Statistical Papers
https://doi.org/10.1007/s00362-022-01296-x

REGULAR ARTICLE

Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

Ramón Ferri-García[1] · María del Mar Rueda[1]

Received: 25 February 2021 / Revised: 3 February 2022 / Accepted: 8 February 2022

Abstract

The development of new survey data collection methods such as online surveys has been particularly advantageous for social studies in terms of reduced costs, immediacy and enhanced questionnaire possibilities. However, many such methods are strongly affected by selection bias, leading to unreliable estimates. Calibration and Propensity Score Adjustment (PSA) have been proposed as methods to remove selection bias in online nonprobability surveys. Calibration requires population totals to be known for the auxiliary variables used in the procedure, while PSA estimates the volunteering propensity of an individual using predictive modelling. The variables included in these models must be carefully selected in order to maximise the accuracy of the final estimates. This study presents an application, using synthetic and real data, of variable selection techniques developed for knowledge discovery in data to choose the best subset of variables for propensity estimation. We also compare the performance of PSA using different classification algorithms, after which calibration is applied. We also present an application of this methodology in a real-world situation, using it to obtain estimates of population parameters. The results obtained show that variable selection using appropriate methods can provide less biased and more efficient estimates than using all available covariates.

Keywords Online surveys · Propensity Score Adjustment · Selection bias · Variable selection · Raking calibration

This study was partially supported by Ministerio de Ciencia e Innovación, Spain [Grant No. PID2019-106861RB-I00/AEI/10.13039/501100011033]. The first-named author also received a FPU grant from Ministerio de Ciencia, Innovación y Universidades, Spain.

✉ María del Mar Rueda
mrueda@ugr.es

Ramón Ferri-García
rferri@ugr.es

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

Published online: 02 March 2022

1 Introduction

In recent years, online surveys have undergone rapid development in a wide variety of fields, including public opinion research (Couper 2000) and life sciences (Thornton et al. 2016; Borodovsky et al. 2018). In contrast to traditional survey modes, which are experiencing issues with response rates (according to Marken (2018), response rates in Gallup Poll Social Series dropped from 28% in 1997 to 7% in 2017) and increasing costs, online surveys offer a faster and cheaper method to measure certain features in individuals. In addition, there is an increasing availability of large sets of data obtained from the Web with automatic procedures (such as web scraping or APIs) that are often used for inference in finite populations.

Non-probabilistic surveys provide us with some advantages over traditional methods but also these surveys have given us many research problems: many problems demand the type of velocity for both data processing and analysis, but the most important is related to the quality of the data. Data quality is far more important than data quantity. Meng (2018) studies some theoretical aspects related to the impact of the application of non-probability online surveys on the estimation quality, and develops a data defect index, a main topic of population inferences from Big Data and concludes it is more important to reduce sampling and non-response biases than non-response rates.

Indeed, non-probabilistic surveys emphasise certain types of nonsampling errors. It is not feasible to obtain a representative sampling frame of the online population except in specific situations where the target population is a well-characterised group (such as company employees or university students each of whom is associated with an e-mail address). For this reason, most online surveys or large volume datasets are based on volunteer samples. In addition, the coverage of this approach is limited by the extent of Internet penetration among the population, which is often subject to demographic characteristics. For instance, according to the Spanish Survey on Equipment and Use of Information and Communication Technologies in Households (National Institute of Statistics of Spain 2018), while 98.5% of the Spanish population aged 16–24 years make regular use of the Internet, only 49.1% of those aged 65–74 years do so. Although the difference has narrowed in the last few years, online surveys are still unable to provide representative samples except when special procedures are used, such as offline recruitment, panels or mixed modes (see Schonlau and Couper 2017 for a review of the available options).

The lack of a probability sampling scheme might lead to significant differences between sampled and nonsampled individuals, which constitutes a selection bias that cannot be redressed with the usual procedures (Elliott and Valliant 2017). Selection bias is a particularly important concern in online surveys because of their intrinsic characteristics (Couper 2000). Statistical adjustments are crucial to obtaining reliable estimates from online survey data; in this context, calibration or Propensity Score Adjustment (PSA) can be used, according to the kind of auxiliary information available. While calibration only needs the vector of population totals for some auxiliary covariates, PSA requires a probability sample drawn from the same target population, even when the nonprobability sample is drawn from a subset of it, which is the case of Internet surveys (not everybody may have access to the Internet in a given population) and imperfect sampling frames in general. This sample is used to estimate the (unknown) participation propensities for the individuals in the nonprobability sample through prediction models. These estimated propensities can be used as inclusion probabilities to build weights for different parametric estimators.

The efficacy of PSA at removing selection bias has been proved, although some considerations should be taken into account. First, PSA is strongly dependent on the covariates used to estimate the propensities. Lee (2006) showed that the use of covariates which are strongly related to the variables of interest in PSA models achieves greater reductions in bias than is the case with nonsignificant variables. Second, further adjustments such as calibration procedures must be applied in order to maximise the effectiveness of PSA (Lee and Valliant 2009; Valliant and Dever 2011; Valliant 2020). Finally, the use of PSA is associated with an increase in the variance of the estimates.

In this study, we focus on the first point raised above: the choice of covariates. Lee (2006) suggested that including all available covariates, as recommended by Rubin and Thomas (1996), might be a reasonable practice. However, statistical models based on modern classification techniques such as Machine Learning algorithms might benefit from feature selection to reduce the complexity of the models (and the variance of their predictions). Variable inclusion in propensity models for treatment weighting has been widely studied (Hirano and Imbens 2001; Brookhart et al. 2006; Austin 2008; Schneeweiss et al. 2009; Austin 2011; Myers et al. 2011; Patrick et al. 2011; Austin and Stuart 2015) and variables are often selected using a stepwise algorithm or they are assessed prior to the study according to their known relationship to the outcome or exposure variables. In this case, better results are obtained when the variables in question are related to the outcome variables or to both the outcome and the exposure variables.

In many real-world applications, there may be very little information about the pre-existing relationships between variables, which increases the difficulty of selecting the best subset of variables for propensity estimation. In the present study, we consider how modern techniques of feature selection (or variable selection) developed for knowledge discovery in data can be used in propensity estimation modelling. These techniques only require an appropriate dataset from which to locate the variables more closely related to a given target variable or that may be more influential with respect to predicted values, according to the behaviour observed in the dataset. The benefits of feature selection, in terms of increased accuracy and reduced computational costs, have been demonstrated in classification tasks (Bolón-Canedo et al. 2013; Xue et al. 2015).

In survey research, feature selection has been studied with respect to the problem of calibration when a large number of variables must be considered. Breidt and Opsomer (2017) reviewed this question and suggested that auxiliary variables for calibration may be too closely correlated or have poor predictive power, and therefore model selection should be employed to improve the estimates obtained and to stabilise the weights. Stepwise and best subsets algorithms have been considered for this purpose, but models from the class of “least absolute shrinkage and selection operator” (LASSO), which perform feature selection by shrinking regression coefficients to zero in non-informative variables, seem to be the most promising methods to improve the weighting. Their efficiency in non-probability samples was highlighted by Chen et al. (2019), who showed that LASSO-weighted estimators have a lower RMSE than PSA-weighted equivalents.

The rest of this paper is organised as follows: Sect. 2 presents the essential aspects of calibration and PSA. The synthetic data and the real survey datasets used in our experiments are then described in Sect. 3. In Sect. 4 we describe the deployment of PSA models with a grid of classifiers and feature selection algorithms for the study data. The results of the experiments in terms of relative bias and efficiency are detailed in Sect. 5, after which the method proposed is applied in a real-world context concerning addiction and dependence, in Sect. 6. Finally, the implications of our findings are discussed in Sect. 7.

2 Adjustments for nonprobability samples

2.1 Calibration

Calibration was developed by Deville and Särndal (1992) as a reweighting method based on the availability of population totals for auxiliary variables measured in a sample, although some later versions addressed missing data situations or the use of dual frames for survey sampling (Ranalli et al. 2016). This adjustment is intended to reduce the coverage error between the target population and the sample, and takes the following form. Let $\mathbf{x}$ be a $n \times p$ matrix of p variables measured in a sample of size n, x_{ij} is the value of the i-th individual in the j-th auxiliary variable, $\mathbf{X} = (X_1, ..., X_j, ..., X_p)$ are the known population totals for the auxiliary variables and $d = (d_1, ..., d_i, ..., d_n)$ is the vector of design weights of the sample. If a probabilistic unbiased sample from the same population is available, estimated population totals can be used for $\mathbf{X}$ as an alternative (see Ferri-García and Rueda 2018 for a study of its efficiency). Calibration then attempts to obtain a new vector of weights $w = w_1, ..., w_i, ..., w_n$ by minimising their distance frp, d (from a class of distances leading to different estimators) subject to the calibration equations:

$$\sum_{k=1}^{n} w_k x_{kj} = X_j,\ j = 1, ..., p \tag{1}$$

When information on population totals is incomplete, and especially when the cross-classification totals (also known as cell counts) are not known, it can be useful to use the raking ratio as defined in Deville et al. (1993), which takes advantage of the estimation of cell counts from the available data in the sample. Here, let $\hat{N}_{ab} = \sum_{k/x_{Ak}=a, x_{Bk}=b} d_k$ be the estimated cell count of ab, which represents the number of individuals whose measured value in the variables A and B is a and b respectively. The raking ratio uses this information to reformulate the calibration equations, thus obtaining the calibrated weights $w_k = d_k \hat{N}^w_{ab} / \hat{N}_{ab}$, where $\hat{N}^w_{ab} = d_k \hat{N}_{ab}$ represents the calibrated estimations of the cell counts. The efficiency of calibration procedures depends on the relevance of the auxiliary information in terms of relationship with the target variable and on the mechanism producing the coverage error. Calibration has also been found to be effective for removing selection bias when the target variable is not related to the selection mechanism (Bethlehem 2010; Rueda 2019).

2.2 Propensity Score Adjustment

Propensity Score Adjustment (PSA) was originally developed by Rosenbaum and Rubin (1983) as a technique for balancing comparison groups in nonrandomised studies, where the inclusion in one group or another might be driven by or associated with variables not controlled by the researchers. PSA was subsequently adapted to the context of online surveys (Taylor 2000; Taylor et al. 2001; Lee 2006; Castro-Martín et al. 2020a) as a means of reducing selection bias when a reference probability sample collected from the same target population is available. In this case, let s_r be the reference sample, s_v the nonprobability sample obtained from the online survey and $s = s_r \cup s_v$. Furthermore, let R be a binary variable measured for U where $R_i = 1$ if $i \in s_v$ and $R_i = 0$ otherwise. PSA assumes that the inclusion probability or propensity score, π, for s_v is conditional on a set of covariates, $\mathbf{x}$, such that:

$$\pi_i = P(R_i = 1|\mathbf{x}_i), \quad i \in U \tag{2}$$

The inclusion probability can therefore be modelled through a proxy of R. Let z be a binary variable measured for s which $z_i = 1$ if $i \in s_v$ and $z_i = 0$ if $i \in s_r$. The propensity score is then estimated by predicting the values of z using a model M:

$$\hat{\pi}_i^* = E_M[z = 1|\mathbf{x}_i], \quad i \in s_v \cup s_r \tag{3}$$

Note that in this case we are not estimating π but π^*, which is the propensity obtained when we predict the measured participation z rather than the true participation R.

The propensity scores are used to reweight the nonprobability sample. In this process, inverse probability weighting formulas can be used, such as the simple inverse probability $w^{PSAIPW1} = 1/\pi$ (Valliant 2020) or the inverse probability allowing weights to be less than one, as proposed by Schonlau and Couper (2017): $w^{PSAIPW2} = (1-\pi)/\pi$. Propensities can also be transformed into weights using the subclassification methods proposed by Lee (2006) and Lee and Valliant (2009). This technique stratifies the vector of propensities into c parts (following Cochran (1968), c is usually taken as 5) with similar propensities, applying the formula:

$$w_i^{PSAsub1} = f_c d_i^v = \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_i^v / \sum_{j \in s_v} d_i^v} d_i^v \tag{4}$$

where d^r, d^v represent the design weights for the reference and volunteer samples respectively and s_r^c, s_v^c are the individuals belonging to the c-th strata of propensities in the reference and volunteer samples respectively. Valliant and Dever (2011) proposed a similar method, but instead of calculating a correction factor,the propensities in each stratum were averaged and then transformed into weights by inverse probability weighting, as follows:

$$w_i^{PSAsub2} = \frac{1}{(\hat{\pi}_g^*)} \tag{5}$$

3 Data

3.1 Artificial data

An experiment with artificial data was performed to evaluate the benefits of feature selection under different conditions. In this experiment, a population U of size $N = 500,000$ was generated with 17 variables: eight variables $\mathbf{x} = (x_1, ..., x_8)$ were used as covariates for PSA algorithms, out of which variables x_1, x_3, x_5 and x_7 were used as calibration variables. Another eight variables $\mathbf{y} = (y_1, ..., y_8)$ were considered as target variables and a variable π measured the probability of each individual of the population being selected in the nonprobability sample.

