic properties of the proposed estimator under the framework of Isaki and Fuller (1982) in which the properties of estimators are established under a given sequence of populations and a corresponding sequence of random sampling designs.

Theorem 1. *Under assumption given in Appendix 1, the proposed estimator $\bar{y}_C \to Y$ in probability as $N \to \infty$, $n_v \to \infty$, $n_r \to \infty$ with $\frac{n_v}{N} = O(1)$ and $\frac{n_r}{N} = O(1)$.*

Proof. Assumptions 1a and 2a give sufficient conditions for the Horvitz-Thompson estimator $\bar{y}_R$ to be consistent (Isaki and Fuller, 1982). Under these conditions $\bar{y}_R \to \overline{Y}$ in probability as the finite population size $N \to \infty$.

Under assumptions 2a-2c Wang et al. (2020) (Appendix A) proves that $\bar{y}_{KW} \to \overline{Y}$ in probability as the finite population size $N \to \infty$, the survey sample size $n_v \to \infty$ and

the probability sample size $n_r \to \infty$ with $n_c/N = O(1)$. Then it is obtained that $\bar{y}_C \to \alpha\overline{Y} + (1-\alpha)\overline{Y}$ the proposed estimator converges to $\overline{Y}$.
Now, we consider the problem of how select the α parameter. A simple selection for α is to weight each estimator by the weight that sample has in the total sample so that $\alpha_n = n_r/(n_r + n_v)$.

An optimal choice of α can be calculated by minimizing the MSE of $\bar{y}_C$, which is given by

$$MSE(\bar{y}_C) = \alpha^2 V(\bar{y}_r) + (1-\alpha)^2 MSE(\bar{y}_{KW}) + 2\alpha(1-\alpha)E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y})).$$

As this equation is a quadratic equation of α, its sole extreme is found straightforwardly. The values of α that minimizes this MSE are given by

$$\alpha_{opt} = \frac{MSE(\bar{y}_{KW}) - E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))}{V(\bar{y}_r) + MSE(\bar{y}_{KW}) - 2E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))}. \tag{11}$$

The optimal α_{opt} can be used to define the optimum expression

$$\bar{y}_{Copt} = \alpha_{opt}\bar{y}_r + (1-\alpha_{opt})\bar{y}_{KW}.$$

The optimal coefficient α_{opt} depends on population parameters, which are unknown in practice, and so $\bar{y}_{Copt}$ cannot be calculated.

Though the sampling procedure of the nonprobability and the probability sample can be treated as independent, the estimator $\bar{y}_{KW}$ uses information from both non-probability and probability sample, and therefore can be correlated with $\bar{y}_r$. If we assume that the term $E((\bar{y}_r - \overline{Y})(\bar{y}_{KW} - \overline{Y}))$ is small relative to $MSE(\bar{y}_{KW})$ and $V(\bar{y}_r)$, and denoting by $\hat{V}(\bar{y}_r)$ the Horvitz-Thompson estimator of $V(\bar{y}_r)$ and $\widehat{MSE}(\bar{y}_{KW})$ an estimator for the $MSE(\bar{y}_{KW})$, we can consider the following estimator for the population mean:

$$\bar{y}_{CO} = \frac{\widehat{MSE}(\bar{y}_{KW})}{\widehat{MSE}(\bar{y}_{KW}) + \hat{V}(\bar{y}_r)}\bar{y}_r + \frac{\hat{V}(\bar{y}_r)}{\widehat{MSE}(\bar{y}_{KW}) + \hat{V}(\bar{y}_r)}\bar{y}_{KW}. \tag{12}$$

An estimator for the variance of $\bar{y}_{KW}$ can be obtained by using resampling methods Wolter (2007). By using resampling techniques, one can incorporate aspects of an estimation process into variance calculations that are not easily captured algebraically. Robbins et al. (2021) consider a delete-a-group jackknife for variance estimation when use weighting methods for blending probability and convenience samples. Rafei, Elliott and Flannagan (2022) and Chen et al. (2022) use bootstrap as the method for variance estimation when the study variable of interest is measured only in the non-probability sample. Wang et al. (2020) considered the jackknife method for calculating an estimator of the $V(\bar{y}_{KW})$. The bias of $\bar{y}_{KW}$ can be estimated by $\bar{y}_r - \bar{y}_{KW}$.

4.2. Blending the samples with coverage bias

Web and social media surveys usually have a significant under-coverage bias. Thus, we consider now a more realistic situation where there is also under-coverage bias in the non-probability sample. Chen et al. (2020) highlight the estimation problems in the scenario of having zero propensity scores for certain units in the target population. According to these authors, the severity of the problem depends on the proportion of the uncovered population units and the discrepancies between the two parts of the population in terms of the response variables. Chen (2020) also discusses issues with incomplete sampling frames where units have zero propensity scores and illustrates the danger of applying regular procedures when the sampling frame is incomplete proposing methods to adjust for under coverage bias from the nonprobability sample.

We will consider that U_r covers the entire finite population but the frame U_v be incomplete ($U_v \subset U$). The population of interest, U, may be divided into two mutually exclusive domains, $ab = U_v$ and $a = U \cap U_v^c$. Units in s_r can be divided as $s_r = s_{ra} \cup s_{rab}$, where $s_{ra} = s_r \cap a$ and $s_{rab} = s_r \cap (ab)$.

Following Hartley's idea (Hartley, 1962), we can obtain a combined estimator of $\overline{Y}$ by weighting the estimators obtained from each sample:

$$\overline{y}_H(\eta) = \frac{1}{N}(\hat{Y}_a + \eta \hat{Y}_{ab} + (1-\eta)\hat{Y}_{KW}), \tag{13}$$

where $\hat{Y}_a = \sum_{i \in s_{ra}} d_i y_i$, $\hat{Y}_{ab} = \sum_{i \in s_{rab}} d_i y_i$ and $\hat{Y}_{KW} = \sum_{i \in s_v} w_i^{KW} y_i$ and $0 < \eta < 1$.

Now, we denote as:

$$d_i^\circ = \begin{cases} d_i & \text{if} \quad i \in s_{ra} \\ \eta d_i & \text{if} \quad i \in s_{rab} \\ (1-\eta) w_i^{KW} & \text{if} \quad i \in s_v \end{cases} \tag{14}$$

then

$$\overline{y}_H(\eta) = \frac{1}{N} \sum_{i \in s} d_i^\circ y_i.$$

Theorem 2. *Under the regularity conditions given in Wang et al. (2020) for the sampling design and the propensity scores, the Hartley estimator $\overline{y}_H(\eta)$ is asymptotically unbiased for $\overline{Y}$.*

Proof. Since each domain is estimated by its Horvitz-Thompson estimator, $\hat{Y}_a + \eta \hat{Y}_{ab}$ is an unbiased estimator of $\sum_{i \in a} y_i + \eta \sum_{i \in ab} y_i$, for a given η. Under the regularity conditions given in Wang et al. (2020) the estimator $\hat{Y}_{KW}$ is asymptotically unbiased for $Y_{ab} = \sum_{i \in ab} y_i$, thus the estimator $\overline{y}_H(\eta)$ is asymptotically unbiased for $\overline{Y}$.

