Mathematics and Computers in Simulation 227 (2025) 272–281

Contents lists available at ScienceDirect

Mathematics and Computers in Simulation

journal homepage: www.elsevier.com/locate/matcom

Original articles

A new technique for handling non-probability samples based on model-assisted kernel weighting

Beatriz Cobo [a,*], Jorge Luis Rueda-Sánchez [b], Ramón Ferri-García [b], María del Mar Rueda [b]

[a] *Department of Quantitative Methods for Economics and Business - University of Granada, Faculty of Economics and Business, Campus Universitario de Cartuja, Granada, 18011, Spain*

[b] *Department of Statistics and Operational Research - University of Granada, Sciences Faculty, Campus Fuentenueva S/N, Granada, 18071, Spain*

ARTICLE INFO

Keywords:
Model-assisted kernel
Kernel weighting
Employment
Confinement period
COVID-19

ABSTRACT

Surveys are going through massive changes, and the most important innovation is the use of non-probability samples. Non-probability samples are increasingly used for their low research costs and the speed of the attainment of results, but these surveys are expected to have strong selection bias caused by several mechanisms that can eventually lead to unreliable estimates of the population parameters of interest. Thus, the classical methods of statistical inference do not apply because the probabilities of inclusion in the sample for individual members of the population are not known. Therefore, in the last few decades, new possibilities of inference from non-probability sources have appeared.

Statistical theory offers different methods for addressing selection bias based on the availability of auxiliary information about other variables related to the main variable, which must have been measured in the non-probability sample. Two important approaches are inverse probability weighting and mass imputation. Other methods can be regarded as combinations of these two approaches.

This study proposes a new estimation technique for non-probability samples. We call this technique model-assisted kernel weighting, which is combined with some machine learning techniques. The proposed technique is evaluated in a simulation study using data from a population and drawing samples using designs with varying levels of complexity for, a study on the relative bias and mean squared error in this estimator under certain conditions. After analyzing the results, we see that the proposed estimator has the smallest value of both the relative bias and the mean squared error when considering different sample sizes, and in general, the kernel weighting methods reduced more bias compared to based on inverse weighting. We also studied the behavior of the estimators using different techniques such us generalized linear regression versus machine learning algorithms, but we have not been able to find a method that is the best in all cases. Finally, we study the influence of the density function used, triangular or standard normal functions, and conclude that they work similarly.

A case study involving a non-probability sample that took place during the COVID-19 lockdown was conducted to verify the real performance of the proposed methodology, obtain a better estimate, and control the value of the variance.

[*] Corresponding author.
E-mail address: beacr@ugr.es (B. Cobo).

https://doi.org/10.1016/j.matcom.2024.08.009
Received 23 October 2023; Received in revised form 24 June 2024; Accepted 7 August 2024
Available online 13 August 2024

1. Introduction

Survey research is one of the most important sources of information on populations. Such research is typically conducted by selecting a probability sample because this enables the generalization of the results of the study to the entire population.

In recent years, however, non-probability samples have boomed, and estimation from non-probability data have been investigated more intensively by the survey methodology community. The analysis of databases that compile surveys carried out during the first year of the COVID-19 pandemic has made it possible to quantify this trend toward non-probability sampling designs. The latter represent 38% of the surveys included in Oxford Supertracker [1] and 92% in a review of surveys related to COVID-19 in Spain [2].

Non-probability sampling is a method that selects the elements of the population under study using non-random methods, i.e. subjective methods. This means that we cannot know precisely the probabilities that each individual belongs to our sample (inclusion probabilities), so we cannot correctly construct the estimators developed in the probability context. These methods are faster, easier, and cheaper because a complete sampling frame is not necessary to obtain the samples, as in the case of probability samples. On the other hand, the probabilities of selection of the units are not known, and this does not allow the generalization of the study results to the entire population through classical statistical inference; as a consequence, the precision of non-probability samples cannot be measured accurately, see Kalton [3].

The primary approaches for inference from non-probability sources are quasirandomization [4], superpopulation modeling [5], and doubly robust estimation [6]. In the quasirandomization approach, the unknown inclusion probabilities are estimated and the inverse of these probabilities is used as weights in the same manner as in design-based inference. In superpopulation modeling, the values of variables of interest for unsampled units are predicted using specified models. The doubly robust estimation combines quasirandomization and superpopulation modeling. These methods for treating non-probability samples are based on modeling either the probability of being included in the sample or, the variable under study, given a set of covariates. Usually, linear regression models are considered for estimation, but recently, some machine learning models have been used in this context. In this field, an immense number of papers have been presented in recent years, reflecting the importance assigned to this question (see e.g [5,7,8] or [9] among others), including simulation studies to compare the efficiency of these alternatives for inference, but no clear results have been obtained regarding the best method to use.

Our work focuses on a new procedure to obtain estimations from non-probability samples data, seeking to achieve more accurate and reliable estimates through model-assisted estimators, which are designed to be approximately unbiased if the distribution that generates the probabilities or the superpopulation model are correctly specified. We will use kernel weighting [10] and some machine learning techniques to create pseudo-weights using a reference probability sample. Section 2 describes the methodology used to construct the new estimators. In Section 3, a simulation study is conducted. A non-probability data selection scenario is implemented, and the behavior of the proposed inference method is evaluated by comparing the estimation results with known population totals. In Section 4, we apply the proposed method to a non-probability sample [11], which was carried out in Spain during the COVID-19 lockdown. Finally, Section 5 presents the conclusion of the work and the results obtained.