The covariates were generated as described in Eq. 6. Four variables (x_1, x_3, x_5, x_7) followed a Bernoulli distribution with $p = 0.5$ and the other four (x_2, x_4, x_6, x_8) followed Normal distributions with a standard deviation of one and a mean parameter dependent on the value of the previous Bernoulli variable for each individual; for instance, if the i-th individual had a value of 1 in x_1, then its value for x_2 was simulated according to a $N(2, 1)$ distribution, and if it had a value of 0, then it was simulated according to a $N(0, 1)$ distribution. This procedure induced a collinearity in the models if all of the covariates were used, an issue that could be addressed by variable selection algorithms.

$$\begin{aligned} x_{1i}, x_{3i}, x_{5i}, x_{7i} &\sim Be(0.5) \qquad\qquad i \in U \\ x_{ji} &\sim N(\mu_{ji}, 1) \qquad\quad i \in U, j = 2, 4, 6, 8 \\ \mu_{ji} &= \begin{cases} 2, \text{ if } x_{(j-1)i} = 1 \\ 0, \text{ if } x_{(j-1)i} = 0 \end{cases} \quad i \in U, j = 2, 4, 6, 8 \end{aligned} \tag{6}$$

The inclusion probability π was made dependent on x_5, x_6, x_7 and x_8 as described in Eq. 7, which allowed the experiment to cover Missing At Random (MAR) situations.

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = -0.5 + 2.5(x_{5i} = 1) + \sqrt{2\pi} x_{6i} x_{8i} - 2.5(x_{7i} = 1), \quad i \in U \tag{7}$$

The target variables were simulated as described in Eqs. 8 to 15. Four types of relationship were considered: no relationship at all with any other variable (y_1 and y_2), a relationship with the selection mechanism (y_3 and y_4), a relationship with some covariates related to the selection mechanism (y_5 and y_6) and a relationship both with the selection mechanism and with some covariates (y_7 and y_8).

$$y_1 \sim Be(0.5) \tag{8}$$

$$y_2 \sim N(10, 1) \tag{9}$$

$$y_{3i} \sim Be\left(\frac{exp(\pi_i)}{1 + exp(\pi_i)}\right), \quad i \in U \tag{10}$$

$$y_{4i} \sim N(10, 1) + 5\pi_i, \quad i \in U \tag{11}$$

$$y_{5i} \sim Be\left(\frac{exp(0.5 + 0.25(x_{5i} = 1) - 0.25(x_{5i} = 0) + x_{6i})}{1 + exp(0.5 + 0.25(x_{5i} = 1) - 0.25(x_{5i} = 0) + x_{6i})}\right), \quad i \in U \tag{12}$$

$$y_{6i} \sim N(10, 1) + 2(x_{5i} = 1) - 2(x_{5i} = 0) + x_{6i}, \quad i \in U \tag{13}$$

$$y_{7i} \sim Be\left(\frac{exp(0.5 + 0.25(x_{7i} = 1) - 0.25(x_{7i} = 0) + x_{8i} + \pi_i)}{1 + exp(0.5 + 0.25(x_{7i} = 1) - 0.25(x_{7i} = 0) + x_{8i} + \pi_i)}\right), \quad i \in U \tag{14}$$

$$y_{8i} \sim N(10, 1) + 2(x_{7i} = 1) - 2(x_{7i} = 0) + x_{8i} + 5\pi_i, \quad i \in U \tag{15}$$

This procedure allowed the target variables to reflect all of the missing data mechanisms; y_1 and y_2 are examples of Missing Completely At Random (MCAR) data, where the outcome is not related to the selection. y_5 and y_6 are examples of Missing At Random (MAR) data, where the outcome is indirectly related to the selection through some variables. Finally, y_3, y_4, y_7 and y_8 are examples of Missing Not At Random (MNAR) data, where the outcome is directly related to the selection mechanism.

3.2 Real data

The experiment was then repeated using a real dataset as a pseudopopulation to examine whether variable selection algorithms might be helpful when more complex relationships are present in the data. The dataset was obtained by the January 2019 Barometer Survey (study number 3238) conducted by the Spanish Centre for Sociological Research (CIS, Spanish initials), a monthly survey that measures political and social opinions among the Spanish adult population (Spanish Center for Sociological Research (2019)). The original dataset of the survey sample made available by the CIS included $n = 2989$ individuals and $p = 203$ variables, out of which 17 variables were finally selected:

- 6 target variables: assessment of the current economic situation in Spain and in their own lives (binary, 1 if "bad" or "very bad", 0 otherwise), score on the ideological self-positioning scale (numeric, 1–10), assessment of the central government's performance (binary, 1 if "Poor" or "Very poor", 0 otherwise), territorial organisation preference (binary, 1 if "State with no autonomous structures", 0 otherwise) and national sentiment (binary, 1 if "Self identification as only Spanish", 0 otherwise).
- 10 variables to be used as covariates in PSA or calibration variables: frequency of attendance at religious acts, gender, age, education level, socioeconomic status, autonomous community of residence, size of the municipality of residence, nationality, marital status and degree to which voting is expected to change things.

Gender, age and size of the municipality were chosen as calibration variables in each simulation run, and were also included as potential covariates for PSA.
- One variable, use of internet in the three months prior to the survey (1 if it was used, 0 otherwise), was taken as a delimiter of the population subset from which nonprobability samples would be drawn. Individuals with a value of 1, but not those with a value of 0, in this variable could belong to the nonprobability sample. The rationale for this delimiter is that it reproduces the conditions that apply in real online surveys, in which people with no internet access cannot be selected to participate.

The pseudopopulation was obtained by bootstrapping the original sample up to $N = 500,000$ individuals through simple random sampling with replacement. Out of the 500,000 individuals, 404,174 (80.83% of the pseudopopulation) had used the internet in the three months prior to the survey. Despite internet's large penentration, the differences between the population with and without access to the internet are noticeable in several target variables, which leads to a remarkable amount of coverage bias when estimating population parameters using only people who had accessed the internet. This coverage bias can be treated using calibration in addition to PSA. Those differences can be observed in Table 1.

The pseudopopulation was obtained by bootstrapping the original sample up to $N = 500,000$ individuals through simple random sampling with replacement. Prior to the bootstrapping, anyone who did not answer ("Does not know"/"Does not answer") any of the 17 items was excluded, as were the persons who answered "Other" for education level, or who gave "Ceuta" or "Melilla" as their autonomous community of residence. The reason for this filtering process was to remove highly uncommon classes that could produce inconsistencies in a simulated sample and provoke errors in the propensity scoring algorithms. Moreover, the education levels "No formal education" and "Primary education" were collapsed into a single class, while missing data in the variable concerning attendance at religious acts was taken as a new class (given that everyone in this group was considered to be atheist or agnostic). After the preprocessing, the sample size before bootstrapping was $n = 2,156$.

4 Methods

4.1 Feature selection algorithms

Feature selection was performed prior to PSA to select those variables that are more relevant for the prediction of a target variable (which must be present in the dataset). In a model-based framework, we only have one variable of interest, y, for which we both predict its values and estimate its population parameters. However, in the design-based framework of PSA, we have two variables that must be considered: the indicator variable z ($z_i = 1$ if $i \in s_v$, $z_i = 0$ otherwise), for which we predict the probability $P(z = 1)$, and the target variable of the study, y, and its population parameters of interest that we want to estimate.

Table 1 Population values of the variables of interest in the real data simulation. All numbers are population proportions of the features of interest described in the "Target variable" column, except for "Ideological self-positioning scale (1–10)" where the numbers correspond to the population mean

Target variable	Population values			Difference between complete and Internet pops.	
	Complete	Internet	Non-internet	Absolute	Relative
Econ. situation in Spain "poor" or "very poor"	14.4%	13.7%	17.5%	0.7%	5.2%
Personal econ. situation "poor" or "very poor"	3.7%	3.3%	5.3%	0.4%	11.6%
Ideological self-positioning scale (1–10)	4.59	4.48	5.04	0.11	2.4%
Central gov. management "poor" or "very poor"	14.3%	14.6%	13.1%	0.3%	−1.9%
Preference for a state without autonomous comm.	17%	15.4%	23.9%	1.6%	10.7%
Feels only Spanish	13.8%	12.3%	19.9%	1.5%	11.9%

Internet population: population who accessed the internet in the three months prior to the survey. Non-internet population: population who did not accessed the internet in the three months prior to the survey. Absolute difference: |Complete - Internet|. Relative difference: Complete/Internet - 1

Given that the predictive model is applied on z in PSA, it would be fair to assume that the relevant variables for prediction should be selected considering z as the target variable in the feature selection algorithms. However, given the previous research on PSA for experimental designs and the fact that the bias must be removed for y (and not necessarily for z), it could also be reasonable to select those variables which are more relevant to y, and therefore to consider y as the target variable in the feature selection algorithms. In this work, we have considered both scenarios: feature selection for the prediction of z, and feature selection for the prediction of y. In the former case, the combination of both samples $s_v \cup s_r$ can be used, while in the latter case only s_v can be used for feature selection, as y has only been measured for individuals in s_v.

The following feature selection algorithms were used in the experiment, and their performance was compared to the use of all variables and to the use of the variables provided by Stepwise:

- CFS (Correlation-based Feature Selection) filter with best first search. This algorithm, proposed by Hall (1999), searches the subset of variables which maximises the correlation with the target variable and minimises that between the variables of the subset. Thus, irrelevant and redundant features are discarded from the optimal subset of features for prediction. Note that Pearson's correlation is used to evaluate the relationships between the variables; if any of the variables within a pair is non-numeric, it is binarised and each of the binary variables is then used separately. The simplicity of the algorithm makes it a fast and intuitive choice for finding the optimal subset of relevant covariates, while at the same time addressing the multicolinearity problems that may arise from focusing only in the correlation between the target variable and the covariates. On the other hand, the use of Pearson's correlation coefficient implies that nonlinear relationships or categorical variables with high cardinality might be erroneously discarded by the algorithm.
- Chi-square filter. This approach calculates the Cramer's V value between the target variable and each independent variable, and so the user must define a cut-off point for selection. In our experiment, the cut-off point was the Cramer's V value with the biggest difference from the V of the next variable in importance (ordered from highest to lowest). This filter performs better at finding relationships between categorical variables rather than continuous ones, given that Cramer's V depends on the number of classes of each pair of variables, and therefore the coefficient might be considerably more sensitive towards continuous variables.
- Gain ratio. This entropy-based filter (Quinlan 1986) is calculated by dividing the information gain by the entropy of the target variable. The information gain is measured as the difference between the sum of the entropies of the independent and the target variables and the entropy of the target variable after introducing the independent variable into the predictive model (defined as a decision tree). The gain ratio, thus, is a relative continuous measure of the predictive performance of a variable. The cut-off point was the gain ratio's value with the biggest difference from the gain ratio of the next variable in importance (ordered from highest to lowest). Gain ratio, as the rest of entropy-based filters, is assumed to identify nonlinear relationships more precisely than correlation coefficients, but require

discretization of continuous variables in order to seize all its potential (Yu and Liu 2003).