Though U_r and U_v are sampled independently, the estimators $\hat{Y}_a + \eta\hat{Y}_{ab}$ and $\hat{Y}_{KW}$ are not independent because, $\hat{Y}_{KW}$ uses information from the probability sample. In the same way as in the previous section, we are going to assume that this dependence is small in relation to the variances of the estimators, and we suppress the covariance term between these two estimators in the calculus of the asymptotic variance of $\overline{y}_H(\eta)$. Under this assumption, the asymptotic variance of the estimator is given by the following expression

$$\begin{aligned} V(\overline{y}_H(\eta)) &= \frac{1}{N^2}(V(\hat{Y}_a + \eta\hat{Y}_{ab}) + V((1-\eta)\hat{Y}_{KW})) \\ &= \frac{1}{N^2}(V(\hat{Y}_a) + \eta^2 V(\hat{Y}_{ab}) + (1-\eta)^2 V(\hat{Y}_{KW})), \end{aligned} \tag{15}$$

where $V(\hat{Y}_a)$ and $V(\hat{Y}_{ab})$ are computed under the sampling design $d = (s_r, p_r)$ and $V(\hat{Y}_{KW})$ under the propensity model π_v.

The choice of the value for η is an important issue. For a fixed value of η, the estimator is simple to implement and gives internal consistency given that the same set of adjusted weights is used for all variables. The value of $\eta = 0.5$ is frequently used in dual frame estimation (Mecatti, 2007). The value of η that minimizes the asymptotic variance in 15 is:

$$\eta_o = \frac{MSE(\hat{Y}_{KW}) - cov(\hat{Y}_a, \hat{Y}_{ab})}{V(\hat{Y}_{ab}) + MSE(\hat{Y}_{KW})}. \tag{16}$$

This value depends on unknown population variances and covariances. By substituting the variances and MSE for its sample based estimators we obtain an estimator that we denote by $\overline{y}_H(opt)$. We note that these modified weights are random variable and their variability needs to be accounted for in standard errors of estimators.

Note. In formula 13, the true population total N is used. It is possible to use an estimator $\hat{N}$ instead of N to construct a type-Hàjek estimator as in the paper of Chen et al. (2020). In our case we would first have to decide which estimator $\hat{N}$ to use. For example based only on the non-probability sample $\hat{N}_1 = \sum_{i \in s_v} w_i^{KW}$, only on the probability sample $\hat{N}_2 = \sum_{i\ ins_r} d_i$ or some estimator based on the two samples. This choice can influence the biasness and the efficiency of the proposed estimator, and adds one more difficulty to the problem.

5. Simulation studies

We have conducted a simulation study to compare the efficiency of some of the proposed estimators based on KW. We are interested in comparing those estimators with some alternative estimators defined in Section 3, in the effect of the machine learning algorithm used in KW, in the effect of the kernel function used in the construction of

KW pseudo-weights and also in the effect of considering coverage bias. In order to illustrate that the superiority of some estimators compared to others depends on the data, we define different setups based on different artificial populations and different sampling strategies.

5.1. Populations and setups

We consider a finite population of size N = 500000. The variables of interest were designed with the objective of having various types of relationships with the covariates and the propensities. We consider 8 auxiliary variables x, 2 variable of interest y and a variable π_{vi} which indicates the probability of being included in the non-probability sample. All of them were simulated as follows:

1. The covariates x_1, x_3, x_5 and x_7 followed a Bernoulli distribution with $p = 0.5$, and x_2, x_4, x_6 and x_8 followed normal distributions with standard deviation of one and a mean parameter of 0 or 2, depending on the value of the previous Bernoulli variable. That is to say, in order to calculate x_2 we relied on the variable x_1 and if this variable was equal to 1, then the mean would be 2, or if the variable was equal to 0, then the mean would be 0. The same procedure was followed for the rest of the variables. The propensity models were fitted using all of the 8 auxiliary variables.

2. The non-probability samples were drawn with a Poisson sampling design where the inclusion probability depends on variables x_5, x_6, x_7 y x_8 as:

$$ln\left(\frac{\pi_{vi}}{1-\pi_{vi}}\right) = -0.5 + 2.5(x_{5i} = 1) + \sqrt{2 \cdot 3.141593}\, x_{6i} x_{8i} - 2.5(x_{7i} = 1), \quad i \in U. \tag{17}$$

3. The target variables were created in order to have different relationships with the covariates and the propensities were simulated according to the formulas:

$$\begin{aligned} y_{1i} &= N(8,2) + 3(x_{5i} = 1) + 5\pi_i, \quad i \in U; \\ y_{2i} &= \begin{cases} 1 & \text{if } y_{1i} > 14.46 \\ 0 & \text{if } y_{1i} \leq 14.46 \end{cases}, \quad i \in U. \end{aligned} \tag{18}$$

   The threshold of 14.46 was chosen because it is equivalent to the median of the variable y_1.

   We considered three setups. In the first setup the probability sample was drawn by simple random sampling without replacement (SRSWOR) from the full population; in the second setup the probability sample was drawn with stratified random sampling by the auxiliar variable x_7 and considering an allocation by strata of 1/3 and 2/3; in the third setup, the probability sample was selected with Midzuno sampling where the probabilities were proportional to a variable following a normal distribution with a mean parameter dependent on the value of the auxiliar variable x_7 and a standard deviation of 0.5.

The aim of the described selection mechanism was to create weights with large variability. As a result, the mean propensity is 0.7050, with a standard deviation of 0.3792, and thus a coefficient of variation of 0.5379. The histogram of propensities $\pi_{vi}, i \in U$, is provided in figure 1.