2. Methodology

Let be s_v the non-probability sample obtained from population U, using a volunteer survey with size n_v. The response variable is denoted by y and the covariate vector is $\mathbf{x} = (x_1, \dots, x_p)$. From the sample s_v, we can estimate the population mean $\overline{Y} = \frac{1}{N}\sum_U y_i$ by the sample mean $\overline{y}_v = \sum_{i\in s_v} \frac{y_i}{n_v} = \sum_U \frac{1_{vi} y_i}{n_v}$ where $1_{vi} = 1$ for $i \in s_v$ and $1_{vi} = 0$ elsewhere. Bias for this estimator $\overline{y}_v$ [12] is given by

$$E(\overline{y}_v - \overline{Y}_N) = \frac{1}{f_v} E\{Cov(1_v, y)\}, \tag{1}$$

where $f_v = n_v/N$. Thus, we have

$$MSE(\overline{y}_v) = \frac{1}{f_v^2} E\{Corr(1_v, y)^2\} Var(1_v) Var(y). \tag{2}$$

The term $E\{Corr(1_v, y)^2\}$ is the term most critical and therefore, a non-probability sampling where $E\{Corr(1_v, y)\} \neq 0$ induces a selection bias to the results. Probability distribution underpinning the notation E() and MSE() is the distribution of the random mechanism for the non-probability sample.

If we seek to eliminate volunteer bias in this type of sampling, we will need accurate auxiliary information related to the topic that we study in our survey. Depending on the auxiliary information available, we may use different bias reduction techniques [13]:

- Population totals of covariates (often called control totals).
- Covariate values for every element in a reference probability sample.
- Covariate values for every element in the entire population.

In this study, we consider the second situation. We denote the probability sample by s_r, also called the reference sample, extracted from the target population U with size n_r and $\mathbf{x} = (x_1, \dots, x_p)$ being the same covariate vector measured in the non-probability sample s_v. For s_r, the probability that have each individual of belonging to our sample is known and greater than zero, which are necessary conditions for a probability sample. These probabilities are also known as first-order inclusion probabilities (π_r), and we can obtain the design weights (w_r) through its inverse, which is a key part of the estimates. In non-probability samples, inclusion probabilities are not known due to the lack of a probability theory to support them since the selection mechanism is non-random. In the quasirandomization approach, we estimate the pseudo-inclusion probabilities. These estimated probabilities are called propensities (π_v) and will be the approach that we will use in this study. We assume an ignorable selection mechanism of the non-probability sample, that is $\forall i \in U$:

$$\pi_{vi} = P\left[1_{vi} = 1|\mathbf{x}_i\right] \text{ where } 1_{vi} = \begin{cases} 1 & \text{for } i \in s_v \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We will elaborate a procedure for defining estimators for the population mean that includes (1) estimation of the propensity scores, (2) use of the estimated propensities to construct weights for the units using kernel weighting, and (3) definition of a model-assisted estimator that uses the propensity score model and regression model for the response variable. This proposal is described below.

2.1. Estimation of propensity scores

We assume that the selection mechanism of s_v follows the following model:

$$\pi_{vi} = P(1_{vi} = 1|\mathbf{x}_i) = p_i(\mathbf{x}) = m(\boldsymbol{\gamma}, \mathbf{x}_i), \quad i = 1, \ldots, N, \tag{4}$$

where $m(\cdot)$ is a given function with second continuous derivatives with respect to $\boldsymbol{\gamma}$.

We aimed to estimate propensity scores using data from both samples. First, we obtain $\hat{\gamma}$ that maximizes the pseudo-likelihood [6]:

$$\tilde{l}(\boldsymbol{\gamma}) = \sum_{s_v} log \frac{m(\boldsymbol{\gamma}, \mathbf{x}_i)}{1 - m(\boldsymbol{\gamma}, \mathbf{x}_i)} + \sum_{s_r} \frac{1}{\pi_{ri}} log(1 - m(\boldsymbol{\gamma}, \mathbf{x}_i)). \tag{5}$$

Once the pseudo-maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ is obtained, the estimated propensities $\hat{\pi}_{vi} = m(\hat{\boldsymbol{\gamma}}, \mathbf{x}_i)$ are used to readjust the propensity bias of the volunteer sample.

Due to the binary nature of 1_{vi}, logistic regression models are usually considered, but in the last few years, other machine learning (ML) models have been used (e.g. [14]. A gradient-boosting machine (GBM) will be used in our work. This technique works as an ensemble of weak classifiers. This algorithm converges toward the minimum value of the loss function by an iterative process. Then the propensity is:

$$\hat{\pi}_{vi} = \mathbf{v}^T J\left(\mathbf{x}_i\right), i \in s_v \tag{6}$$

where $J\left(\mathbf{x}_i\right)$ is a matrix of terminal nodes of m decision trees used for the boosting and $\mathbf{v}$ is a vector indicating the weight of each tree.

Once we have computed the propensities $\hat{\pi}_v$, we will explain the technique for designing the weights for the units included in the non-probability sample. The most used technique is inverse weighting [6], which is provided by the estimator $\hat{\bar{Y}}_{PSA} = \sum_{i \in s_v} \frac{1}{N} \frac{y_i}{\hat{\pi}_{vi}}$. This method is sensitive to the propensity model specification and produces highly uncertain estimates when there are extreme weights [10]. In this work, we will focus on the kernel weighting (KW) method, which is less sensitive to model misspecification and avoids extreme weights. The KW method has produced promising results [15].