- One-R. This algorithm, developed by Holte (1993), is based on very simple rules of association, by which each independent variable is tabulated with the target variable. The number of errors is then determined and interpreted such that higher values represent a stronger predictive power. OneR automatically divides continuous variables into categories using discretization functions, hence it is a more suitable filter in datasets with categorical and continuous variables. However, the algorithm is prone to overfitting if the discretization aims to obtain "pure" classes where all individuals take the same values.
- Random Forest importance filter. This algorithm computes the mean importance value across the trees created in a Random Forest model (Breiman 2001) for each independent variable. In our experiment, the importance value taken was the mean decrease in accuracy when the variable was discarded from the Random Forest model. The cut-off point was the importance value with the biggest difference from the importance value of the next variable (ordered from highest to lowest according to their importance value). This algorithm is suitable for any kind of target variable and covariates, and its bagging configuration can be advantageous in more complex situations. In addition, the use of the mean decrease in accuracy instead of node impurity (measured with the Gini index) avoids the overestimation of the importance value of continuous attributes. However, the method is still sensitive to multicolinearity problems (Nicodemus et al. 2010).
- Boruta algorithm. This algorithm is based on the Random Forest importance measure, but it considers a set of non-informative variables created from the random shuffling of each independent variable included in the model. As a result, the algorithm selects the variables that have greater importance than non-informative variables. To obtain statistically valid results, the procedure is repeated until every variable has been deemed as "important" or "unimportant". Further details on this algorithm can be consulted in Kursa and Rudnicki (2010). This algorithm can be considered an improved version of the Random Forest importance filter and has the advantage of automatically selecting the relevant variables and discarding the irrelevant ones. However, the computational costs of the algorithm are large.
- LASSO regression (Tibshirani 1996). This regression model performs a variable selection based on introducing penalisation terms into the Ordinary Least Squares equations. As a result, a regression model is provided but only the variables selected have non-zero coefficients. In the present study, we take advantage of the LASSO variable selection technique by extracting the variables with non-zero coefficients and using them as inputs for the propensity estimation models. When all the coefficients of the LASSO model are zero, no PSA is performed and therefore the weights remain unitary. The LASSO algorithm has provided better results than other predictive methods in nonprobability sampling contexts (Chen et al. 2019; Castro-Martín et al. 2020a), but the variable selection is performed considering a specific model and optimization criteria. The advantages on the use of subsets of relevant variables according to LASSO as input variables for other predictive algorithms are unclear.

4.2 Estimation with Propensity Score Adjustment and calibration

Once the optimal subset of variables had been selected, the Propensity Score Adjustment (PSA) was performed. As well as logistic regression (LR), the standard algorithm in PSA, several other algorithms were also tested for propensity estimation, namely: k-Nearest Neighbours (kNN), Gradient Boosting Machine (GBM) and feed-forward neural networks (NN). Parameter tuning was performed for these three algorithms. Ten-fold cross-validation was applied to the model, predicting z prior to PSA; the following parameter grids were used for each algorithm:

- k-Nearest Neighbours (kNN): $k = 5, 7, 9$.
- Gradient Boosting Machine (GBM): number of trees = 50, 100, 150, learning rate $= 0.1$, interaction depth $= 1, 2, 3$.
- Feed-forward neural networks (NN): number of units in the hidden layer = 1, 3, 5, weight decay $= 0.1, 0.0001, 0$.

The choice of the kNN and GBM algorithms is based on their performance in the previous study developed in Ferri-García and Rueda (2020), where they showed a better performance than logistic regression in some situations in terms of bias and MSE. The use of neural networks is based on providing more diversity in the approaches, and the possible predictive advantage than neural networks may provide in modeling (Breidt and Opsomer 2017). k-Nearest Neighbors is a simple algorithm that can provide good results when the number of covariates is low, while its performance decreases in contexts of high dimensionality. For this reason, variable selection might be highly recommendable to boost kNN performance. Gradient Boosting Machines are better in those high-dimensionality situations because of their boosting algorithm that is able to internally select the best predictors.

4.3 Experiment settings

In both scenarios, the same procedure was followed to measure the effects of variable selection in PSA and calibration on the estimation from nonprobability samples. This procedure, repeated across 400 simulation runs for each dataset (artificial and real), can be sequentially described as follows:

1. Two samples of size n = 1,000 are drawn. The first one, s_r, is the probability sample and is drawn by simple random sampling without replacement (SRSWOR) from the full population. The second sample, s_v, is the nonprobability sample and is drawn according to the following schemes:
   - Artificial dataset: unequal probability sampling where π is the vector of inclusion probabilities, calculated as described in Equation 7.
   - Real dataset (two schemes):
     - SRSWOR from the subset of the population who had accessed the internet during the three months prior to the survey.
     - Unequal probability sampling from the subset of the population who had accessed the internet during the three months prior to the survey, with

inclusion probabilities proportional to the age:

$$\pi_i = \frac{(200 - \text{Age}_i)^5}{(200-10)^5}, i \in U_I \tag{16}$$

where U_I is the subset of the pseudopopulation who used the Internet in the three months prior to the survey.

2. Propensity of belonging to s_v is estimated with PSA, using the variable selection algorithms described in Sect. 4.1 to select the input covariates for propensity prediction models, and the four choices of algorithms described in Sect. 4.2 to model propensities. We also consider the choice where no variable selection algorithm is applied and all covariates are included in the models.
3. Estimated propensities are transformed into weights using the inverse probability weighting formula $w_i = 1/\pi_i$.
4. Weights are used to estimate the population mean of each target variable with and without applying Raking calibration, on which the propensity weights w obtained in step 3 are used as initial weights.

The resulting 400 estimates of the population mean for each combination of methods are subsequently used to obtain the relative bias of a given combination of methods:

$$RB(\%) = \left| \frac{\sum_{i=1}^{400} \frac{\hat{\overline{y_i}}}{400} - \overline{Y}}{\overline{Y}} \right| \cdot 100 \tag{17}$$

where $\overline{Y}$ is the population mean of the target variable, and $\hat{\overline{y_i}}$ is the estimate of the population mean of the i-th simulation obtained after applying bias reduction methods. Together with the relative bias, the efficiency of each variable selection method with respect to the case in which all variables are used is also shown, given a propensity model m (Log. reg., GBM, kNN or NN), a Raking calibration choice (yes or no) r, and a choice for the target variable (exposure or outcome) in selection algorithms v:

$$\text{Effect}_{k|m,r,v} = \frac{MSE_{k,m,r,v}}{MSE_{\text{All vars.},m,r,v}} \tag{18}$$

where $k = \{$Boruta, CFS, Chi-squared, Gain ratio, LASSO, StepWise, OneR, Random Forest importance$\}$ is the variable selection algorithm and MSE is the Mean Squared Error observed for the combination of methods:

$$MSE = \text{Bias}^2 + \text{Variance} = \left(\frac{\sum_{i=1}^{400} \hat{\overline{y}}_i}{400} - \overline{Y} \right)^2 + \frac{\sum_{i=1}^{400} \left(\hat{\overline{y}}_i - \frac{\sum_{i=1}^{400} \hat{\overline{y}}_i}{400} \right)^2}{399} \tag{19}$$

An effect greater than 1 means that the use of the variable selection method k is inefficient in comparison with using all covariates, while if it remains below 1 the selector k provides more efficient estimates, provided all other adjustments remain equal. This "Effect" can be seen as the MSE of a certain variable selection method in relation to the reference case in which all variables are used.

The statistical significance of each effect was tested using bootstrapping techniques. Basic resampling (with 1000 replications) was performed to obtain each effect number, so the standard deviation of the effect could be estimated from the bootstrapped samples. The standard deviation was used to perform t-tests on the following null hypothesis:

$$\begin{aligned} H0 &: \text{Effect}_{k|m,r,v} \geq 1 \\ H1 &: \text{Effect}_{k|m,r,v} < 1 \end{aligned} \tag{20}$$

The t-tests used the standard deviation calculated from the bootstrap procedure, and the confidence level was fixed at 95% for all effects. Rejection of the null hypothesis would mean that there are statistical evidences that the variable selection method k provides more efficient estimates given that the rest of conditions remain unchanged. Resampling was performed using the *resample* package available in R (Hesterberg 2015).

Finally, for each feature selection algorithm and Raking choice (Raking used or not used after PSA), we computed the estimated mean and median relative bias and effect. For relative bias, we also computed the number of times that the estimates provided by a feature selection algorithm have been among the best (Relative Bias less than 1% greater than the minimum) conditional to a given variable of interest, target variable in the feature selection algorith, PSA predictive algorithm and Raking strategy. For the effect, we computed the number (and percentage) of times that the effect has been below 1 (the MSE after applying a given feature selection algorithm was lower than the MSE using all variables) and below 0.9 (the MSE after applying a given feature selection algorithm was more than 10% lower than the MSE using all variables).

5 Results

5.1 Artificial data

The relative bias results obtained in the simulation with artificial data are shown in Tables 8 and 9. For the MCAR variable y_1, variable selection was useful when neural nets were used as the predictive model and Raking was applied after PSA, although the improvements were not dramatic. The least biased estimates were provided by PSA with kNN using all variables in y_1 and variables selected by OneR (selecting on the variable of interest) with neural nets and no Raking in y_2, although this result was closely followed by the Gain ratio score in the latter case. However, the differences are too small to be considered relevant.

With the MAR variables (y_5 and y_6), Raking calibration markedly reduced the bias in the estimates. Regarding variable selection, almost all the methods in y_5 and some of them in y_6 reduced the bias when the predictive model was logistic regression, although some reductions were also observed when other methods were applied in different models. In the case of y_6, the chi-square filter, the Gain Ratio and Random Forest all reduced the bias from 2.88 (when using all available covariates) to 2.01 in y_2 if logistic regression and Raking calibration were applied.

Finally, in NMAR situations (y_3, y_4, y_7 and y_8), the application of Raking calibration also reduced bias but not as much as for MAR variables. For y_3 and y_8, the best choice for the target variable in the selection algorithms was the variable of interest (y), while fixing the target variable in the indicator variable of inclusion in s_v (z) provided better results in y_4. The largest reductions in bias in y_3 were obtained with the LASSO algorithm, although CFS, Chi-square and the Gain Ratio also worked well when combined with Raking.

The effect of each variable selection method in comparison to using all variables, if the rest of methods remain equal, is detailed in Tables 10 and 11. These results are in line with those for relative bias in each case, although they reflect some improvement in a much larger set of situations. With the MCAR variables (y_1 and y_2), MSE could be slightly reduced using variable selection in some cases, but in general the effect improvements were small. Reductions were always below 10% of the MSE using all variables, and only 7.03% of the efficiencies were below 0.95 (this is, reductions above 5% of the baseline MSE), all of them achieved selecting variables with regard to the variable of interest.

Regarding the MAR variables (y_5 and y_6), very noticeable improvements in efficiency were obtained in the estimation of y_6. When Raking calibration was applied, the use of the Chi-square filter, Gain Ratio or Random Forest reduced MSE by 11% (if k-NN was used in PSA) to 50% (if LR was used in PSA), when they selected variables using the variable of interest as the target. Reductions in MSE with the same variable selection methods were also observed when Raking was not applied. Other methods, too, provided larger effect values when y_5 was estimated for the cases in which logistic regression was used to estimate the propensities. It is worth noting that the hypothesis of greater efficiency of selection methods was accepted in several cases, which gives some evidence that variable selection methods can work in practice.

Finally, regarding the NMAR variables, the reductions in MSE were noticeable: 12.5%, 11.7% and 24.2% of the times the effect was below 0.9 (improvements in the MSE above 10%) in the estimation of y_3, y_7 and y_8 respectively. In the estimation of y_3, all the combinations of methods, except for those involving Boruta selection algorithm, provided efficiencies significantly lower than 1 (with a mean effect of 0.922) when selecting variables using the variable of interest as the target in feature selection algorithms. The opposite situation was observed in y_4: selecting variables using the indicator variable of inclusion in s_v as the target provided better results, although the improvements were considerably lower despite many of them being statistically significant. In y_7, feature selection algorithms provided more efficient estimates mainly in those cases where logistic regression was used, although OneR (selecting on the variable of interest) provided gains when using other algorithms for

PSA in the cases where Raking was not applied. For y_8, there are statistical evidences that all of the feature selection algorithms provide more efficient estimates (except for StepWise and LASSO) when the selection is done using the variable of interest as the target, regardless of the classifier used in PSA or the further use of Raking calibration.