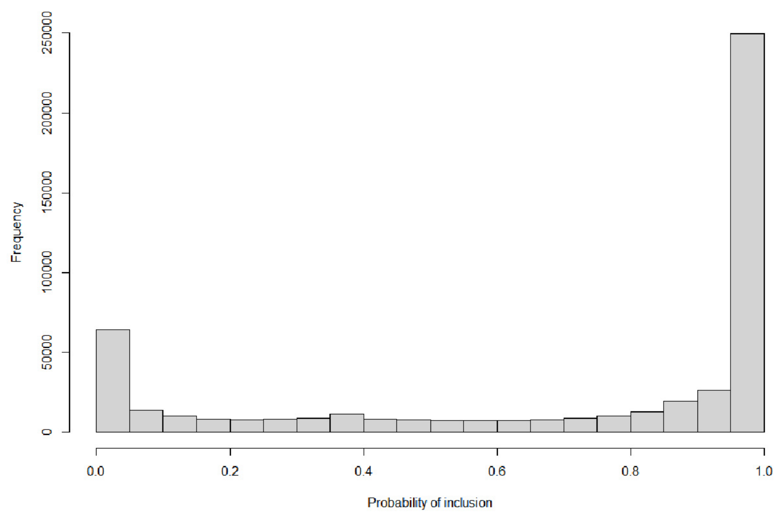

Figure 1. *Histogram of the population propensities.*

5.2. The simulation procedure

The first simulation study evaluates the performance of some estimators for $\overline{Y}$ when there is selection bias in the estimates. We focused on the proposed estimator discussed in the paper, $\overline{y}_{CO}$, and we compared it with others estimators based on propensities. As a reference estimator we have considered the naive estimator that weights the estimators simply by their sizes $\overline{y}_{REF} = \frac{n_r}{N_r}\overline{y}_r + \frac{n_v}{N_v}\overline{y}_v$. We also evaluate the estimators $\overline{y}_{RDR1}$ (7), $\overline{y}_{RDR2}$ (8) and $\overline{y}_{CPSA}$ (9) that do not use calibration.

We considered the XGBoost (Chen and Guestrin, 2016) algorithm among several machine learning approaches for estimating the propensities in all estimators. This algorithm builds decision trees ensembles that optimize an objective function via gradient tree boosting (Friedman, 2001). Literature shows that PSA with gradient boosting machines provides better results than other machine learning approaches (Lee, Lessler and Stuart, 2010, 2011; McCaffrey, Ridgeway and Morral, 2004, 2013; Ferri-García and Rueda, 2020; Rueda et al., 2022). The method depends on several hyperparameters for a proper functioning and in order to avoid overfitting. We have considered the following hyperparameters: the number of trees forming the ensemble (50, 100 or 150), the weight shrinkage applied after each boosting step (0.3 or 0.4), the maximum number of splits that each tree can contain (1, 2 or 3), the proportion of variables used in each step (0.6 or 0.8) and the proportion of data used in each step (0.5, 0.75 or 1).

For each setup we select 500 probability samples of size $n_r = 250$ and 500 non-probability samples of sizes $n_v = 500; 1000; 2000$. We compute the Monte Carlo relative bias of the estimators:

$$|RB| = \frac{1}{B}\sum_{b=1}^{B}\frac{|\overline{y}_b - \overline{Y}|}{\overline{Y}} \cdot 100, \tag{19}$$

and the Monte Carlo root mean square relative error (RMSRE):

$$RMSRE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\overline{y}_b - \overline{Y}}{\overline{Y}}\right)^2} \cdot 100. \tag{20}$$

where B is the number of iterations, and $\overline{y}_b$ is an estimate of $\overline{Y}$, by the method under study, computed for the b-th sample.

We also examine the behaviour of variance estimators. We consider the jackknife method used in Wang et al. (2020) to account for all sources of variability. The performance of a variance estimator along with the point estimator $\overline{y}_i$ is assessed by the length of the intervals obtained at 95% confidence level and their real coverage.

Variance estimators for $\overline{y}_{KW}$ is also calculated based on bootstrap methods. We have obtained similar results for RB and RMSRE for the proposed estimator $\overline{y}_{CO}$ and we observed that the behaviour with respect to the other estimators is barely influenced by the variance estimation method used. In the work only the results of the jackkniffe method are shown.

The simulation study has been carried out using the statistical software R, and for its implementation we have needed the use of specific packages of the area, such as NonProbEst (Castro, Ferri and Rueda, 2020), KWML (Kern et al., 2021), sampling (Tillé and Matei, 2021) and caret (Kuhn et al., 2022).

5.3. Results

Tables 1 and 2 contain the simulation results for y_1 and y_2 respectively for the three setups considering different sample sizes. In all setups, as expected, the proposed estimator with gradient boosting and kernel weighting ($\overline{y}_{CO}$) provides lower values of both |RB| and RMSRE. The second best estimator is $\overline{y}_{CPSA}$, which obtains results similar to the first and with the rest of the estimators we obtain higher values of the |RB| and RMSRE. It is also observed that the behaviour pattern in terms of reduction |RB| and RMSRE is similar in the three sample designs considered for the probabilistic sample.

Table 1. *Monte Carlo bias and root mean square relative error. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{REF}$	4.772	4.847	4.732	4.795	4.801	4.872
$\bar{y}_{RDR1}$	3.081	3.231	2.736	2.895	2.770	2.952
$\bar{y}_{RDR2}$	3.246	3.389	2.888	3.045	2.907	3.088
$\bar{y}_{CPSA}$	1.251	1.554	1.197	1.512	1.341	1.663
$\bar{y}_{CO}$	1.173	1.457	1.232	1.559	1.213	1.576
Stratified sampling						
$\bar{y}_{REF}$	4.800	4.880	4.864	4.942	4.730	4.788
$\bar{y}_{RDR1}$	2.998	3.190	2.913	3.114	2.752	2.910
$\bar{y}_{RDR2}$	3.651	3.788	3.601	3.746	3.435	3.546
$\bar{y}_{CPSA}$	1.448	1.798	1.595	2.003	1.326	1.665
$\bar{y}_{CO}$	1.224	1.521	1.322	1.671	1.162	1.431
Midzuno sampling						
$\bar{y}_{REF}$	4.771	4.845	4.766	4.827	4.735	4.792
$\bar{y}_{RDR1}$	3.100	3.257	2.801	2.947	2.766	2.912
$\bar{y}_{RDR2}$	3.381	3.520	3.122	3.250	3.069	3.198
$\bar{y}_{CPSA}$	1.219	1.526	1.261	1.554	1.2390	1.573
$\bar{y}_{CO}$	1.010	1.393	1.141	1.412	1.124	1.425