2.2. Constructing weights for the units using the kernel weighting method (KW)

The KW method [10] constructs its new weights by weighting the design weights w_r from the probability sample, which are known and well-constructed, according to the similarity between the individuals from both samples. The idea is that if an individual in the non-probability sample has similar characteristics (according to its covariates' values) to those in the probability sample, then both will have similar inclusion probabilities and therefore will have similar design weights. To obtain these similarities, we compute the distance between individuals through the difference in the propensity scores, which is based on the values of the covariates:

$$d_{ij} = \widehat{\pi}_{vi} - \widehat{\pi}_{vj}, \ i \in s_v, \ j \in s_r \tag{7}$$

These distances are symmetric, continuous, and positive; therefore they can be density functions of random distributions such as the normal function or the triangular function. As we want to obtain proportions, called kernel weights, that measure the similarity between individuals, the following expression is used:

$$k_{ij} = \frac{K\{d_{ij}/h\}}{\sum_{i \in s_v} K\{d_{ij}/h\}} \tag{8}$$

being $K\{.\}$ a kernel function centered at zero, and h the corresponding bandwidth. These weights will take values between zero and one, and their sum for all individuals in the non-probability sample must be equal to one [10]. In this case, the larger the value of these weights, the smaller the difference between individuals. Finally, to compute the pseudoweights, we multiply the design weights w_{rj}, $\forall j \in s_r$, by the kernel weights k_{ij}, which measure the similarity between individuals, and do the sum for each individual $i \in s_v$. In this way, we obtain a pseudoweight for each individual in the non-probability sample of the form:

$$w_i^{KW} = \sum_{j \in s_r} w_{rj} k_{ij} \tag{9}$$

We can define the new estimator:

$$\widehat{\overline{Y}}_{KW} = \frac{1}{N}\sum_{i\in s_v} w_i^{KW} y_i. \tag{10}$$

Wang et al. [10] showed that the KW estimator $\widehat{\overline{Y}}_{KW} \to \overline{Y}$ in probability under certain conditions of the kernel function (see Appendix A).

2.3. The proposed new model-assisted estimator

We can also construct a model-assisted estimator using the reduction bias technique called statistical matching [16]. This is a model-assisted estimator because it is based on the modelization of the relationship between y and the vector $\mathbf{x}$, using a training set the non-probability data obtained from s_v. For this, we have to assume that the target population is a realization of a superpopulation model M:

$$y_i = M\left(\mathbf{x_i}\right) + e_i,\ \ i = 1,\dots,N \tag{11}$$

being $M\left(\mathbf{x_i}\right) = E_M[y_i|\mathbf{x_i}]$ and $e \sim N(0,\sigma)$. Once we have our model that explains the relationship between both variables, we will predict the value of the response variable in s_r using the covariate values from individuals that belong to this sample:

$$\hat{y}_i = E_M\left[y_i|\mathbf{x_i}, 1_i\right],\ \ i \in s_r. \tag{12}$$

This estimation will depend on the model used, producing different results. We can use models like general linear models (GLM) or a more innovative approach using machine learning models. One of the techniques that produces better results is extreme gradient boosting (XGBoost), which has been shown to produce better results than traditional models [17].

The mean of the variable under study can be expressed as

$$\overline{Y} = \frac{1}{N}(\sum_{k\in U}\hat{y}_k + \sum_{k\in U}\left(y_k - \hat{y}_k\right)). \tag{13}$$

Thus, we can compute an estimator using the generalized difference predictor given in Cassel et al. [18], as:

$$\widehat{\overline{Y}}_{MAKW} = \frac{1}{N}(\sum_{j\in s_r} w_{rj}\hat{y}_j + \sum_{j\in s_v} w_j^{KW}\left(y_j - \hat{y}_j\right)), \tag{14}$$

where w_j^{KW} are the weights from kernel weighting and the first term of the function is the Horvitz-Thompson estimator of the mean of the predicted values.

We call this estimator the model-assisted kernel weighting estimator (MAKW).

Now we discussed the properties of the proposed estimator. We assume the regularity conditions given in Appendix A, and we also assume a parametric regression model $M\left(\mathbf{x_i}\right) = f(\mathbf{x_i}, \boldsymbol{\beta})$. Let $\widehat{\overline{Y}}_{HT} = \frac{1}{N}\sum_{j\in s_r} w_{rj}y_j$ the Horvitz-Thompson estimator, note that

$$\widehat{\overline{Y}}_{MAKW} - \widehat{\overline{Y}}_{HT} = \frac{1}{N}\sum_{i\in s_v} w_i^{KW}\hat{e}_i - \frac{1}{N}\sum_{j\in s_r} w_{rj}\hat{e}_j, \tag{15}$$

where $\hat{e}_i = y_i - f(\mathbf{x_i}, \hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ a consistent estimator of $\boldsymbol{\beta}$.

If model M is correctly specified, under the regularity conditions given in Appendix A, we have:

$$E(\widehat{\overline{Y}}_{MAKW} - \widehat{\overline{Y}}_{HT}) \approx \frac{1}{N}(\sum_{j\in U} e_j - \sum_{j\in s_r} w_{rj}e_j), \tag{16}$$

where $e_j = y_j - f(\mathbf{x_j}, \boldsymbol{\beta}^*)$ and $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}$. The second term is asymptotically design unbiased for $\frac{1}{N}\sum_{j\in U} e_j$, so the estimator $\widehat{\overline{Y}}_{MAKW}$ is asymptotically unbiased under the pseudo-inclusion model.

On the other hand, if the outcome regression model is correctly specified, then $E_M(\hat{e}_j) \approx 0$, thus $E_M(\widehat{\overline{Y}}_{MAKW} - \widehat{\overline{Y}}_{HT}) \approx 0$ and $\widehat{\overline{Y}}_{MAKW}$ is an approximately unbiased estimator of the population mean. Therefore, we have established double robustness of $\widehat{\overline{Y}}_{MAKW}$, our proposed estimator is approximately unbiased if the pseudo-inclusion or superpopulation models are correctly specified.