A summary of the relative bias and effect results observed in the artificial data simulation can be found in Table 2. When Raking calibration is not applied, results on relative bias are very similar across variable selection strategies, although the most featured method among the best ones is the strategy of using all variables. When Raking calibration is applied, the situation slightly changes, with the gain ratio score appearing the most in the set of algorithms that provide the best results, followed by the OneR algorithm. These results can be explained by the fact that each Bernoulli covariate is related to a Gaussian covariate, but the different Bernoulli and Gaussian covariates are independent of each other. This scenario can be more suitable for simple algorithms that test one variable at a time, such as OneR, because it is only necessary to retain one of the two variables that compose each Bernoulli-Gaussian relationship, while in more complex scenarios the choice would not be that clear. Regarding the effect, none of the variable selection algorithms seem to outperform the case where all variables are used when Raking calibration has not been applied. If it has been applied, there is a certain evidence towards the effectiveness of variable selection, with all of the algorithms except for StepWise and Boruta achieving lower MSE values than the case where all variables are used more than a half of the times they are included in the preprocessing step.

5.2 Real data

The relative bias results obtained by each combination of methods in the simulation using CIS data and considering SRSWOR from the Internet population to obtain the nonprobability samples are listed in Table 12. Interestingly, the best choice in variable selection differed according to the propensity estimation model considered. For example, PSA using k-NN provided the best results when using all the available covariates, except for the variable measuring central government performance. In the remaining cases, the use of certain variable selection algorithms was associated with a decrease in relative bias. This was especially apparent for the variables measuring the economic situation in Spain, central government management and the preference for a unitary national state without autonomous communities. In these cases, the largest reductions in relative bias (compared to the case in which all variables were used) were obtained when the variable selection algorithms used the variable of interest as the target variable. Raking calibration had a modest positive effect on the variables measuring the ideological self-positioning scale, the preference for a unitary national state without autonomous communities and whether the respondent self identified as only Spanish, while its impact on relative bias in the other variables was non-relevant or negative. The efficiency of each variable selection algorithm for a given combination of adjustments (propensity model, use of calibration and target variable choice for selection), in comparison with the case in which all variables are used, is shown in Table 13. For all variables, one or more selection algorithms increased the efficiency, in comparison

Table 2 Estimated mean and median of RB and Effect of estimates using PSA for each algorithm, and number of times its estimates have been among the best (RB less than 1% greater than the minimum) or have been more efficient than using all variables (Effect under 1 and under 0.9) in the artificial data simulation

Use of	Algorithm	RB (%)			Effect			
raking		Mean	Median	Best	Mean	Median	Eff. < 1	Eff. < 0.9
No Raking	All vars.	10.097	11.819	27 (42.2%)				
	Boruta	10.106	11.924	14 (21.9%)	1.001	1.000	31 (48.4%)	0 (0%)
	CFS	10.166	11.825	13 (20.3%)	1.011	1.003	29 (45.3%)	2 (3.1%)
	Chi-sq.	10.159	11.902	6 (9.4%)	1.009	1.006	28 (43.8%)	5 (7.8%)
	Gain r.	10.083	11.803	21 (32.8%)	0.998	0.999	33 (51.6%)	5 (7.8%)
	LASSO	10.255	11.911	7 (10.9%)	1.016	1.006	24 (37.5%)	5 (7.8%)
	OneR	10.208	11.704	21 (32.8%)	1.014	0.998	33 (51.6%)	4 (6.2%)
	RF imp.	10.160	11.913	9 (14.1%)	1.009	1.000	30 (46.9%)	5 (7.8%)
	StepWise	10.193	11.947	14 (21.9%)	1.013	1.000	31 (48.4%)	0 (0%)
Raking	All vars.	5.945	7.014	18 (28.1%)				
	Boruta	5.974	7.068	10 (15.6%)	1.002	1.002	27 (42.2%)	2 (3.1%)
	CFS	5.929	7.212	9 (14.1%)	1.040	0.996	33 (51.6%)	7 (10.9%)
	Chi-sq.	5.889	7.286	16 (25%)	0.987	0.976	38 (59.4%)	10 (15.6%)
	Gain r.	5.826	7.085	25 (39.1%)	0.978	0.977	37 (57.8%)	12 (18.8%)
	LASSO	6.018	7.227	13 (20.3%)	1.040	0.996	33 (51.6%)	5 (7.8%)
	OneR	6.005	7.566	23 (35.9%)	1.054	0.985	37 (57.8%)	12 (18.8%)
	RF imp.	5.992	7.295	9 (14.1%)	1.012	0.994	34 (53.1%)	8 (12.5%)
	StepWise	6.009	7.137	11 (17.2%)	1.027	1.003	29 (45.3%)	3 (4.7%)

with the case in which all variables were used. The estimated mean and median effect of all the possible combinations of methods shown in Table 13 is below 1 for all the target variables except for the ideological self-positioning scale. However, the numbers are strongly dependent on the target fixed in the feature selection algorithm. For instance, the percentage of combinations with an effect below 0.9 (decrease of more than 10% in the MSE in comparison to the case where all covariates are used) when fixing the variable measuring the inclusion in $s_v(z)$ as the target is 0% when estimating the economic situation in Spain, central government management and the preference for a unitary national state without autonomous communities, but if the variable of interest is fixed as the target in feature selection algorithms, the percentage rises up to 12.5%, 29.7% and 20.3% respectively. On the other hand, feature selection provides better results in the estimation of ideological self-positioning scale scores when fixing the variable measuring the inclusion in $s_v(z)$ as the target (estimated median effect: 0.942; percentage of combinations with an effect below 0.9: 20.3%). Some statistical evidences can be observed regarding the effectiveness of several methods, such as the chi-square filter, LASSO and OneR for PSA with logistic regression when estimating the economic situation in Spain, or CFS and Gain ratio when estimating the ideological self-positioning scale mean score, among other examples.

A summary of the relative bias and effect results can be found in Table 3. When Raking calibration is not applied, results on relative bias are very similar across variable selection strategies, although the most featured method among the best ones is CFS, followed by using all variables. When Raking calibration is applied, the situation slightly changes, with OneR appearing the most in the set of algorithms that provide the best results, followed by CFS. Regarding effect, it can be observed that all the variable selection algorithms (except for StepWise and Boruta when applying Raking calibration) provide more efficient estimates in more than half of the cases, a percentage that goes above 60% and even 80% of the cases for CFS and OneR when Raking is applied. In both situations mentioned, along with the case of chi-square filter, the percentage of cases where the effect is below 0.9 (reduction of the MSE above 10% in comparison to the case where all covariates are used) is almost 20% (18.8%). These results, along with the statistical evidence observed in hypothesis testing, suggest an advantage of variable selection methods in comparison to the use of all the available covariates.

The relative bias results obtained by each combination of methods in the simulation using CIS data and considering probabilities proportional to the age to obtain the nonprobability samples from the Internet population are listed in Table 14. It is worth noting that the behavior of relative bias changes for some variables; the non-raked estimates of the personal economic situation are less biased than the case where the nonprobability sample is obtained via SRSWOR from the Internet population, but more biased when estimating the rest of the variables of interest. On the other hand, feature selection algorithms offer a very similar performance to the previous case, providing less biased estimates in a variety of scenarios. The largest reductions in relative bias (compared to the case in which all variables were used) were again obtained when the variable selection algorithms used the variable of interest as the target variable, especially (but not exclusively) if Raking calibration was applied. The effect of

Table 3 Estimated mean and median of Relative Bias and Effect of estimates using PSA for each algorithm, and number of times its estimates have been among the best (Relative Bias less than 1% greater than the minimum) or have been more efficient than using all variables (Effect under 1 and under 0.9) in the real (bootstrapped) data simulation with SRSWOR from the Internet population to obtain the nonprobability sample

Use of	Algorithm	RB (%)			Effect			
raking		Mean	Median	Best	Mean	Median	Eff. < 1	Eff. < 0.9
No Raking	All vars.	7.184	8.579	16 (33.3%)				
	Boruta	7.262	8.653	5 (10.4%)	1.028	0.996	27 (56.2%)	0 (0%)
	CFS	7.200	8.706	20 (41.7%)	0.975	0.970	32 (66.7%)	4 (8.3%)
	Chi-sq.	7.026	8.536	9 (18.8%)	0.984	0.965	38 (79.2%)	3 (6.2%)
	Gain r.	7.365	9.133	10 (20.8%)	1.001	0.968	32 (66.7%)	5 (10.4%)
	LASSO	7.159	8.677	0 (0%)	0.990	0.970	32 (66.7%)	3 (6.2%)
	OneR	7.004	8.049	15 (31.2%)	0.984	0.959	38 (79.2%)	3 (6.2%)
	RF imp.	7.184	8.708	9 (18.8%)	0.996	0.984	33 (68.8%)	2 (4.2%)
	StepWise	7.404	8.903	1 (2.1%)	1.023	1.010	19 (39.6%)	0 (0%)
Raking	All vars.	9.050	7.476	15 (31.2%)				
	Boruta	9.143	7.564	4 (8.3%)	1.029	1.002	21 (43.8%)	1 (2.1%)
	CFS	8.923	7.610	14 (29.2%)	0.951	0.948	39 (81.2%)	9 (18.8%)
	Chi-sq.	8.827	7.310	13 (27.1%)	0.973	0.947	34 (70.8%)	8 (16.7%)
	Gain r.	9.000	7.601	7 (14.6%)	0.984	0.952	35 (72.9%)	9 (18.8%)
	LASSO	8.877	7.449	7 (14.6%)	0.977	0.962	32 (66.7%)	7 (14.6%)
	OneR	8.776	7.189	17 (35.4%)	0.970	0.949	39 (81.2%)	9 (18.8%)
	RF imp.	9.039	7.414	10 (20.8%)	0.991	0.972	34 (70.8%)	3 (6.2%)
	StepWise	9.264	7.566	2 (4.2%)	1.013	1.007	20 (41.7%)	2 (4.2%)

each variable selection algorithm for a given combination of adjustments (propensity model, use of calibration and target variable choice for selection), in comparison with the case in which all variables are used, is shown in Table 15. In general, the effect was noticeably lower in this scenario in comparison to the previous one (SRSWOR from the Internet population to obtain the nonprobability samples). The percentage of combinations that provided effects below 0.9 was more than 10% in the estimation all the variables of interest, and the estimated median was below 1 except for ideological self-positioning scale and feeling only Spanish variables. In those cases, it can be shown that the effect is largely different depending on the choice for the target variable in the feature selection algorithm; when estimating ideological self-positioning scale, it is better to fix the indicator variable of inclusion in s_v as the target (estimated median effect: 0.933 against 1.37 when using the variable of interest), and vice versa for feeling only Spanish (estimated median effect: 0.990 against 1.02 when using the indicator variable of inclusion in s_v). In addition, the number of statistically significant results for the effect is large, up to the point that for each variable of interest and Raking choice (except personal economic situation with no Raking, and feeling only Spanish with Raking) there is a feature selection algorithm that has a positive effect on estimators' efficiency. The aforementioned results are summarized in Table 4. When Raking calibration is not applied, the estimated mean and median relative bias observed is smaller for certain feature selection methods, more precisely: chi-square filter, OneR and Random Forest importance. These algorithms are also the ones that appear the most among the best approaches for feature selection (in terms of relative bias). The high performance of these algorithms remains in the case where Raking calibration is applied. Regarding effect, it can be noticed that chi-squared filter, OneR, Gain ratio and Random Forest importance are the ones that provide the best results, providing efficient estimates around 70% of the times or even more than 80% sometimes, while other algorithms also seem to offer a good performance, such as CFS. It is particularly relevant to observe that the effect on the estimates of using chi-square and OneR algorithms was below 0.9 more than 40% of the times when Raking calibration was used.

6 Application study

This section presents an application of variable selection for PSA in a real-world context, to estimate the population mean of two variables using a probability and a nonprobability sample. The application takes place within a study on abuse and dependence in a population of university students.