Table 2. *Monte Carlo bias and root mean square relative error. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{REF}$	16.229	16.653	16.199	16.547	16.396	16.775
$\bar{y}_{RDR1}$	10.025	10.769	9.130	9.809	9.188	9.971
$\bar{y}_{RDR2}$	10.650	11.367	9.600	10.281	9.605	10.398
$\bar{y}_{CPSA}$	5.188	6.406	5.136	6.400	5.759	7.176
$\bar{y}_{CO}$	4.538	5.665	4.681	5.920	5.343	6.642
Stratified sampling						
$\bar{y}_{REF}$	16.317	16.738	16.703	17.133	16.222	16.537
$\bar{y}_{RDR1}$	11.010	11.734	11.012	11.772	10.518	11.068
$\bar{y}_{RDR2}$	12.819	13.412	12.821	13.447	12.326	12.785
$\bar{y}_{CPSA}$	5.647	7.110	5.866	7.613	5.126	6.408
$\bar{y}_{CO}$	5.119	6.444	5.198	6.704	4.612	5.684
Midzuno sampling						
$\bar{y}_{REF}$	16.421	16.829	16.248	16.581	16.690	17.030
$\bar{y}_{RDR1}$	10.738	11.437	9.866	10.512	10.182	10.824
$\bar{y}_{RDR2}$	11.634	12.271	10.771	11.382	11.020	11.632
$\bar{y}_{CPSA}$	5.246	6.652	5.052	6.206	5.632	7.004
$\bar{y}_{CO}$	4.706	5.903	4.490	5.583	4.965	6.163

Table 3. *Confidence invervals' real coverage and length. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\bar{y}_{REF}$	0.000	0.481	0.000	0.453	0.000	0.438
$\bar{y}_{RDR1}$	0.108	0.517	0.168	0.492	0.164	0.475
$\bar{y}_{RDR2}$	0.088	0.522	0.146	0.505	0.146	0.490
$\bar{y}_{CPSA}$	0.956	0.853	0.962	0.854	0.918	0.855
$\bar{y}_{CO}$	0.962	0.811	0.960	0.807	0.928	0.781
	Stratified sampling					
$\bar{y}_{REF}$	0.002	0.503	0.002	0.477	0.000	0.463
$\bar{y}_{RDR1}$	0.190	0.552	0.178	0.525	0.178	0.509
$\bar{y}_{RDR2}$	0.054	0.533	0.052	0.507	0.030	0.494
$\bar{y}_{CPSA}$	0.944	0.939	0.898	0.948	0.954	0.948
$\bar{y}_{CO}$	0.958	0.844	0.906	0.822	0.952	0.788
	Midzuno sampling					
$\bar{y}_{REF}$	0.000	0.488	0.000	0.462	0.000	0.445
$\bar{y}_{RDR1}$	0.124	0.529	0.146	0.505	0.140	0.487
$\bar{y}_{RDR2}$	0.090	0.528	0.090	0.509	0.094	0.493
$\bar{y}_{CPSA}$	0.958	0.886	0.962	0.887	0.956	0.886
$\bar{y}_{CO}$	0.952	0.820	0.964	0.808	0.950	0.769

Table 4. *Confidence invervals' real coverage and length. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\bar{y}_{REF}$	0.008	0.076	0.006	0.070	0.002	0.066
$\bar{y}_{RDR1}$	0.276	0.077	0.286	0.070	0.236	0.066
$\bar{y}_{RDR2}$	0.242	0.077	0.252	0.071	0.232	0.068
$\bar{y}_{CPSA}$	0.968	0.130	0.954	0.130	0.930	0.131
$\bar{y}_{CO}$	0.950	0.118	0.954	0.119	0.904	0.116
	Stratified sampling					
$\bar{y}_{REF}$	0.018	0.080	0.002	0.074	0.002	0.071
$\bar{y}_{RDR1}$	0.198	0.079	0.174	0.072	0.132	0.069
$\bar{y}_{RDR2}$	0.108	0.078	0.084	0.071	0.052	0.068
$\bar{y}_{CPSA}$	0.944	0.139	0.924	0.140	0.976	0.140
$\bar{y}_{CO}$	0.932	0.126	0.916	0.121	0.944	0.114
	Midzuno sampling					
$\bar{y}_{REF}$	0.010	0.077	0.002	0.071	0.002	0.068
$\bar{y}_{RDR1}$	0.232	0.077	0.232	0.071	0.162	0.067
$\bar{y}_{RDR2}$	0.168	0.078	0.178	0.072	0.126	0.068
$\bar{y}_{CPSA}$	0.950	0.133	0.988	0.134	0.958	0.134
$\bar{y}_{CO}$	0.950	0.122	0.960	0.118	0.924	0.115

Table 5. *Monte Carlo bias and root mean square relative error of estimators changing the ML method. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{CO}$	1.156	1.406	1.162	1.428	1.260	1.570
$\overline{Y}_{CO-NNET}$	1.165	1.422	1.243	1.521	1.317	1.676
$\overline{Y}_{CO-K}$	1.165	1.418	1.165	1.438	1.270	1.610
$\overline{Y}_{CO-LR}$	1.197	1.468	1.279	1.568	1.339	1.695
Stratified sampling						
$\bar{y}_{CO}$	1.250	1.547	1.261	1.595	1.257	1.578
$\overline{Y}_{CO-NNET}$	1.389	1.713	1.379	1.773	1.474	1.829
$\overline{Y}_{CO-K}$	1.234	1.527	1.250	1.582	1.240	1.550
$\overline{Y}_{CO-LR}$	1.467	1.814	1.477	1.891	1.557	1.923
Midzuno sampling						
$\bar{y}_{CO}$	1.254	1.567	1.191	1.478	1.307	1.615
$\overline{Y}_{CO-NNET}$	1.331	1.665	1.277	1.605	1.490	1.884
$\overline{Y}_{CO-K}$	1.272	1.592	1.203	1.495	1.337	1.658
$\overline{Y}_{CO-LR}$	1.382	1.732	1.313	1.650	1.529	1.929

Table 6. *Monte Carlo bias and root mean square relative error of estimators changing the ML method. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
Simple random sampling without replacement						
$\bar{y}_{CO}$	5.144	6.264	4.932	6.105	5.115	6.378
$\overline{Y}_{CO-NNET}$	5.510	6.760	5.315	6.554	5.618	6.840
$\overline{Y}_{CO-K}$	5.107	6.278	5.023	6.240	5.057	6.334
$\overline{Y}_{CO-LR}$	5.739	7.011	5.558	6.903	5.842	7.101
Stratified sampling						
$\bar{y}_{CO}$	5.045	6.334	5.151	6.566	5.200	6.455
$\overline{Y}_{CO-NNET}$	5.449	6.848	5.870	7.472	6.058	7.461
$\overline{Y}_{CO-K}$	4.967	6.312	5.240	6.716	5.553	6.837
$\overline{Y}_{CO-LR}$	5.593	7.028	5.939	7.508	6.197	7.615
Midzuno sampling						
$\bar{y}_{CO}$	4.781	5.868	4.9700	6.309	5.137	6.327
$\overline{Y}_{CO-NNET}$	5.290	6.467	5.683	6.972	5.432	6.698
$\overline{Y}_{CO-K}$	4.922	6.078	5.175	6.454	5.086	6.292
$\overline{Y}_{CO-LR}$	5.520	6.717	5.807	7.124	5.592	6.916