2.4. Variance estimation

There is no agreement on how to determine the efficiency of estimators based on non-probability samples. The variance of the proposed estimators could be calculated with respect to the distribution of the probability sample design, with respect to the random mechanism of the pseudo-inclusion model, and with the superpopulation model M. Therefore, there are three sources of variation.

It should be noted that the sample design used for obtaining s_r influences the calculation of the propensities and must be considered as a source of variation in the estimation of the variance component due to the use of s_v. Valliant [7] proposed an estimator of variance for the propensity score adjustment (PSA) estimator $\widehat{\overline{Y}}_{PSA} = \sum_{i\in s_V}\frac{1}{N}\frac{y_i}{\hat{\pi}_{vi}}$ using linearization, when $\hat{\pi}_{vi}$ is estimated using the logistic model. However, this estimator does not consider the distribution of the probability sample design, and consequently, the variance is underestimated. This author also proposed a grouped jackknife in which both the random mechanism of

Table 1
$|RB|$ and RMSRE of the estimators.

	n_r = 1000, n_v = 500		n_r = 1000, n_v = 1000		n_r=1000, n_v = 1500	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\hat{\hat{Y}}_{PSA}$	2,2513	2,3776	2,2514	2,3399	2,2278	2,2830
$\hat{\hat{Y}}_{KW}$	0,8980	1,1029	0,8220	1,0152	0,7570	0,9251
$\hat{\hat{Y}}_{DR}$	1,1243	1,3513	1,1677	1,3649	1,1612	1,3373
$\hat{\hat{Y}}_{MAKW}$	0,8080	0,9806	0,7781	0,9676	0,7546	0,9244

Table 2
$|RB|$ and RMSRE of the proposed estimator for different techniques.

	n_r = 1000, n_v = 500		n_r = 1000, n_v = 1000		n_r = 1000, n_v = 1500	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\hat{\hat{Y}}_{MAKW\_GLM}$	0,8080	0,9806	0,7781	0,9676	0,7546	0,9244
$\hat{\hat{Y}}_{MAKW\_KNN}$	0,8347	1,0249	0,6977	0,8891	0,6677	0,8184
$\hat{\hat{Y}}_{MAKW\_NN}$	0,9644	1,2140	0,8124	0,9918	0,7433	0,9461
$\hat{\hat{Y}}_{MAKW\_GBM}$	0,8372	1,0114	0,8007	0,9827	0,7917	0,9606

Table 3
$|RB|$ and RMSRE for normal kernel density.

	n_r = 1000, n_v = 500		n_r = 1000, n_v = 1000		n_r = 1000, n_v = 1500	
	|RB|	RMSRE	|RB|	RMSRE	|RB|	RMSRE
$\hat{\hat{Y}}_{KW\_SN}$	0,8644	1,0587	0,7808	0,9655	0,6900	0,8517
$\hat{\hat{Y}}_{MAKWN\_SN}$	0,8142	0,9888	0,7530	0,9346	0,6915	0,8600

the pseudo-inclusion model and the sampling design are repeated for each replicate. This replication method is more computationally complex but can implicitly reflect the variation due to the estimated weights. However, its consistency has not been formally proven.

Variance estimation for the model-assisted estimator is a challenging problem. It is difficult to construct a consistent variance estimator with unknown scenarios on the model specifications [9]. Chen et al. [6] described a technique for the variance of the estimator $\widehat{\hat{Y}}_{DR} = \frac{1}{N}(\sum_{j\in s_r} w_{rj}\hat{y}_j + \sum_{j\in s_v} \frac{1}{\hat{\pi}_{vj}}\left(y_j - \hat{y}_j\right))$, when using linear models for the propensities and for modeling the y values. The technique is delicate with various issues for practical applications. On the other hand, there are hardly any results on the theoretical properties of variance estimators when more complex machine learning models and techniques are used, as in our case.

In the real application described in the following section, we have chosen a jackknife methodology [19] similar to that used by Valliant [7]. In this replication variance estimation method, each s_v/s_r combination is divided into random groups, and the model for predicting the propensities is readjusted in every group, i.e., both the quasirandomization given by propensities and regression model estimation steps are calculated for each replicate.

3. Simulation study

To test the behavior of the inference methods described in the previous section, we simulate a population U of size 10000 with one variable of interest and six auxiliary variables $\mathbf{x} = (x_1, \dots, x_6)$ that we use to estimate the propensities, specifically x_1, x_3, x_5, followed a Bernoulli distribution with $p = 0.6$ and x_2, x_4, x_6 followed a Normal distribution with a standard deviation of 0.5 and a mean parameter dependent on the value of the previous Bernoulli variable for each individual

$$x_{1i}, x_{3i}, x_{5i} \sim Be(0.6) \qquad i \in U$$

$$x_{ji} \sim N(\mu_{ji}, 0.5) \qquad i \in U, j = 2, 4, 6$$

$$\mu_{ji} = \begin{cases} 1 & \text{if} \;\; x_{(j-1)i} = 1 \\ 0 & \text{otherwise} \end{cases}, \quad i \in U, j = 2, 4, 6. \tag{17}$$