The probability sample used as the reference sample was obtained from a survey conducted in 2015,targeting students at the University of Granada (UGR), Spain. The sample was composed of $n_r = 856$ respondents, recruited in face-to-face interviews under a three-stage cluster sampling design, which produced an estimated sampling error of ±3.3% in the case of $p = q = 0.5$ with a confidence level of 95%. The survey questionnaire included screening instruments for abuse and dependence, namely the Spanish Mobile Phone Abuse Questionnaire (ATeMo) (Olivencia-Carrión et al. 2018), which provides a score between 0 and 100 points that reflects the level of mobile phone abuse of the participant. The ATeMo

Table 4 Estimated mean and median of Relative Bias and Effect of estimates using PSA for each algorithm, and number of times its estimates have been among the best (Relative Bias less than 1% greater than the minimum) or have been more efficient than using all variables (Effect under 1 and under 0.9) in the real (bootstrapped) data simulation considering inclusion probabilities proportional to age in the nonprobability sample

Use of	Algorithm	RB (%)			Effect			
raking		Mean	Median	Best	Mean	Median	Eff. < 1	Eff. < 0.9
No Raking	All vars.	10.798	8.891	8 (16.7%)				
	Boruta	10.954	8.984	1 (2.1%)	1.045	1.008	20 (41.7%)	0 (0%)
	CFS	10.742	9.006	8 (16.7%)	0.978	0.992	26 (54.2%)	9 (18.8%)
	Chi-sq.	10.431	8.363	16 (33.3%)	0.977	0.959	37 (77.1%)	10 (20.8%)
	Gain r.	10.514	8.892	20 (41.7%)	0.967	0.943	38 (79.2%)	12 (25%)
	LASSO	10.800	9.025	3 (6.2%)	1.002	1.003	21 (43.8%)	7 (14.6%)
	OneR	10.497	8.341	20 (41.7%)	0.981	0.947	37 (77.1%)	10 (20.8%)
	RF imp.	10.398	8.690	22 (45.8%)	0.978	0.968	38 (79.2%)	12 (25%)
	StepWise	11.078	9.188	0 (0%)	1.032	1.024	2 (25%)	0 (0%)
Raking	All vars.	9.759	9.141	11 (22.9%)				
	Boruta	9.911	9.126	0 (0%)	1.055	1.007	22 (45.8%)	1 (2.1%)
	CFS	9.536	8.735	11 (22.9%)	0.947	0.941	34 (70.8%)	16 (33.3%)
	Chi-sq.	9.207	8.442	20 (41.7%)	0.951	0.916	39 (81.2%)	20 (41.7%)
	Gain r.	9.757	8.665	3 (6.2%)	1.015	0.966	35 (72.9%)	10 (20.8%)
	LASSO	9.568	8.922	4 (8.3%)	1.002	0.985	30 (62.5%)	10 (20.8%)
	OneR	9.073	8.282	20 (41.7%)	0.952	0.923	40 (83.3%)	21 (43.8%)
	RF imp.	9.378	8.515	21 (43.8%)	0.957	0.938	41 (85.4%)	17 (35.4%)
	StepWise	9.815	8.886	0 (0%)	1.011	1.006	22 (45.8%)	5 (10.4%)

instrument contains 25 items of type 5-point Likert scale, where the possible values are 0, 1, 2, 3 and 4. The survey also included the Cannabis Abuse Screening Test (CAST)(Legleye et al. 2007) and the Severity of Dependence Scale (SDS) (Gossop et al. 1995), together with subscales regarding internet and videogames addiction from the MULTICAGE-CAD4 instrument (Pedrero-Pérez et al. 2007). The survey also recorded the age, gender and university faculty of each participant.

The nonprobability sample was derived from a survey completed by self-selected respondents, conducted in January 2018 and also targeting UGR students. The sample was composed of $n_v = 176$ respondents, who were recruited via snowball sampling performed by the students themselves. All of the variables included in this survey were measured in the reference sample. However, some data preprocessing was performed prior to the analysis; four respondents were ruled out because they were under 18 years old, as were another 43, who left more than 85% of the questionnaire items unanswered or who left blank all of the items of any of the scales. The final sample size, therefore, was $n_v = 129$ individuals. Missing data present in the sample was imputed using the Classification and Regression Trees (CART) algorithm (Breiman et al. 1984).

Age, gender and faculty were used as calibration variables in Raking, as the population totals (but not the cross-probabilities) were available. The covariates eligible for PSA were the total score for the CAST and SDS scales,the MULTICAGE subscales (internet and videogames), and the variables used for calibration: age, gender, and faculty. In total, seven variables were eligible for propensity modelling. The two variables of interest were present in both samples; this is not a feasible situation in real-world applications of PSA (the target variable would not be available in the probability sample) but in this case it allowed us to compare the estimations from both samples. These variables were:

- Mean score on the total ATeMo scale, which was 30.066 units in the reference sample (with a sample standard deviation of 15 units) and 32.558 units in the unweighted convenience sample (with a sample standard deviation of 13.99 units).
- Mean score on the item "I have tried to spend less time using my mobile phone but I cannot do it" (number 16 in the ATeMo instrument). The mean score of this item in the reference sample was 0.776 (with a sample standard deviation of 1.002) while in the unweighted convenience sample it was 1.217 (with a sample standard deviation of 1.132), this being the greatest difference observed in in any ATeMo item between the reference sample and the convenience sample.

Table 5 shows the distributions of the covariates available for PSA in both samples. Except for gender, the values differ greatly in the distribution of the covariates between the two samples. Overall, respondents to the online sample were younger and more prone to cannabis consumption. In addition, their score for the MULTICAGE subscales of internet and videogames addiction tended to be higher than those of the reference sample members. Finally, the Science Faculty at the UGR was clearly overrepresented in the online sample, as to a lesser extent was the

Medicine Faculty, while the other faculties were underrepresented. Given that the variability between samples can be identified in the covariates, it seems likely that PSA might be helpful to obtain more efficient estimates out of the online sample.

Estimation of the population means followed the same procedure as described in Sect. 4.3: each algorithm for variable selection was applied before PSA (with the same predictive models -and hyperparameter optimisation- as in the simulations: logistic regression, GBM, k-NN and neural networks) using the described reference and convenience samples, and the resulting weights were used directly in the estimators or as initial weights for Raking calibration. The estimated population means for each combination of methods and the estimated Leave-One-Out jackknife variance (Quenouille 1956) are shown in Tables 6 and 7 respectively.

In all cases, the use of variable selection algorithms made the estimates closer to the value observed in the reference sample. For estimation of Item number 16, selecting variables that set the indicator variable of inclusion in s_v (z) as the target variable gave subsets that provided the closest estimates for each predictive algorithm, while for the ATeMo score the best choice was to set the variable of interest (y) as the target in the variable selection algorithms. Raking calibration also helped provide estimates that were closer to the reference sample one, especially in the case of Item number 16. On the other hand, the application of these methods increased the variance of the estimator, although in general this increase was greater when any variable selection algorithm was used (with some exceptions).

7 Discussion and conclusions

In propensity estimation models for online surveys, the question of the variables to be included has been widely discussed, and in some cases questions have been included specifically to distinguish between the potentially covered population and target population individuals (Schonlau et al. 2007). Informative variables can be selected by the practitioner prior to the study, especially when there is some knowledge on the relationships between variables. However, there is often no information at all on the relationships present in the variables prior to the study, and this circumstance is even more likely in high dimensional contexts, which are becoming ever-more frequent with the development of Big Data methods in survey sampling.

In such cases, variable or feature selection algorithms may contribute to identifying the most informative subset of variables. The simulations performed in our study, using synthetic data and a real survey, reveal the impact of variable selection. In building the models, we also considered machine learning classification algorithms and the subsequent application of Raking calibration, in order to determine which alternatives are most effective in terms of bias removal.

Our analysis shows that feature selection makes a significant contribution to reducing relative bias. However, the best feature selection algorithm, in this respect

Table 5 Distributions of covariates in online and reference samples

Variable	Level	Online sample	Reference sample	p-value
Gender				
	Male	43.4%	37.6%	0.2443[b]
	Female	56.6%	62.4%	
Age				
	Mean age	20.39 ± 2.78[a]	21.12 ± 3.05[a]	0.0068[c]
Faculty				
	Computing	3.9%	9.7%	< 2.2e-16[d]
	Science	58.9%	10.6%	($\chi^2 = 209$)
	Business	3.9%	8.9%	
	Law	3.9%	8.3%	
	Humanities	5.4%	7.5%	
	Medicine	10.9%	4.3%	
	Other faculties	13.2%	50.7%	
MULTICAGE				
(internet)	0	9.3%	32.8%	3.84e-07[d]
	1	30.2%	27.3%	($\chi^2 = 35.4$)
	2	31.0%	19.7%	
	3	17.1%	14.1%	
	4	12.4%	6.0%	
MULTICAGE				
(videogames)	0	72.1%	81.7%	< 2.2e-16[d]
	1	15.5%	8.5%	($\chi^2 = 246.9$)
	2	9.3%	6.0%	
	3	3.1%	2.6%	
	4	0.0%	1.3%	
CAST				
	No consumption	21.7%	86.6%	< 2.2e-16[d]
	No issues	42.6%	4.7%	($\chi^2 = 308.8$)
	Few issues	27.1%	4.6%	
	Considerable issues	5.4%	3.0%	
	Many issues	3.1%	1.2%	
SDS				
	No consumption	22.5%	86.6%	< 2.2e-16[d]
	No issues	53.5%	8.3%	($\chi^2 = 279.7$)
	Few issues	15.5%	3.4%	
	Considerable issues	3.9%	1.4%	
	Many issues	4.7%	0.4%	

[a] Standard deviation of the age

[b] Two sample test for equality of proportions with continuity correction

[c] Welch two sample t-test

[d] Pearson's Chi-squared test

Table 6 Mean estimates of the population mean for both target variables after applying of each combination of adjustments

		Estimation of pop. mean for Item number 16				Estimation of pop. mean for ATeMo score			
		No Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	No Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)
Convenience sample		1.217				32.62			
Reference sample		0.776				30.07			
Log. reg.	All vars.	1.424	1.424	0.816	0.816	30.32	30.32	30.64	30.64
	Stepwise	1.411	1.105	0.819	0.812	30.43	30.10	30.58	30.20
	CFS	1.329	1.105	0.810	0.812	31.65	30.10	31.86	30.20
	Chi-sq.	1.329	1.329	0.810	0.810	31.65	32.00	31.86	31.83
	Gain r.	1.271	0.972	0.810	0.862	33.62	31.18	31.96	31.66
	OneR	1.262	1.329	0.827	0.810	33.59	32.00	32.59	31.83
	RF imp.	0.972	1.105	0.862	0.812	31.18	30.10	31.66	30.20
	Boruta	1.328	1.222	0.770	0.760	30.23	30.22	30.86	30.34
	LASSO	0.939	1.217	0.805	0.862	28.93	30.10	29.94	30.20
GBM	All vars.	1.093	1.242	0.768	0.797	29.98	31.19	30.84	31.12
	Stepwise	1.190	1.143	0.776	0.858	30.37	30.42	30.70	30.05
	CFS	1.129	1.151	0.772	0.864	31.95	30.05	31.68	29.97
	Chi-sq.	1.243	1.222	0.777	0.801	31.36	31.32	31.42	31.15
	Gain r.	1.267	1.000	0.821	0.862	33.58	31.26	32.08	31.66
	OneR	1.259	1.168	0.832	0.797	33.54	32.31	32.61	31.46
	RF imp.	0.995	1.144	0.862	0.854	31.32	30.25	31.66	29.97
	Boruta	1.300	1.295	0.831	0.797	29.62	29.89	30.55	30.36
	LASSO	1.017	1.217	0.836	0.862	29.09	30.37	30.11	30.11

Table 6 continued

		Estimation of pop. mean for Item number 16				Estimation of pop. mean for ATeMo score			
		No Raking		Raking		No Raking		Raking	
		Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)
Convenience sample		1.217				32.62			
Reference sample		0.776				30.07			
k-NN	All vars.	1.192	1.192	0.860	0.860	32.50	32.50	31.72	31.72
	Stepwise	0.860	1.217	0.891	0.862	27.76	32.62	31.60	31.66
	CFS	1.140	1.217	0.823	0.862	31.55	32.62	30.91	31.66
	Chi-sq.	1.140	1.140	0.823	0.823	31.55	32.40	30.91	31.54
	Gain r.	1.201	1.217	0.837	0.862	32.46	32.62	31.51	31.66
	OneR	1.217	1.140	0.862	0.823	32.62	32.40	31.66	31.54
	RF imp.	1.217	1.217	0.862	0.862	32.62	32.62	31.66	31.66
	Boruta	1.191	1.217	0.855	0.862	32.53	32.63	31.67	31.51
	LASSO	1.276	1.217	0.899	0.862	33.51	32.62	32.44	31.66
Neural nets	All vars.	1.200	1.249	0.808	0.787	30.81	35.99	31.09	33.36
	Stepwise	1.452	1.142	0.838	0.841	32.33	30.54	32.78	30.16
	CFS	1.237	1.142	0.778	0.841	31.59	30.54	31.37	30.16
	Chi-sq.	1.216	1.249	0.825	0.811	31.40	32.08	31.58	32.12
	Gain r.	1.263	0.992	0.835	0.862	33.55	31.32	32.33	31.66
	OneR	1.257	1.216	0.837	0.788	33.51	32.06	32.61	32.15
	RF imp.	0.992	1.142	0.862	0.841	31.32	30.54	31.66	30.16
	Boruta	1.269	1.315	0.765	0.837	31.19	30.18	31.19	30.48
	LASSO	0.945	1.217	0.809	0.862	29.54	30.54	30.39	30.15