Tables 3 and 4 show the real coverages and lengths of the corresponding 95% confidence intervals. The coverage of intervals based on estimators $\bar{y}_{REF}$, $\bar{y}_{RDR1}$ and $\bar{y}_{RDR2}$ are very low, as expected, due to the bias in the estimates. On the contrary, the proposed es-

Table 7. *Confidence invervals' real coverage and length changing the ML method. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\overline{y}_{CO}$	0.974	0.812	0.948	0.805	0.924	0.780
$\overline{Y}_{CO-NNET}$	0.970	0.830	0.956	0.839	0.918	0.858
$\overline{Y}_{CO-K}$	0.970	0.823	0.960	0.820	0.936	0.821
$\overline{Y}_{CO-LR}$	0.970	0.854	0.956	0.867	0.916	0.876
	Stratified sampling					
$\overline{y}_{CO}$	0.946	0.845	0.930	0.820	0.916	0.784
$\overline{Y}_{CO-NNET}$	0.926	0.905	0.932	0.915	0.926	0.929
$\overline{Y}_{CO-K}$	0.954	0.854	0.938	0.849	0.936	0.849
$\overline{Y}_{CO-LR}$	0.918	0.936	0.924	0.951	0.920	0.963
	Midzuno sampling					
$\overline{y}_{CO}$	0.914	0.823	0.952	0.804	0.918	0.770
$\overline{Y}_{CO-NNET}$	0.922	0.860	0.940	0.875	0.892	0.882
$\overline{Y}_{CO-K}$	0.930	0.835	0.952	0.827	0.912	0.829
$\overline{Y}_{CO-LR}$	0.918	0.893	0.950	0.901	0.908	0.911

timator $\overline{y}_{CO}$ and $\overline{y}_{CPSA}$ have good performance, having the intervals a real coverage close to the nominal coverage. With respect to the length of the intervals, as we expected, the $\overline{y}_{CO}$ estimator is the one with the shortest length for all types of sampling considered, sample sizes and type of variable. The KW is intended to reduce variance and indeed it succeeds for these scenarios and variables.

Table 8. *Confidence intervals' real coverage and length changing the ML method. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\overline{y}_{CO}$	0.940	0.118	0.938	0.120	0.926	0.119
$\overline{Y}_{CO-NNET}$	0.906	0.124	0.926	0.125	0.900	0.126
$\overline{Y}_{CO-K}$	0.954	0.120	0.944	0.120	0.938	0.120
$\overline{Y}_{CO-LR}$	0.920	0.128	0.916	0.130	0.900	0.130
	Stratified sampling					
$\overline{y}_{CO}$	0.950	0.127	0.928	0.120	0.898	0.116
$\overline{Y}_{CO-NNET}$	0.944	0.138	0.930	0.139	0.942	0.139
$\overline{Y}_{CO-K}$	0.960	0.128	0.952	0.127	0.950	0.127
$\overline{Y}_{CO-LR}$	0.938	0.140	0.920	0.140	0.946	0.143
	Midzuno sampling					
$\overline{y}_{CO}$	0.952	0.120	0.934	0.120	0.900	0.114
$\overline{Y}_{CO-NNET}$	0.964	0.130	0.914	0.130	0.958	0.133
$\overline{Y}_{CO-K}$	0.962	0.123	0.944	0.121	0.958	0.122
$\overline{Y}_{CO-LR}$	0.958	0.134	0.930	0.135	0.952	0.137

5.4. Influence of the machine learning method

In the previous simulation we used gradient boosting machine as a machine learning method, but different methods can be used. In this case we are going to make a comparison of the most used machine learning methods to see if the results are influenced by them. Specifically, we are going to compare neural networks (NNET), K-nearest neighbours (K) and logistic regression (LR) with respect to gradient boosting machine for qualitative and quantitative variables y_1 and y_2 considering the three types of sampling and for the different sample sizes. The results obtained in the comparative study can be seen in the Tables 5, 6, 7 and 8.

When comparing the $|RB|$ and the RMSRE values for y_1 for all sample sizes (Table 5), we can see that in simple random sampling and Midzuno sampling the smallest values are found for $\overline{y}_{CO}$, in the case of stratified sampling, the smallest values are found in $\overline{Y}_{CO-K}$. For y_2 (Table 6) the results obtained for the gradient boosting machine and K-nearest neighbours method are similar if we compare the $|RB|$ and the RMSRE values. When looking at the Tables 7 and 8 for y_1 it can be observed that the greatest coverage (0.91-0.97) obtained is given in the case of the gradient boosting machine and K-nearest neighbours methods. For y_2 the K-nearest neighbours method obtains the greatest coverage (0.93-0.96). With respect to the length of the confidence interval, gradient boosting machine obtains the smallest values and logistic regression model obtains the largest. The performance of the logistic regression was to be expected since the propensities do not depend on all the covariates and there is an error in the propensity model specification.

Table 9. *Monte Carlo bias and root mean square relative error of estimators changing the kernel. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	$\|RB\|$	RMSRE	$\|RB\|$	RMSRE	$\|RB\|$	RMSRE
	Simple random sampling without replacement					
$\overline{y}_{CO}$	1.160	1.448	1.144	1.437	1.261	1.578
$\hat{\overline{Y}}_{CO-SN}$	1.164	1.449	1.140	1.435	1.264	1.577
$\hat{\overline{Y}}_{CO-TSN}$	1.161	1.451	1.145	1.437	1.261	1.577
	Stratified sampling					
$\overline{y}_{CO}$	1.245	1.573	1.414	1.734	1.210	1.492
$\hat{\overline{Y}}_{CO-SN}$	1.250	1.579	1.389	1.703	1.206	1.492
$\hat{\overline{Y}}_{CO-TSN}$	1.256	1.597	1.389	1.719	1.110	1.489
	Midzuno sampling					
$\overline{y}_{CO}$	1.221	1.540	1.229	1.513	1.312	1.631
$\hat{\overline{Y}}_{CO-SN}$	1.220	1.536	1.232	1.518	1.308	1.626
$\hat{\overline{Y}}_{CO-TSN}$	1.230	1.548	1.231	1.518	1.320	1.632