By using simple random sampling without replacement, we extracted $n_r = 1000$ samples and non-probability samples were extracted by Poisson sampling with probability

$$ln\left(\frac{\pi_v}{1-\pi_v}\right) = -1 + \sqrt{3\pi}x_2x_3 \tag{18}$$

for different sample sizes $n_v = 500, 1000$, and 1500. Finally, the variable of interest is defined as

$$y = N(3, 0.5) + \pi_v + 2x_5. \tag{19}$$

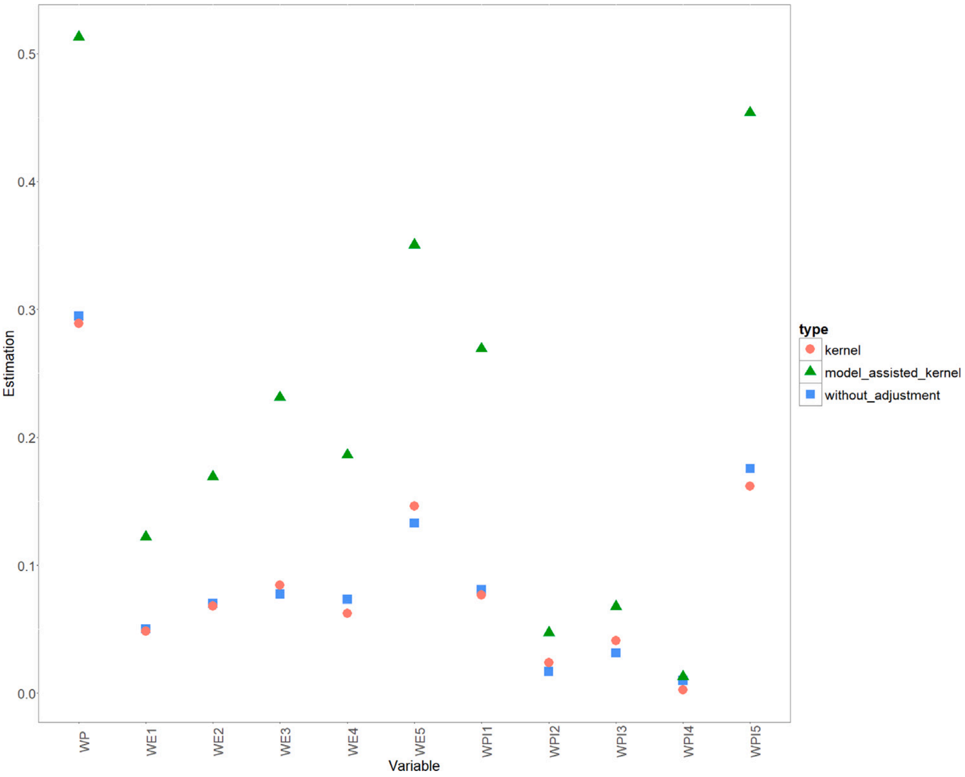

Fig. 1. Without adjustment, kernel and model-assisted kernel estimations for variables of interest.

where π_v is the probability defined by Poisson sampling, so it would be an example of missing not at random data, where the outcome is directly related to the selection mechanism. We use generalized linear models to estimate the propensities and to obtain the predicted values $\hat{y}_i$. We used the triangular density on (−3, 3) (previously used by [10]) as the selected kernel function. We focus on the model-assisted kernel weighting (MAKW) discussed in this paper. We also calculate three other estimators: propensity score adjustment (PSA), kernel weighting (KW) and doubly robust (DR) estimators. To compare the results, we calculate the relative bias ($|RB|$) of the estimators and the root mean square relative error ($RMSRE$)

$$|RB| = \frac{1}{1000}\sum_{i=1}^{1000}\frac{|\bar{y}_i - \bar{Y}|}{\bar{Y}} * 100 \tag{20}$$

$$RMSRE = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\frac{\bar{y}_i - \bar{Y}}{\bar{Y}}\right)^2} * 100 \tag{21}$$

being $\bar{y}_i$ an estimate of $\bar{Y}$ computed for the ith sample.

The R packages sampling, survey, caret, NonProbEst, and KWML are used to calculate propensities, sampling weights, and estimators.

The simulation results in various values of n_v are reported in Table 1.

Results from Table 1 can be summarized in the following points: (1) the proposed estimator $\hat{\bar{Y}}_{MAKW}$ has the smallest of both |RB| and RMSRE. (2) In general, the KW- methods (including $\hat{\bar{Y}}_{KW}$ and $\hat{\bar{Y}}_{MAKW}$) reduced bias more than methods based on inverse weighting ($\hat{\bar{Y}}_{PSA}$ and $\hat{\bar{Y}}_{DR}$). (3) As sample size n_v increases, the biases and errors of the estimators decrease.

The second study examines the behavior of the estimator when using different techniques, including generalized linear regression (GLM) and machine learning techniques, specifically, K nearest neighbor (KNN), neural network (NN), and gradient boosting machine (GBM). The simulation results are reported in Table 2.

Linear models seem to be the most robust compared to machine learning models. In some cases, other methods achieve good results, but we do not find a method that is the best in all cases.

Finally, we will study the influence of the density function used. We also considered the standard normal density (SN). Results are shown in Table 3.

In view of the results, we see that both density functions work in a similar way with respect to $|RB|$ and RMSRE, although it seems that using the normal density produces slightly smaller values. Therefore, we can say that there is little influence of the kernel function in this study.

In summary, this simulation study shows the robustness of the proposed estimator, which is already capable of reducing bias in situations in which the model is poorly specified.

4. Survey on habits and living conditions in Spain during home confinement

The proposed method is applied to data obtained from a survey conducted in Spain during the COVID-19 lockdown [11]. This survey was conducted between 28th April and 14th May in 2020, and it is a non-probability sample because they used a snowball sampling via email and social media, so we can expect a significant lack of representativity and the estimates obtained will be highly biased. With this survey, they wanted to get information about how the people in Spain lived and felt during these lockdowns, so they asked about household habits, employment issues, health, and politics. They also obtain some sociodemographic variables that define each individual who participated in the survey (auxiliary variables). Some of these variables are the same as those used in the CIS Barometer of May 2020 [20]. We will obtain our reference sample from this survey because it was conducted by an official institution following strict methodologies and we consider it a probability survey. The CIS is the Spanish Sociological Research Center; therefore, both surveys have the same target population and both were developed during the same period.