Table 7 Estimated variance of the estimators, obtained via Leave-One-Out jackknife, after applying each combination of methods

		Estimation of pop. mean for Item number 16				Estimation of pop. mean for ATeMo score			
		No Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	No Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Raking Indicator variable of inclusion in s_v (z)	Variable of interest (y)
Convenience sample		0.010				1.51			
Log. regr.	All vars.	0.171	0.171	0.036	0.036	14.12	14.12	6.71	6.71
	Stepwise	0.179	0.009	0.036	0.022	14.03	1.24	6.43	6.14
	CFS	0.125	0.009	0.034	0.022	11.76	1.24	5.38	6.14
	Chi-sq.	0.125	0.125	0.034	0.034	11.76	10.07	5.38	5.51
	Gain r.	0.026	0.022	0.034	0.024	4.26	3.28	5.40	5.08
	OneR	0.025	0.125	0.036	0.034	4.16	10.07	5.17	5.51
	RF imp.	0.022	0.009	0.024	0.022	3.28	1.24	5.08	6.14
	Boruta	0.116	0.026	0.030	0.031	14.45	1.73	6.69	6.73
	LASSO	0.021	0.010	0.022	0.024	4.00	1.24	6.48	6.14
GBM	All vars.	0.077	0.060	0.027	0.027	8.21	9.88	6.57	6.08
	Stepwise	0.032	0.009	0.025	0.024	3.71	1.62	5.98	7.46
	CFS	0.086	0.009	0.029	0.024	8.11	1.72	5.86	7.69
	Chi-sq.	0.051	0.032	0.028	0.026	4.81	3.88	5.45	4.87
	Gain r.	0.024	0.021	0.034	0.024	4.02	3.08	5.22	5.08
	OneR	0.024	0.033	0.037	0.025	4.02	3.64	5.26	5.32
	RF imp.	0.021	0.009	0.024	0.024	3.13	1.71	5.08	7.65
	Boruta	0.050	0.027	0.026	0.031	6.00	1.78	6.26	6.45
	LASSO	0.019	0.010	0.023	0.024	3.84	1.56	6.81	7.24
k-NN	All vars.	0.012	0.012	0.024	0.024	1.93	1.93	5.62	5.62

Table 7 continued

		Estimation of pop. mean for Item number 16				Estimation of pop. mean for ATeMo score			
		No Raking		Raking		No Raking		Raking	
		Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)	Indicator variable of inclusion in s_v (z)	Variable of interest (y)
Convenience sample		0.010				1.51			
	Stepwise	0.012	0.010	0.028	0.024	2.48	1.51	6.17	5.08
	CFS	0.010	0.010	0.022	0.024	1.60	1.51	5.39	5.08
	Chi-sq.	0.010	0.099	0.022	0.022	1.60	1.59	5.39	5.47
	Gain r.	0.011	0.010	0.024	0.024	1.49	1.51	4.80	5.08
	OneR	0.010	0.010	0.026	0.022	1.56	1.59	5.28	5.47
	RF imp.	0.010	0.010	0.024	0.024	1.51	1.51	5.08	5.08
	Boruta	0.012	0.011	0.024	0.025	1.81	3.41	5.61	6.11
	LASSO	0.016	0.010	0.027	0.024	3.29	1.51	5.32	5.08
Neural nets	All vars.	0.101	0.129	0.029	0.030	8.89	15.04	5.15	5.69
	Stepwise	0.061	0.009	0.030	0.023	11.61	1.51	6.50	6.97
	CFS	0.125	0.009	0.029	0.023	10.16	1.51	5.16	6.97
	Chi-sq.	0.091	0.067	0.028	0.031	7.85	7.61	6.28	5.37
	Gain r.	0.023	0.021	0.035	0.024	3.93	3.13	5.29	5.08
	OneR	0.023	0.091	0.035	0.033	3.81	8.03	5.14	5.62
	RF imp.	0.021	0.009	0.024	0.023	3.13	1.51	5.08	6.97
	Boruta	0.121	0.025	0.030	0.038	10.32	1.54	5.72	7.22
	LASSO	0.019	0.010	0.022	0.024	3.02	1.51	5.35	6.97

and regarding its effect on the estimation, varies according to the dataset considered and the adjustment choices made. The best variable selection method depends on the dataset, meaning there is no one-size-fits-all solution. However, the reduction of model complexity associated with variable selection consistently produced more efficient estimators. As expected, selecting variables according to their impact on the outcome variable provided the best results overall. In line with Austin and Stuart (2015), we find that the propensity score works on the covariates included in the model, so it is preferable to include prognostically important variables (related to the outcome) as the probability to mitigate the bias in the estimation of the target variable will also be higher. In view of these results, in practice the combination of several variable selection approaches, rather than just one, might be useful to identify the best subset in each situation.

Regarding other adjustment methods, Raking calibration after PSA proved to be the most efficient technique in almost all cases. The redundancy of variables between adjustments can reduce the efficiency of their combination in some cases, as observed by Lee and Valliant (2009), who reported that the use of the same variables for PSA and calibration resulted in estimates which, despite being less biased than estimates using only PSA, underperformed versus adjustments with no redundancy.

On the other hand, the use of classification algorithms instead of logistic regression for estimating propensities was advantageous overall, but only for certain algorithms and with no clear view as to which was the best algorithm for estimation. The application of this sort of algorithm in nonprobability sampling was recently studied by Buelens et al. (2018) as an option for model-based estimates, and by Castro-Martín et al. (2020b), Ferri-García and Rueda (2020) and Ferri-Garca et al. (2020) for PSA in online surveys. It has also been studied for PSA in nonresponse adjustment (Phipps and Toth 2012; Buskirk and Kolenikov 2015), with promising results. Further studies should take into account this approach, together with the use of a wider range of algorithms, and should consider how preprocessing (such as the feature selection applied in the present study) might influence their performance in propensity estimation.

Further research is needed regarding the implications of variable selection on nonprobability samples, as our study presents certain limitations. Most importantly, relatively few covariates were available for each simulation and for the application study. Originally, feature selection algorithms were intended to reduce dimensionality in large data sets, facilitating the selection of only the most significant variables for prediction. Further research into these algorithms in PSA for selection bias treatment using a larger number of covariates would enhance our understanding of these questions. However, our results also support their use in a low dimensional context, meaning that the value of these algorithms could extend beyond computing optimisation. For example, the use of variable selection algorithms could be extended to calibration; although research has shown their potential and some methods have been developed in this area (Chen et al. 2019), further study is needed to consider this topic, as calibration requires little information and therefore can be more widely applied. Finally, the use of more powerful algorithms for propensity estimation, such as deep learning techniques, should be considered in future studies, as these methods usually involve automatic variable selection and could provide more precise estimates.

Funding This study was partially supported by Ministerio de Ciencia e Innovación, Spain [Grant No. PID2019-106861RBI00/AEI/10.13039/501100011033]. The first-named author also received a FPU grant from Ministerio de Ciencia, Innovación y Universidades. Funding for open access charge: Universidad de Granada / CBUA Spain.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Table 8 Mean Relative Bias of the estimates of population means for variables y_1, y_2, y_3, y_4 in the artificial data simulation for each combination of methods

		Rakinge = No				Raking = Yes				Raking = No				Raking = Yes			
		GLM	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_1 (MCAR)								y_2 (MCAR)							
Indicator variable of inclusion in s_v (z)	All vars.	0.158	0.139	0.087	0.186	0.205	0.182	0.096	0.228	0.048	0.049	0.049	0.047	0.053	0.054	0.056	0.051
	Boruta	0.159	0.17	0.176	0.188	0.209	0.218	0.195	0.23	0.047	0.048	0.048	0.049	0.052	0.054	0.051	0.053
	CFS	0.166	0.161	0.188	0.161	0.201	0.214	0.179	0.195	0.049	0.048	0.051	0.049	0.053	0.055	0.053	0.056
	Chi-sq.	0.181	0.186	0.224	0.192	0.205	0.223	0.231	0.22	0.049	0.052	0.054	0.052	0.053	0.058	0.056	0.059
	Gain r.	0.153	0.185	0.184	0.182	0.19	0.235	0.181	0.232	0.049	0.051	0.048	0.049	0.053	0.059	0.052	0.055
	LASSO	0.161	0.195	0.209	0.191	0.202	0.225	0.208	0.211	0.049	0.047	0.056	0.049	0.052	0.054	0.061	0.054
	StepWise	0.159	0.17	0.156	0.187	0.211	0.212	0.15	0.194	0.047	0.049	0.056	0.049	0.053	0.055	0.061	0.054
	OneR	0.164	0.18	0.171	0.209	0.191	0.233	0.198	0.282	0.05	0.048	0.05	0.05	0.053	0.054	0.057	0.057
	RF imp.	0.171	0.18	0.201	0.163	0.208	0.228	0.181	0.182	0.05	0.051	0.059	0.053	0.053	0.059	0.065	0.06
Variable of interest (y)	All vars.	0.158	0.166	0.09	0.187	0.205	0.212	0.093	0.208	0.048	0.049	0.049	0.046	0.053	0.054	0.056	0.05
	Boruta	0.186	0.193	0.178	0.214	0.205	0.216	0.202	0.206	0.048	0.051	0.053	0.053	0.054	0.054	0.06	0.062
	CFS	0.163	0.16	0.159	0.161	0.166	0.166	0.166	0.169	0.082	0.074	0.058	0.077	0.068	0.067	0.055	0.07
	Chi-sq.	0.175	0.17	0.159	0.17	0.166	0.161	0.167	0.167	0.047	0.046	0.048	0.046	0.052	0.052	0.052	0.052
	Gain r.	0.168	0.166	0.159	0.162	0.166	0.166	0.166	0.166	0.046	0.046	0.048	0.046	0.052	0.052	0.052	0.052
	LASSO	0.161	0.186	0.17	0.188	0.163	0.182	0.179	0.179	0.048	0.048	0.048	0.048	0.052	0.052	0.052	0.052
	StepWise	0.183	0.169	0.146	0.167	0.202	0.193	0.137	0.178	0.048	0.05	0.047	0.049	0.052	0.054	0.052	0.053
	OneR	0.195	0.19	0.156	0.188	0.182	0.181	0.166	0.172	0.047	0.047	0.048	0.045	0.052	0.052	0.052	0.05
	RF imp.	0.163	0.162	0.148	0.182	0.213	0.18	0.171	0.197	0.048	0.051	0.056	0.049	0.056	0.057	0.062	0.055