Table 10. *Monte Carlo bias and root mean square relative error of estimators changing the kernel. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
	Simple random sampling without replacement					
$\overline{y}_{CO}$	4.641	5.762	4.839	6.070	5.012	6.378
$\hat{\hat{Y}}_{CO-SN}$	4.567	5.679	4.832	6.088	5.002	6.335
$\hat{\hat{Y}}_{CO-TSN}$	4.627	5.776	4.783	6.041	5.040	6.409
	Stratified sampling					
$\overline{y}_{CO}$	5.215	6.627	4.902	6.175	5.069	6.298
$\hat{\hat{Y}}_{CO-SN}$	5.199	6.612	4.991	6.250	5.064	6.332
$\hat{\hat{Y}}_{CO-TSN}$	5.271	6.631	4.988	6.230	5.099	6.377
	Midzuno sampling					
$\overline{y}_{CO}$	4.657	5.873	5.122	6.274	4.966	6.211
$\hat{\hat{Y}}_{CO-SN}$	4.736	5.896	5.202	6.311	5.014	6.263
$\hat{\hat{Y}}_{CO-TSN}$	4.617	5.870	5.220	6.375	4.993	6.256

Table 11. *Confidence intervals' real coverage and length changing the kernel. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\overline{y}_{CO}$	0.946	0.812	0.962	0.807	0.918	0.782
$\hat{\hat{Y}}_{CO-SN}$	0.956	0.814	0.966	0.811	0.920	0.790
$\hat{\hat{Y}}_{CO-TSN}$	0.950	0.812	0.968	0.810	0.918	0.789
	Stratified sampling					
$\overline{y}_{CO}$	0.946	0.843	0.912	0.828	0.948	0.785
$\hat{\hat{Y}}_{CO-SN}$	0.954	0.851	0.930	0.831	0.936	0.791
$\hat{\hat{Y}}_{CO-TSN}$	0.932	0.843	0.932	0.831	0.946	0.793
	Midzuno sampling					
$\overline{y}_{CO}$	0.930	0.821	0.958	0.807	0.912	0.777
$\hat{\hat{Y}}_{CO-SN}$	0.942	0.825	0.960	0.813	0.910	0.782
$\hat{\hat{Y}}_{CO-TSN}$	0.932	0.821	0.958	0.810	0.914	0.785

5.5. Influence of the kernel function

In the previous simulations we used the tringular distribution as kernel function in the construction of KW pseudo-weights, but different distributions can be used. In this case we are going to make a comparison of the distribution implemented in the R package Boosted Kernel Weighting (Kern et al., 2021) to see if the results are influenced by them. Specifically, we are going to compare triangular, standard normal (SN) and truncated standard normal (TSN) for qualitative and quantitative variables y_1 and y_2 considering

Table 12. *Confidence intervals' real coverage and length changing the kernel. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	Coverage	Length	Coverage	Length	Coverage	Length
	Simple random sampling without replacement					
$\bar{y}_{CO}$	0.944	0.118	0.940	0.118	0.932	0.117
$\hat{\bar{Y}}_{CO-SN}$	0.958	0.120	0.934	0.121	0.930	0.119
$\hat{\bar{Y}}_{CO-TSN}$	0.956	0.119	0.936	0.121	0.932	0.119
	Stratified sampling					
$\bar{y}_{CO}$	0.924	0.125	0.954	0.121	0.922	0.114
$\hat{\bar{Y}}_{CO-SN}$	0.926	0.126	0.944	0.122	0.926	0.117
$\hat{\bar{Y}}_{CO-TSN}$	0.938	0.126	0.944	0.122	0.916	0.117
	Midzuno sampling					
$\bar{y}_{CO}$	0.952	0.121	0.942	0.119	0.920	0.117
$\hat{\bar{Y}}_{CO-SN}$	0.950	0.123	0.942	0.122	0.918	0.117
$\hat{\bar{Y}}_{CO-TSN}$	0.948	0.122	0.936	0.120	0.916	0.117

the three types of sampling and for the different sample sizes. The results obtained in the comparative study can be seen in the Tables 9, 10, 11 and 12.

Table 13. *Monte Carlo bias and root mean square relative error of estimators with coverage bias. Variable y_1.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\bar{y}_{REF}$	5.541	5.615	5.581	5.649	5.554	5.619
$\bar{y}_{RDR1}$	3.279	3.427	3.295	3.421	3.175	3.304
$\bar{y}_{RDR2}$	3.233	3.409	3.198	3.358	2.999	3.166
$\bar{y}_{CPSA}$	1.267	1.574	1.213	1.535	1.220	1.543
$\bar{y}_{CO}$	1.258	1.563	1.209	1.529	1.204	1.520
$\bar{y}_H(opt)$	1.195	1.486	1.125	1.426	1.132	1.446

Table 14. *Monte Carlo bias and root mean square relative error of estimators with coverage bias. Variable y_2.*

	$n_r = 250, n_v = 500$		$n_r = 250, n_v = 1000$		$n_r = 250, n_v = 2000$	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\bar{y}_{REF}$	19.689	20.085	19.689	20.085	19.689	20.085
$\bar{y}_{RDR1}$	12.210	12.828	12.210	12.828	12.210	12.828
$\bar{y}_{RDR2}$	12.118	12.794	12.118	12.794	12.118	12.794
$\bar{y}_{CPSA}$	5.359	6.623	5.359	6.623	5.359	6.623
$\bar{y}_{CO}$	5.375	6.648	5.375	6.648	5.375	6.648
$\bar{y}_{H(opt)}$	5.258	6.453	5.258	6.453	5.258	6.453

The values of |RB| and the RMSRE are similar for the kernel functions used, so we can say that there is no influence of the kernel function in this study. Regarding coverage, we see that in all cases it is quite good, moving around 0.91–0.96, obtaining the shortest length of the interval in most cases in the $\bar{y}_{CO}$ estimator.

5.6. Results under coverage bias

In order to check the behaviour of the Hartley estimator $\bar{y}_H(opt)$, proposed in section 4.2, we have repeated the previous simulation but now we include a mechanism to reproduce coverage bias in our simulation. This context is compared with the same estimators considered in the first simulation.

The probability sample is selected by SRSWOR from the full population but the non-probability sample is now selected from a frame U_v created from the population U containing only individuals whose variable $x_5 = 1$ (related to target variables).

In Tables 13 and 14 values of |RB| and the RMSRE can be seen for each of the considered estimators.