The auxiliary variables that both samples have in common are: age, autonomous community, education level, confidence in the government during the pandemic, employment status, intended electoral vote, last electoral vote, province, sex, and urban density.

In Fig. B.2 (Appendix B), the relative frequencies of some auxiliary variables can be obtained for both samples and the population values obtained from official national organizations when these are available. If we observe the autonomous community, we can see that in most communities, the non-probability sample under-represents this value; however, in the community of Ceuta, it is highly over-represented. In the case of the urban density variable, the non-probability sample under-represents the population value in all categories, except in the last category, that is, in large urban centers. With respect to sex, in the man category, both samples were underrepresented with respect to the population value. Regarding the age variable, we have considered three age ranges, with the lower limit being 18 years, since we are conducting a study on economic variables, specifically on work issues. In this case, in the first two categories, the non-probability sample under-represents the population, and in the third category, which includes workers over 45 years of age, it is over-represented in both samples. We compute three different estimators: The naive estimator without adjustment $\bar{y}_v$, the kernel estimator $\hat{\hat{Y}}_{KW}$, and the model-assisted kernel estimator, $\hat{\hat{Y}}_{MAKW}$.

To obtain the results, we analyzed the questions with binary answers Yes/No from Pérez et al. [11] related to productivity and work experience during confinement, specifically, 3001 A (which we will called WP), 3001B (WE1, WE2, WE3, WE4, WE5), and 3001C (WPI1, WPI2, WPI3, WPI4, WPI5).

The differences between the estimations obtained considering the $\bar{y}_v$, $\hat{\hat{Y}}_{KW}$, and $\hat{\hat{Y}}_{MAKW}$ estimators for each of the main questions are observed in Fig. 1.

In Fig. 1 it can be seen that in all the variables of interest, the estimate using the model-assisted kernel is higher than that obtained without adjustment or using only KW. Regarding the variance, if we look at the Table C.4 (Appendix C), we see that the two methods that use the kernel have greater variance than the naive estimator, as we expected, since we must include the variability due to the calculation of the propensities.

5. Conclusions

The use of non-probability samples for academic research and official statistics requires a valid inference framework based on sample data and possible auxiliary information. We consider the situation in which this auxiliary information is provided by a probability sample representing the population well but excludes the study variable.

The proposed estimation method employs explicit assumptions for the outcome regression and quasirandomization models. The method is based on kernel weighting methodology that uses estimated propensities as a measure of similarity and is therefore less sensitive to model misspecification than the usual inverse propensity method [7]. The proposed estimator has good robustness properties against deviations in the model specification of the propensity score model or the outcome model. We have also confirmed with a simulation study that the use of this method produces an effective bias reduction while controlling variance. The practical application carried out shows the importance of the differences in the estimates obtained between the different methods.

CRediT authorship contribution statement

Beatriz Cobo: Writing – review & editing, Writing – original draft, Software, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jorge Luis Rueda-Sánchez:** Writing – original draft, Software, Formal analysis, Data curation. **Ramón Ferri-García:** Writing – original draft, Methodology, Funding acquisition. **María del Mar Rueda:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Conceptualization.

Acknowledgments

This work is part of grant PDC2022-133293-I00 funded by the MCIN/AEI/10.13039/5011000 11033, Spain and the European Union "NextGenerationEU"/PRTR"1 and the grant FEDER C-EXP-153-UGR23 funded by the Consejería de Universidad, Investigación e Innovación and by the ERDF Andalusia Program 2021–2027, Spain. The research was also partially supported from IMAG-María de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and PPJIA2023-030 of the University of Granada, Spain. Funding for open access charge: Universidad de Granada / CBUA.

Appendix A. Regularity conditions

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 $r^{-1}+k^{-1}=1$
The kernel function $K(u)$, the bandwidth h, and the sampling schemes verify the following:
(2a) $K(u)$, $\int K(u)du=1$, $\sup_u|K(u)|<\infty$, and $\lim_{|u|\to\infty}|u||K(u)|=0$
(2b) $h=h(n_v)$, $h\to 0$, $n_v\ h\to\infty$ when $n_v\to\infty$
The distributions of the estimated propensity scores in the probability and non-probability samples are interchangeable.

Appendix B. Relative frequencies in both surveys and the population values for the covariables

See Fig. B.2.