Table 8 continued

		Rakinge = No				Raking = Yes				Raking = No				Raking = Yes			
		GLM	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_3 (MNAR)								y_4 (MNAR)							
Indicator variable of inclusion in s_v (z)	All vars.	11.55	11.67	11.11	11.74	8.96	9.51	9.12	9.8	12.39	12.52	11.9	12.61	9.67	10.25	9.83	10.57
	Boruta	11.55	11.71	11.28	11.82	8.96	9.57	9.36	9.9	12.39	12.56	12.05	12.68	9.67	10.33	10.05	10.7
	CFS	11.46	11.56	11.18	11.64	8.83	9.27	9.1	9.47	12.29	12.39	12.01	12.48	9.52	10.02	9.84	10.22
	Chi-sq.	11.51	11.63	11.27	11.72	8.83	9.37	9.22	9.59	12.34	12.47	12.08	12.57	9.53	10.12	9.95	10.33
	Gain r.	11.42	11.54	11.17	11.61	8.84	9.29	9.12	9.43	12.26	12.37	12	12.45	9.53	10.02	9.83	10.19
	LASSO	11.53	11.64	11.26	11.75	8.87	9.41	9.23	9.66	12.36	12.49	12.07	12.61	9.56	10.15	9.97	10.43
	StepWise	11.54	11.7	11.25	11.82	8.96	9.56	9.3	9.93	12.38	12.56	12.07	12.68	9.66	10.32	10.08	10.68
	OneR	11.41	11.43	11.1	11.49	8.78	9.08	8.95	9.18	12.23	12.26	11.92	12.32	9.47	9.79	9.65	9.91
	RF imp.	11.45	11.65	11.27	11.74	8.88	9.51	9.32	9.71	12.29	12.5	12.09	12.61	9.59	10.26	10.03	10.5
Variable of interest (y)	All vars.	11.55	11.66	11.1	11.72	8.96	9.5	9.11	9.78	12.39	12.51	11.9	12.62	9.67	10.24	9.83	10.59
	Boruta	11.46	11.65	11.2	11.69	8.95	9.52	9.23	9.69	12.37	12.55	12.03	12.67	9.65	10.3	10.01	10.65
	CFS	11.28	11.27	10.54	11.26	8.66	8.66	8.64	8.67	12.2	12.5	12.07	12.61	9.65	10.37	10.05	10.54
	Chi-sq.	11.29	11.28	10.54	11.26	8.65	8.66	8.64	8.65	12.38	12.59	12.12	12.7	9.64	10.36	10.15	10.72
	Gain r.	11.26	11.26	10.54	11.24	8.66	8.67	8.64	8.67	12.36	12.56	12.13	12.68	9.56	10.29	10.09	10.6
	LASSO	10.51	10.51	10.5	10.5	8.62	8.62	8.62	8.62	12.39	12.58	12.16	12.72	9.55	10.27	10.09	10.59
	StepWise	11.41	11.38	10.9	11.43	8.85	8.92	8.78	8.92	12.4	12.58	12.11	12.7	9.59	10.28	10.05	10.6
	OneR	10.98	11.03	10.78	11.09	8.77	9.01	8.84	9.12	12.39	12.6	12.12	12.7	9.66	10.38	10.15	10.72
	RF imp.	11.25	11.38	11.02	11.44	8.9	9.23	9.04	9.33	12.38	12.58	12.15	12.72	9.58	10.33	10.13	10.65

The closer to zero a value, the less biased the mean estimate obtained

Table 9 Mean Relative Bias of the estimates of population means for variables y_5, y_6, y_7, y_8 in the artificial data simulation for each combination of methods

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_5 (MAR)								y_6 (MAR)							
Indicator variable of inclusion in s_v (z)	All vars.	18.12	16.55	14.92	15.84	8.56	7.7	6.69	7.18	16.91	14.97	13.51	14.44	2.88	2.51	2.11	2.31
	Boruta	18.13	16.41	14.95	15.76	8.57	7.52	6.74	7.03	16.91	14.89	13.47	14.39	2.89	2.45	2.15	2.27
	CFS	18.02	17.28	15.82	17.15	8.26	8.42	7.45	8.33	16.83	15.53	14.14	15.45	2.81	2.7	2.34	2.66
	Chi-sq.	18.12	17.04	15.64	16.89	8.18	8.06	7.2	7.94	16.99	15.42	14.06	15.34	2.76	2.6	2.26	2.55
	Gain r.	17.97	17.24	15.82	17.16	8.36	8.44	7.48	8.39	16.69	15.44	14.11	15.4	2.87	2.72	2.36	2.69
	LASSO	18.12	16.93	15.48	16.66	8.26	7.91	7.09	7.66	16.98	15.34	13.91	15.18	2.79	2.56	2.23	2.46
	StepWise	18.1	16.38	14.97	15.82	8.54	7.46	6.78	7.09	16.89	14.9	13.48	14.46	2.87	2.43	2.15	2.28
	OneR	17.92	17.74	16.2	17.9	8.01	8.93	7.83	9.34	16.89	15.94	14.52	16.05	2.72	2.85	2.45	2.96
	RF imp.	18.06	16.63	15.26	16.29	8.68	7.86	6.99	7.41	16.56	14.9	13.64	14.71	3	2.58	2.25	2.43
Variable of interest (y)	All vars.	18.12	16.57	14.92	15.83	8.56	7.72	6.68	7.14	16.91	14.97	13.51	14.44	2.88	2.51	2.12	2.3
	Boruta	18.03	16.71	15.27	16.2	8.62	7.99	7.11	7.62	16.78	15.33	13.11	14.93	2.65	2.34	2.11	2.2
	CFS	18.15	16.8	15.41	16.5	8.3	7.8	6.98	7.46	15.8	15.4	14.28	15.49	3.24	3.16	2.71	3.21
	Chi-sq.	17.9	17.13	15.74	16.99	8.73	8.59	7.63	8.5	16.43	16.4	12.15	16.38	2	2	2	2
	Gain r.	18.14	16.6	15.23	16.32	8.28	7.57	6.81	7.21	16.43	16.4	12.15	16.38	2	2	2	2
	LASSO	17.82	17.85	16.34	18.04	8.47	9.34	8.13	9.65	16.88	16.03	14.58	16.15	2.71	2.89	2.48	3.02
	StepWise	17.9	17.5	15.89	17.41	8.36	8.87	7.63	8.85	16.89	15.78	14.25	15.77	2.73	2.8	2.37	2.82
	OneR	17.54	17.6	16.21	17.7	8.99	9.47	8.26	9.62	16.88	16.03	14.59	16.16	2.71	2.89	2.48	3.01
	RF imp.	17.64	17.77	16.38	17.93	9	9.62	8.39	9.83	16.43	16.4	12.15	16.38	2	2	2	2

Table 9 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_7 (MNAR)								y_8 (MNAR)							
Indicator variable of inclusion in s_v (z)	All vars.	10.88	10.46	9.27	10.23	7.7	6.75	6.2	6.88	15.53	15.65	14.2	15.39	11.11	11.08	10.55	11.44
	Boruta	10.88	10.42	9.38	10.23	7.7	6.69	6.33	6.83	15.54	15.65	14.39	15.47	11.11	11.15	10.78	11.52
	CFS	10.61	10.57	9.66	10.55	6.86	6.96	6.45	6.95	15.69	15.7	14.65	15.75	10.52	10.91	10.59	11.05
	Chi-sq.	10.76	10.55	9.66	10.54	6.91	6.8	6.37	6.79	15.89	15.77	14.73	15.86	10.56	10.95	10.66	11.1
	Gain r.	10.58	10.54	9.66	10.54	6.88	6.95	6.46	6.94	15.63	15.65	14.64	15.74	10.53	10.91	10.59	11.04
	LASSO	10.82	10.54	9.62	10.49	7.11	6.76	6.38	6.75	15.86	15.78	14.66	15.83	10.7	11	10.7	11.2
	StepWise	10.87	10.39	9.44	10.23	7.69	6.66	6.4	6.85	15.52	15.62	14.4	15.44	11.1	11.14	10.82	11.52
	OneR	10.2	10.49	9.66	10.59	6.53	7.15	6.57	7.31	15.15	15.43	14.53	15.55	10.31	10.78	10.47	10.91
	RF imp.	10.92	10.49	9.6	10.41	7.19	6.68	6.29	6.52	16.05	15.83	14.72	15.9	10.72	11.04	10.7	11.16
Variable of interest (y)	All vars.	10.88	10.46	9.26	10.21	7.7	6.74	6.2	6.83	15.53	15.64	14.2	15.39	11.11	11.08	10.54	11.43
	Boruta	10.78	10.4	9.4	10.27	7.71	6.86	6.41	6.96	15.08	15.12	14.02	14.96	10.85	10.85	10.54	11.16
	CFS	10.75	10.34	9.45	10.26	7.84	6.96	6.47	6.96	15.01	14.61	13.69	14.72	10.52	10.47	10.17	10.49
	Chi-sq.	10.5	10.33	9.41	10.32	7.37	7.05	6.53	7.16	14.1	14.12	13.38	14.05	10.29	10.31	10.22	10.42
	Gain r.	10.8	10.36	9.48	10.3	7.91	6.93	6.49	6.96	13.46	13.47	12.95	13.44	9.94	9.94	9.94	9.94
	LASSO	10.13	10.22	9.27	10.29	6.99	7.2	6.5	7.26	15.53	15.63	14.43	15.48	11.11	11.17	10.88	11.53
	StepWise	10.4	10.32	9.37	10.4	7.18	6.96	6.33	6.95	15.52	15.61	14.37	15.45	11.1	11.14	10.8	11.5
	OneR	9.59	9.71	9	9.78	6.55	6.85	6.31	6.9	15.01	14.64	13.69	14.72	10.52	10.47	10.18	10.5
	RF imp.	10	10.16	9.23	10.29	6.76	7.24	6.53	7.35	14.47	14.46	13.7	14.53	10.5	10.51	10.42	10.8

The closer to zero a value, the less biased the mean estimate obtained

Table 10 Effect of the estimates of population means for variables y_1, y_2, y_3, y_4 in the artificial data simulation for each combination of methods

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_1 (MCAR)								y_2 (MCAR)							
Indicator variable of inclusion in s_v (z)	Boruta	0.998	1.018	1.008	0.988	1	1.032	1.02	1.006	0.996	0.974	1.032	0.99	0.998	0.966	1.032	0.997
	CFS	1.006	1.03	1.062	1.004	1.013	1.031	1.094	1.012	0.993	0.996	1.041	0.977	0.985	0.985	1.02	0.969
	Chi-sq.	1.003	1.027	1.051	1.014	1.006	1.032	1.05	1.036	0.989	0.989	1.036	0.972	0.981	0.991	1.018	0.982
	Gain r.	1.003	1.015	1.037	0.995	1.014	1.023	1.065	1.001	0.987	0.988	1.068	0.974	0.99	1.003	1.068	0.98
	LASSO	1.002	1.025	1.044	1.007	0.995	1.031	1.05	1.026	0.999	1	1.025	0.995	0.986	0.991	1.009	0.994
	StepWise	0.999	1.02	0.998	1.001	1.002	1.027	1.012	1.018	1	1.006	1.02	0.969	1.002	1.004	1.023	0.965
	OneR	1.008	1.032	1.085	1	1.006	1.026	1.11	1	0.984	0.998	1.021	0.983	0.972	0.995	1.017	0.989
	RF imp.	1	1.031	1.034	1.017	1.014	1.054	1.065	1.034	0.996	0.993	1.047	0.973	0.996	0.987	1.04	0.959
Variable of interest (y)	Boruta	1.002	0.986	0.999	0.983	1.004	0.944	0.994	0.926	0.996	0.99	1.024	0.997	0.995	0.985	1.009	0.998
	CFS	1	1.017	0.975	1.001	0.996	0.974	0.975	0.931	0.988	1.01	0.982	1.003	0.979	1.009	0.981	0.991
	Chi-sq.	1.006	1.024	0.975	1.007	0.996	0.976	0.975	0.93	0.971	0.978	0.926	0.984	0.96	0.971	0.93	0.959
	Gain r.	1.001	1.019	0.975	0.999	0.996	0.974	0.975	0.932	0.968	0.978	0.926	0.971	0.96	0.97	0.93	0.958
	LASSO	0.971	0.981	0.976	0.959	0.997	0.967	0.971	0.923	0.936	0.944	0.926	0.941	0.961	0.968	0.93	0.955
	StepWise	0.993	1.007	1.015	0.972	0.995	0.976	1.023	0.919	0.987	0.992	0.977	0.998	0.968	0.982	0.951	0.97
	OneR	1.025	1.059	0.976	1.042	0.987	0.985	0.975	0.94	0.968	0.975	0.926	0.979	0.96	0.967	0.93	0.958
	RF imp.	0.988	1.008	1.006	0.984	0.999	0.976	1.01	0.932	0.958	0.979	1	0.975	0.99	0.983	0.99	0.989
		y_3 (MNAR)								y_4 (MNAR)							
Indicator variable of inclusion in s_v (z)	Boruta	1	1.009	1.028	1.013	1	1.013	1.048	1.023	1	1.007	1.024	1.011	1	1.016	1.045	1.023
	CFS	0.984	0.983	1.013	0.985	0.972	0.959	1.002	0.945	**0.984***	**0.981***	1.018	**0.98***	**0.971***	**0.956***	1.002	**0.934***
	Chi-sq.	0.993	0.994	1.028	0.998	0.972	0.976	1.025	0.967	0.993	0.993	1.031	0.993	**0.972***	**0.976***	1.025	**0.955***
	Gain r.	0.978	0.981	1.012	0.98	0.972	0.963	1.007	**0.939***	**0.979***	**0.978***	1.017	**0.976***	**0.973***	**0.956***	1.001	**0.929***