As expected, all the estimators considered now have greater bias than in the previous simulation. We observe that the estimators $\bar{y}_{CPSA}$ and $\bar{y}_{CO}$ continue to be better than the other PSA-based estimators in terms of |RB| and RMSRE reduction. As expected, the estimator based on dual frames, $\bar{y}_H(opt)$, is the one that produces estimates with less |RB|, and consequently is also able to reduce the RMSRE compared to its competitors.

6. Discussion

In the last decade, survey research has witnessed the surge of non-probability sampling as a feasible alternative to probability sampling. In theory, the superiority of probability sampling should be clear, as it has a theoretical basis in design-based inference allowing for unbiased estimation of population parameters along with the calculation of exact sampling error. However, they are very expensive and usually have small sizes. Non-probability samples can offer some advantages in that sense, as they can be deployed in many relatively inexpensive ways, but they lack an underlying mathematical theory given their usual lack of design. This is troublesome with respect to achieving accuracy and representativeness for estimates derived from such samples.

Given their potential, many efforts have been undertaken in recent years to combine both probability and nonprobability samples to produce a single inference which may be able to overcome the limitations of each method, resulting in a rich literature on data integration in finite populations. Most of this literature is based on considering a framework where the variables of interest have not been observed in the probability sample. In this paper, we have considered the problem of observed study variables in both the non-probability sample and the probability sample, in presence of auxiliary information.

Since both samples contain the same variables, we propose a methodology to combine two surveys based on probability and non-probability samples with the help of machine learning algorithms, in order to obtain reliable estimations with small variance. We have introduced a general class of estimators, based on the kernel weighting method, and studied theoretically their bias properties. Using simulations we have also compared the proposed estimators with other methods for integrating probability and non-probability samples developed in the literature in different simulation setups, both in terms of $|RB|$ and RMSRE.

The simulation study indicates that $|RB|$ and RMSRE of estimators can be reduced when combining the probability and the non-probability sample using the KW method proposed here in the case where there is a relationship between the variable of interest and the participation probability. We also observed that the choice of the ML method used for propensity predictions is very important and can influence the estimates obtained. However, the kernel function in the construction of KW pseudo-weights does not influence the estimates obtained. From our simulation study we also deduce that in case the sample of volunteers has a coverage bias, it is appropriate to use an estimator based on dual frames that allows this bias to be treated as well.

These methods can be implemented using freely available statistical packages such as R. The R code used for the simulation study and the computation of the results are available on request. However, the computational cost of resampling should be mentioned. Many of the proposed methods rely on variance estimation techniques which involve resampling. For each iteration, a new model has to be trained and the calculations have to be repeated, considerably slowing down the process. Therefore, they should be avoided when execution time is of the essence and many variables are involved.

Some other papers (Elliot (2009), Dever (2018)) also combine the pseudo-weighted nonprobability and probability samples first and estimate the finite population mean from the combined sample. When pseudo-weighted samples are combined, the assigned weights only depend on the sample sizes, the design weights and the estimated propensities, which do not depend on the variable under study. Thus, the same weights are used to make estimates for all variables, but for some variables the procedure may not be able to eliminate voluntariness biases. On the contrary, the method that we propose depends on each variable under study, and takes into account the voluntariness bias that may be important for those variables that are correlated with the probability of participating in the survey of volunteers, which is the case that interests us.

In our proposal we have considered non parametric methods to estimate the underlying propensity model that reflect the self-selection process, which provides added flexibility over logistic regression-based methods. Some recent works also use nonparametric methods to make inferences for non-probability samples. Chen et al. (2022) use kernel smoothing while Yang, Kim and Hwang (2021) use nearest neighbor for mass imputation for the probability sample using the non-probability data as the training set. Our method differs from these works fundamentally in two aspects: in our case the variable under study is observed in the two samples, and we use the inverse propensity weighting methodology while they use mass imputation.

Our advice to practitioners is that the use of probability samples remains essential to obtain reliable estimates based on an accepted theory such as sampling theory (Beaumont, 2020), but complementing the probability sample with a non-probability sample can serve a means to reduce the errors in the estimates.

There is a lot of room for future research to improve estimation by mean integration: other similarity measures and other weighting adjustment methods such as weight smoothing for multipurpose surveys (Ferri-García et al., 2022) can be considered. In this work only the estimation of means and totals has been considered, but the method can be applied, with certain adjustments, to the case of other non-linear parameters such as distribution functions or quantiles. In addition, new alternative methods for estimation from a nonprobability sample continue to emerge. Liu and Valliant (2023) introduces one method of weighting that assign a unit in the nonprobability sample the weight from its matched case in the probability sample. These new methods can be used as an alternative to kernel weighting to build estimators similar to our proposal. These issues will be future research topics.

Acknowledgments

This work is part of grant PID2019-106861RB-I00 funded by MCIN/ AEI/10.13039/50 1100011033, by grant PDC2022-133293-I00 funded by MCIN/AEI/10.13039/5011000 11033 and the European Union "NextGenerationEU"/PRTR"1 and the grant FEDER C-EXP-153-UGR23 funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021-2027. The research was also partially supported from IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.

Conflict of interest

The authors declare no potential conflict of interests.

References

Bailey, M.A. (2022). Comments on "Statistical inference with non-probability survey samples" - Non-probability samples: An assessment and way forward. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00003-eng.htm.

Beaumont, J.F. (2020). Are probability surveys bound to disappear for the production of official statistics?. *Survey Methodology, Statistics Canada*, 46, 1. http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-eng.htm.

Buelens, B., Burger, J. and vanden Brakel, J.A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review* 86(2), 322-343.

Burakauskaitė, I., Čiginas, A. (2022). Non-probability sample integration in the survey of lithuanian census. *Workshop on Survey Statistics, Tartu, 2022*. http://isi-iass.org/home/wp-content/uploads/I.Burakauskaite_A.Ciginas_presentation.pdf.

Castro, L., Ferri, R. and Rueda, M.M. (2020). *N*onProbEst: Estimation in Nonprobability Sampling. https://CRAN.R-project.org/package=NonProbEst.

Castro, L., Rueda, M., Ferri-García, R. and Hernando-Tamayo, C. (2021). On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *M*athematics 9, 2991. https://doi.org/10.3390/math9232991

Castro L., Rueda M. and Ferri-García R. (2022). Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys *J*ournal of Computational and Applied Mathematics, 404, 113414.

Chen, S., Yang, S. and Kim, J.K. (2022). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*, 10(1), 1-24.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA* 785-794.

Chen, Y. (2020). Statistical Analysis with Non-Probability Survey Samples. Ph Thesis

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Chu, K.C.K. and Beaumont, J.F. (2019). The use uf 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, Calgary, AB, Canada.

Dever, J. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference 1-15.

Deville, J.C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association* 87(418), 376–382.