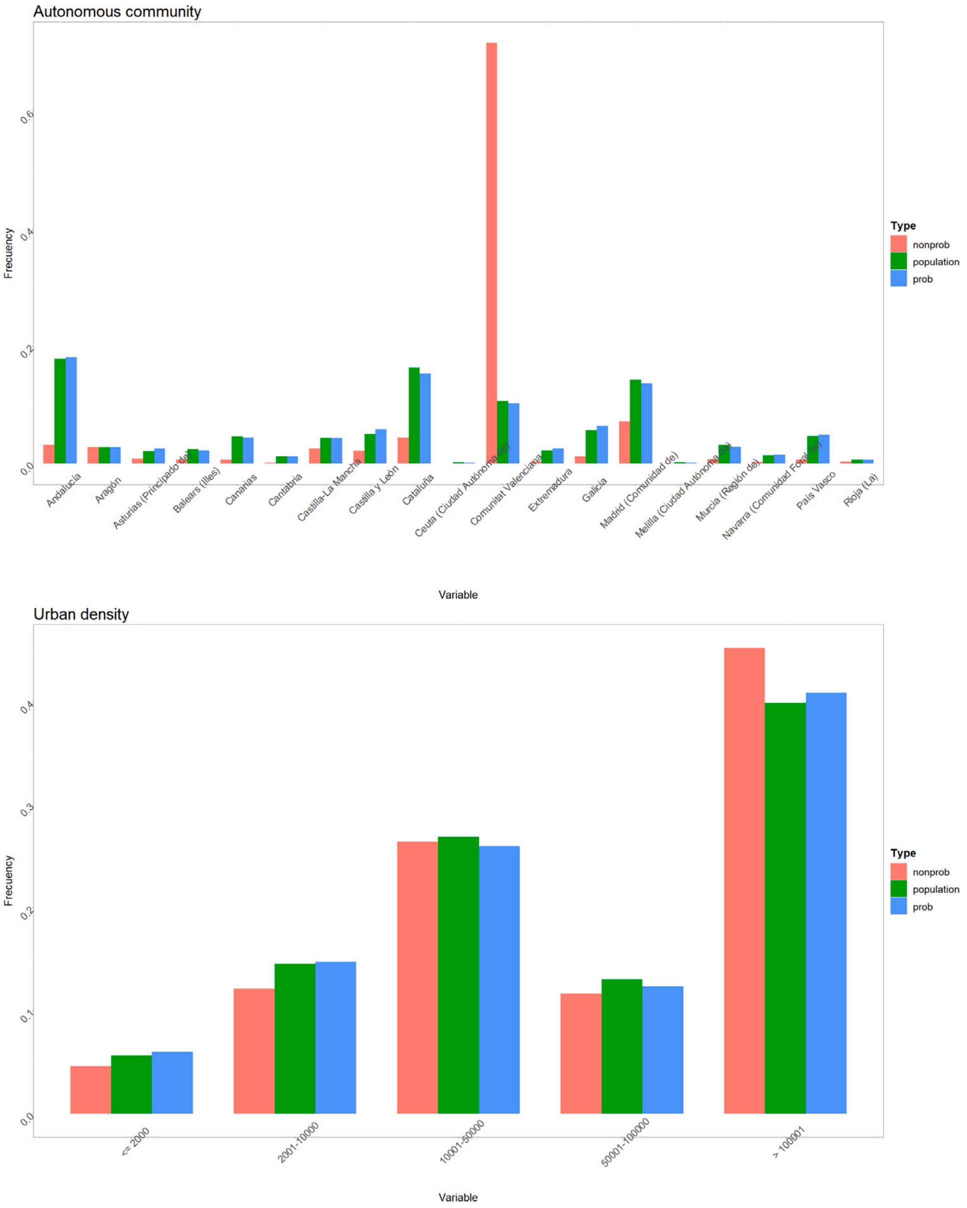

Fig. B.2. Relative frequencies.