Disogra, C., Cobb, C.L., Chan, E.K. and Dennis, J.M. (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. *In Proceedings of the American Statistical Association, Section on Survey Research. Joint Statistical Meetings (JSM).*

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* 2(6), 2982.

Elliott, M.R. (2022). Comments on "Statistical inference with non-probability survey samples".. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00004-eng.htm.

Elliott, M. and Haviland, A. (2007). Use of a Web-Based Convenience Sample to Supplement a Probability Sample. *Survey Methodology* 33, 211-215.

Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14(1), 153-158.

Fahimi, M., Barlas, F.M., Thomas, R.K. and Buttermore, N. (2015). Scientific surveys based on incomplete sampling frames and high rates of nonresponse. *Survey Practice* 8(5), 1-11.

Ferri-García, R. and Rueda, M.M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat. Oper. Res. T., 42(2), 159–162.

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

Ferri-García, R., Beaumont, J.F., Bosa, K., Charlebois, J. and Chu, K. (2022). Weight smoothing for nonprobability surveys. *TEST* 31(3), 619-643.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* 29, 1189-1232.

Hartley, H.O. (1962). Multiple frame surveys. *In Proceedings of the Social Statistics Section, American Statistical Association* 203-206.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 77(377), 89-96.

Kennedy, C. and Hartig, H. (2019). Response rates in telephone surveys have resumed their decline. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

Kern, C., Li, Y. and Wang, L. (2021). Boosted Kernel Weighting - Using Statistical Learning to Improve Inference from Nonprobability Samples. *Journal of Survey Statistics and Methodology* 9(5), 1088-111.

Kim, J.K. and Tam, S.M. (2021). Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference. *International Statistical Review* 89, (2), 382401.

Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review* 87, 177-191.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T. (2022). *caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret

Lee, B.K., Lessler, J. and Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* 29, 337-346.

Lee, B.K., Lessler, J. and Stuart, E.A. (2011). Weight trimming and propensity score weighting. *PLoS ONE 6* e18174.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J. Off. Stat.* 22(2), 329-349

Lee, S. and Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociol. Method. Res.* 37(3), 319-343.

Liu, Z. and Valliant, R. (2023). Investigating an alternative for estimation from a non-probability sample: Matching plus calibration. *Journal of Official Statistics* 39, 45-78.

Lohr, S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology* 37(2), 197-213.

Lohr, S.L. (2022). Comments on "Statistical inference with non-probability survey samples". Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00005-eng.htm.

Marken, S. (2018). Still Listening: The State of Telephone Surveys. Gallup Methodology Blog. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx.

McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R. and Burgette, L.F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* 32(19), 3388-3414.

McCaffrey, D.F., Ridgeway, G. and Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* 9(4), 403.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology* 33(2), 151-157.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.F., Dessertaine, A. and Puech, P. (2022). QR Prediction for Statistical Data Integration. *TSE Working Paper* 22, 13-44.

Meng, X.-L. (2022). Comments on "Statistical inference with non-probability survey samples" - Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00006-eng.htm.

Nekrašaitė-Liegė, V., Čiginas, A. and Krapavickaitė, D. (2022). Usage of non-probability sample and scraped data to estimate proportions. *Workshop on survey statistics 2022*. https://vb.vgtu.lt/object/elaba:138918140/index.html

Rao, J.N.K.(2020). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B* 83(1), 242-272.

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

Rivers, D. (2007). Sampling for web surveys. *In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA*.

Robbins, M.W., Ghosh-Dastidar, B. and Ramchand, R. (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology* 9(5), 1114-1145.

Rueda, M.D.M., Ferri-García, R. and Castro-Martín, L. (2022). Combining Big Data with probability survey data: a comparison of methodologies for estimation from non-probability surveys. *Padua Research Archive-Institutional Repository* 711.

Rueda, M.D.M., Pasadas-del-Amo, S., Cobo, B., Castro-Martín, L. and Ferri-García, R. (2022). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*. https://doi.org/10.1002/bimj.202200035

Särndal, C.E. and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics* 146-156.

Servy, E., Cuesta, C.B., Marí, G.P.D. and Armida, M.L. (2006). Utilización del paquete "kernsmooth" de r para construir suavizados loess y bandas de variabilidad a datos de la encuesta de ocupación hotelera. http://hdl.handle.net/2133/8792

Tillé, Y. and Matei, A. (2021). *sampling: Survey Sampling.* https://CRAN.R-project.org/package=sampling

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* 8(2), 231-263.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference: A prediction approach*. New York: John Wiley.

Wang, L., Graubard, B.I., Katki, H.A. and Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, https://doi.org/10.1111/rssa.12564.

Wang, Z. and Kim, J.K. (2022). Comments on "Statistical inference with non-probability survey samples". Survey Methodology, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00007-eng.htm.

Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A. and Blom, A.G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology* 8(1), 120-147. https://doi.org/10.1093/jssam/smz051.

Wolter, K. (2007). *Introduction to variance estimation.* New York: Springer.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, Statistics Canada, 48 (2). Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm.

Xi, W., Hinton, A., Lu, B., Krotki, K., Keller-Hamilton, B., Ferketich, A. and Sukasih, A. (2022). Analysis of combined probability and nonprobability samples: a simulation evaluation and application to a teen smoking behaviour survey, *Communications in Statistics - Simulation and Computation* DOI: 10.1080/03610918.2022.2102181.

Yang, S. and Kim, J.K. (2020). Statistical data integration in survey sampling: a review. *Jpn J Stat Data Sci* 3, 625-650. https://doi.org/10.1007/s42081-020-00093-w

Yang, S., Kim, J. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation. *Survey Methodology* 47, 29-58

A. Appendix 1

Regularity conditions for the HT estimator

The first and second order probabilities verify:

1a) $N^{-2}\sum_{i\neq j=1}^{N}(\pi_i\pi_j-\pi_{ij})^r=O(n^{-2r\delta})$

2a) $N^{-1}\sum_{i=1}^{N}(y_i/\pi_i-Y/n)^{2k}<M<\infty$ for $\delta>0$ and $r^{-1}+k^{-1}=1$

Regularity conditions for the KW estimator:

The kernel function $K(u)$, the bandwidth h and the sampling schemes verify:

2a) $K(u)$, $\int K(u)du=1$, $\sup_u|K(u)|<\infty$, y $\lim_{|u|\to\infty}|u||K(u)|=0$

2b) $h=h(n_v)$, $h\to 0$, but $n_v h\to\infty$ as $n_v\to\infty$ and the distributions of the estimated propensity scores in the probability and non-probability samples are interchangeables.