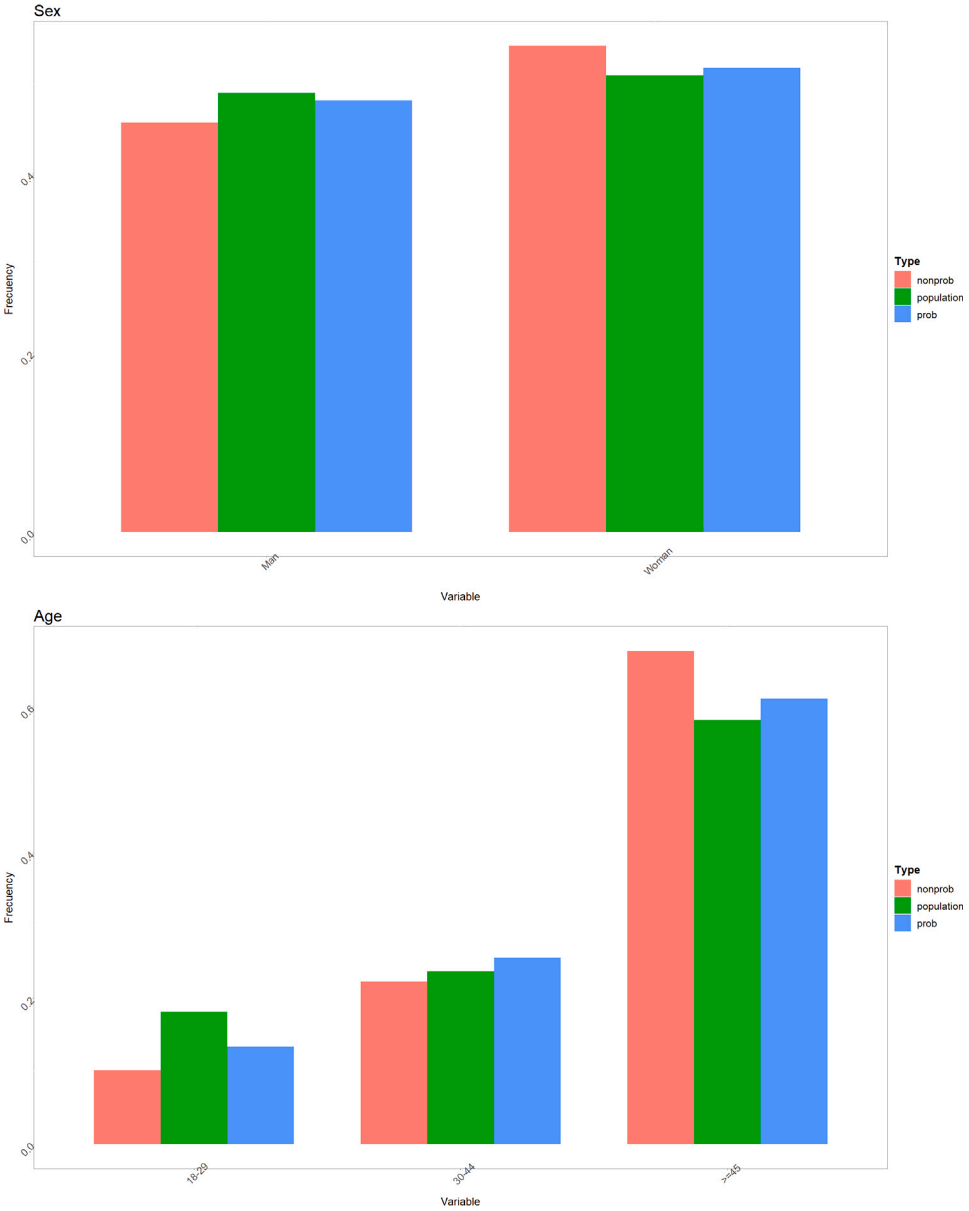

Fig. B.2. (*continued*).

Appendix C. Without adjustment, kernel and model-assisted kernel estimations and variances for variables of interest

See Table C.4.

Table C.4
Without adjustment, kernel, and model-assisted kernel estimations and variances for variables of interest.

Variable	Without adjustment		Kernel		Model-assisted kernel	
	Estimation	Variance	Estimation	Variance	Estimation	Variance
WP	0,2949	1,24E−05	0,2835	8,99E−04	0,5071	2,16E−04
WE1	0,0503	4,98E−06	0,0510	7,93E−04	0,1229	7,56E−05
WE2	0,0702	6,39E−06	0,0685	1,09E−03	0,1722	1,19E−04
WE3	0,0775	6,82E−06	0,0808	1,15E−03	0,2292	1,18E−04
WE4	0,0736	6,59E−06	0,0644	1,23E−03	0,1860	1,24E−04
WE5	0,1329	8,74E−06	0,1469	1,34E−03	0,3500	1,27E−04
WPI1	0,0807	7,00E−06	0,0759	9,75E−04	0,2696	8,32E−05
WPI2	0,0168	1,89E−06	0,0266	7,17E−04	0,0486	5,17E−05
WPI3	0,0315	3,35E−06	0,0399	9,13E−04	0,0658	7,66E−05
WPI4	0,0099	1,14E−06	0,0030	6,60E−06	0,0131	7,48E−07
WPI5	0,1755	8,52E−06	0,1674	1,28E−03	0,4542	1,09E−04

References

[1] M. Daly, B. Ebbinghaus, L. Lehner, M. Naczyk, T. Vlandas, Oxford Supertracker: The Global Directory for COVID Policy Trackers and Surveys, Department of Social Policy and Intervention, 2020, https://supertracker.spi.ox.ac.uk/.

[2] C. Sánchez-Cantalejo, D. Yucumá, M.M. Rueda, et al., Scoping review of the methodology of large health surveys conducted in Spain early on the COVID-19 pandemic, Front. Public Health 11 (2023) 1217519.

[3] G. Kalton, Introduction to Survey Sampling, Sage Publications, Newbury Park, CA, ISBN: 9781544338569, 1983.

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

[5] B. Buelens, J. Burger, J.A. van den Brakel, Comparing inference methods for non-probability samples, Internat. Statist. Rev. 86 (2) (2018) 322–343.

[6] Y. Chen, P. Li, C. Wu, Doubly robust inference with non-probability survey samples, J. Amer. Statist. Assoc. 115 (532) (2020) 2011–2021.

[7] R. Valliant, Comparing alternatives for estimation from non-probability samples, J. Surv. Stat. Methodol. 8 (2) (2020) 231–263.

[8] J.N.K. Rao, On making valid inferences by integrating data from surveys and other sources, Sankhya B 83 (2022) 242–272.

[9] C. Wu, Statistical inference with non-probability survey samples, Surv. Methodol. Statist. Canada 48 (2) (2022) 283–311.

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

[11] V. Pérez, C. Aybar, J.M. Pavía, Dataset of the COVID-19 lockdown survey conducted by GIPEyOP in Spain, Data Brief 40 (2022) 107700.

[12] J.K. Kim, Z. Wang, Sampling techniques for big data analysis, Internat. Statist. Rev. 87 (2019) 177–191.

[13] M.D.M. Rueda, R. Ferri-García, L. Castro, The R package Non-ProbEst for estimation in non-probability surveys, R J. 12 (1) (2020) 406–418.

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

[15] C. Kern, Y. Li, L. Wang, Boosted kernel weighting–using statistical learning to improve inference from non-probability samples, J. Surv. Stat. Methodol. 9 (5) (2021) 1088–1113.

[16] D. Rivers, Sampling for web surveys, in: Handbook of Web Surveys, 2007, CorpusID: 12615042.

[17] L. Castro-Martín, M.M. Rueda, R. Ferri-García, C. Hernando-Tamayo, On the use of Gradient Boosting methods to improve the estimation with data obtained with self-selection procedures, Mathematics 9 (23) (2021) 2991.

[18] C.M. Cassel, C.E. Sarndal, J.H. Wretman, Some results on generalized difference estimation and generalized regression estimation for finite populations, Biometrika 63 (1976) 615–620.

[19] K.M. Wolter, Introduction to Variance Estimation, 2nd ed., Springer, Inc., New York, NY, USA, ISBN: 978-0-387-32917-8, 2007.

[20] Centro de Investigaciones Sociológicas (CIS), Barómetro de mayo 2020, Estudio n^o 3281, 2020, https://www.cis.es/cis/export/sites/default/-Archivos/Marginales/3280_3299/3281/es3281mar.pdf.