Table 10 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	LASSO	0.996	0.996	1.026	1.002	0.979	0.983	1.027	0.978	0.995	0.996	1.029	1.001	**0.979***	**0.982***	1.03	**0.975***
	StepWise	0.999	1.006	1.025	1.014	0.999	1.012	1.039	1.028	0.999	1.007	1.029	1.011	0.999	1.014	1.051	1.021
	OneR	0.976	0.962	1.001	0.961	0.961	**0.924***	0.976	**0.894***	**0.975***	**0.959***	1.003	**0.955***	**0.959***	**0.914***	**0.965***	**0.878***
	RF imp.	0.984	0.997	1.028	1.001	0.983	1.002	1.043	0.986	**0.984***	0.998	1.031	1.001	**0.984***	1.003	1.041	**0.987***
Variable of	Boruta	0.984	0.999	1.017	0.995	0.998	1.005	1.026	0.983	0.998	1.006	1.023	1.007	0.997	1.012	1.038	1.011
interest (y)	CFS	**0.955***	**0.938***	**0.905***	**0.925***	**0.935***	**0.844***	**0.906***	**0.798***	**0.97***	0.998	1.029	0.998	0.997	1.025	1.046	0.991
	Chi-sq.	**0.957***	**0.939***	**0.905***	**0.925***	**0.934***	**0.843***	**0.907***	**0.796***	1	1.013	1.038	1.012	0.995	1.023	1.067	1.024
	Gain r.	**0.952***	**0.935***	**0.904***	**0.923***	**0.934***	**0.844***	**0.907***	**0.798***	0.996	1.008	1.039	1.008	**0.979***	1.011	1.054	1.002
	LASSO	**0.834***	**0.82***	**0.898***	**0.81***	**0.928***	**0.837***	**0.902***	**0.79***	1.001	1.012	1.045	1.015	**0.976***	1.007	1.055	1.001
	StepWise	0.976	**0.956***	0.968	**0.954***	0.975	**0.892***	0.938	**0.845***	1.002	1.011	1.036	1.012	**0.985***	1.008	1.047	1.002
	OneR	**0.915***	**0.909***	**0.951***	**0.909***	0.962	**0.918***	0.952	**0.89***	1	1.014	1.037	1.012	0.999	1.027	1.067	1.025
	RF imp.	**0.951***	0.957	0.988	0.957	0.987	0.951	0.986	**0.918***	0.999	1.012	1.043	1.015	**0.983***	1.018	1.063	1.012

Values greater than one indicate inefficiency, while values below one show that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used. Values in **bold** indicate that the hypothesis of the effect being equal or greater than 1 can be rejected ($\alpha = 0.05$) for that combination of methods

*Reject the null hypothesis that the effect is equal or greater than 1 for this combination of methods ($\alpha = 0.05$)

Table 11 Effect of the estimates of population means for variables y_5, y_6, y_7, y_8 in the artificial data simulation for each combination of methods

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		y_5 (MAR)								y_6 (MAR)							
Indicator	Boruta	1.001	0.983	1.003	0.989	1.002	0.957	1.016	0.962	1	0.989	0.995	0.993	1.003	0.957	1.031	0.963
variable of	CFS	0.99	1.089	1.123	1.169	**0.936***	1.18	1.223	1.323	0.991	1.077	1.097	1.145	0.957	1.164	1.225	1.327
inclusion in	Chi-sq.	1	1.06	1.099	1.136	**0.919***	1.09	1.146	1.213	1.009	1.062	1.084	1.129	**0.921***	1.074	1.145	1.213
s_v (z)	Gain r.	0.985	1.083	1.123	1.17	0.958	1.187	1.236	1.341	**0.976***	1.063	1.092	1.136	0.995	1.181	1.256	1.357
	LASSO	1.001	1.047	1.077	1.106	**0.937***	1.061	1.125	1.148	1.008	1.05	1.061	1.105	**0.938***	1.043	1.115	1.136
	StepWise	0.998	0.98	1.006	0.996	0.995	0.946	1.028	0.98	0.998	0.991	0.996	1.002	0.994	**0.941***	1.039	0.97
	OneR	**0.978***	1.146	1.178	1.27	**0.882***	1.309	1.336	1.612	0.998	1.134	1.156	1.233	**0.892***	1.277	1.337	1.599
	RF imp.	0.994	1.01	1.047	1.058	1.026	1.048	1.095	1.072	**0.961***	0.992	1.021	1.038	1.082	1.063	1.135	1.116
Variable of	Boruta	0.991	1.017	1.049	1.047	1.015	1.077	1.133	1.142	**0.985***	1.051	**0.945***	1.072	**0.869***	**0.883***	0.996	**0.917***
interest (y)	CFS	1.004	1.028	1.067	1.085	0.947	1.031	1.097	1.1	**0.873***	1.056	1.118	1.148	1.245	1.544	1.622	1.878
	Chi-sq.	**0.977***	1.069	1.117	1.152	1.042	1.252	1.313	1.423	**0.943***	1.197	**0.808***	1.283	**0.487***	**0.634***	**0.89***	**0.745***
	Gain r.	1.002	1.004	1.042	1.062	**0.94***	0.972	1.043	1.024	**0.943***	1.198	**0.808***	1.283	**0.487***	**0.634***	**0.89***	**0.745***
	LASSO	**0.968***	1.156	1.199	1.29	0.992	1.421	1.445	1.723	0.996	1.145	1.166	1.248	**0.888***	1.307	1.359	1.667
	StepWise	**0.976***	1.114	1.134	1.207	0.965	1.299	1.281	1.49	0.998	1.111	1.114	1.191	**0.902***	1.232	1.252	1.48
	OneR	**0.938***	1.126	1.181	1.244	1.09	1.458	1.486	1.717	0.996	1.145	1.167	1.25	**0.888***	1.308	1.366	1.662
	RF imp.	**0.948***	1.146	1.206	1.274	1.096	1.491	1.531	1.778	**0.943***	1.198	**0.808***	1.283	**0.487***	**0.634***	**0.89***	**0.745***
		y_7 (MNAR)								y_8 (MNAR)							
Indicator	Boruta	1.001	0.991	1.025	0.998	0.999	0.981	1.039	0.984	1.002	1	1.027	1.01	0.999	1.011	1.045	1.014
variable of	CFS	**0.956***	1.02	1.086	1.059	**0.804***	1.052	1.074	1.013	1.021	1.006	1.063	1.047	**0.897***	**0.97***	1.007	**0.933***
inclusion in	Chi-sq.	0.981	1.016	1.085	1.057	**0.816***	1.011	1.049	0.971	1.047	1.015	1.074	1.06	**0.904***	**0.976***	1.02	**0.94***

Table 11 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
s_v (z)	Gain r.	**0.949***	1.014	1.086	1.057	**0.809***	1.05	1.077	1.015	1.013	0.999	1.062	1.044	**0.898***	**0.968***	1.007	**0.93***
	LASSO	0.991	1.014	1.078	1.048	**0.862***	1	1.056	0.964	1.044	1.018	1.067	1.058	**0.928***	**0.985***	1.029	**0.959***
	StepWise	0.998	0.986	1.039	1	0.996	0.975	1.068	0.991	0.999	0.997	1.028	1.007	0.998	1.01	1.051	1.013
	OneR	**0.88***	1.002	1.085	1.065	**0.728***	1.106	1.11	1.11	**0.949***	**0.97***	1.045	1.018	**0.86***	**0.945***	**0.985***	**0.909***
	RF imp.	1.008	1.006	1.073	1.033	**0.877***	0.981	1.029	**0.902***	1.068	1.023	1.075	1.066	**0.931***	0.992	1.029	**0.951***
Variable of interest (y)	Boruta	0.983	0.989	1.03	1.01	1.006	1.032	1.068	1.034	**0.946***	**0.938***	**0.977***	**0.948***	**0.956***	**0.961***	1	**0.957***
	CFS	0.978	0.979	1.042	1.009	1.033	1.063	1.086	1.036	**0.934***	**0.873***	**0.93***	**0.914***	**0.895***	**0.892***	**0.93***	**0.843***
	Chi-sq.	**0.935***	0.976	1.032	1.02	**0.925***	1.089	1.102	1.089	**0.829***	**0.82***	**0.889***	**0.838***	**0.858***	**0.867***	**0.94***	**0.834***
	Gain r.	0.987	0.982	1.049	1.017	1.053	1.052	1.091	1.035	**0.753***	**0.743***	**0.83***	**0.764***	**0.8***	**0.805***	**0.889***	**0.756***
	LASSO	**0.873***	**0.956***	1.005	1.015	**0.83***	1.13	1.099	1.12	1	0.998	1.034	1.011	1	1.016	1.065	1.019
	StepWise	**0.917***	0.975	1.024	1.036	**0.877***	1.072	1.051	1.046	0.999	0.997	1.025	1.007	0.999	1.011	1.05	1.014
	OneR	**0.781***	**0.864***	**0.945***	**0.918***	**0.73***	1.022	1.03	1.008	**0.934***	**0.877***	**0.931***	**0.915***	**0.895***	**0.893***	**0.932***	**0.843***
	RF imp.	**0.848***	**0.946***	0.995	1.015	**0.779***	1.138	1.102	1.141	**0.872***	**0.86***	**0.933***	**0.895***	**0.894***	**0.901***	**0.977***	**0.897***

Values greater than one indicate inefficiency, while values below one show that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used. Values in **bold** indicate that the hypothesis of the effect being equal or greater than 1 can be rejected ($\alpha = 0.05$) for that combination of methods

*Reject the null hypothesis that the effect is equal or greater than 1 for this combination of methods ($\alpha = 0.05$)

Table 12 Mean relative bias of the estimates of population means in the real data simulation for each combination of methods, considering SRSWOR from the Internet population to obtain the nonprobability samples

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		Econ. situation in Spain "poor" or "very poor"								Personal econ. situation "poor" or "very poor"							
Indicator variable of inclusion in s_v (z)	All vars.	6.99	6.00	5.02	5.61	7.37	6.30	5.07	6.15	11.30	10.82	10.31	10.51	24.17	23.02	20.17	22.32
	Boruta	7.00	5.94	6.02	5.49	7.24	6.18	6.23	5.77	10.74	10.36	11.13	10.40	23.71	22.88	21.62	22.63
	CFS	7.79	6.27	5.29	6.25	7.61	6.16	5.42	6.09	12.90	11.87	10.62	12.06	24.01	23.41	20.64	23.55
	Chi-sq.	7.00	5.84	5.27	5.52	7.06	5.86	5.39	5.48	10.59	10.34	10.80	10.22	23.43	22.88	20.80	22.66
	Gain r.	8.09	6.46	5.25	6.60	7.86	6.26	5.37	6.37	14.11	12.93	10.48	13.18	24.30	23.76	20.47	23.96
	LASSO	7.05	5.91	5.62	5.60	7.14	5.98	5.82	5.62	10.68	10.45	11.13	10.31	23.69	23.13	21.39	22.93
	StepWise	7.20	6.05	6.15	5.70	7.24	6.09	6.35	5.68	10.86	10.59	11.61	10.42	23.80	23.19	21.95	22.97
	OneR	6.58	5.64	5.27	5.30	6.74	5.81	5.34	5.49	9.72	9.63	10.80	9.52	22.74	22.25	20.84	22.09
	RF imp.	7.22	5.97	5.31	5.70	7.20	5.98	5.46	5.68	11.15	10.72	10.68	10.57	23.40	22.90	20.68	22.68
Variable of interest (y)	All vars.	6.99	6.00	5.00	5.68	7.37	6.25	5.07	6.19	11.30	10.70	10.31	10.72	24.17	22.88	20.17	22.46
	Boruta	6.79	5.90	5.42	5.84	6.95	6.03	5.46	6.01	11.69	11.00	10.31	10.97	24.27	23.08	20.24	22.68
	CFS	6.63	5.96	5.77	5.94	6.52	5.97	5.63	5.94	10.26	9.90	11.12	9.63	22.39	21.65	21.19	21.27
	Chi-sq.	5.96	5.57	5.23	5.41	6.12	5.74	5.23	5.62	10.48	10.12	10.95	10.07	21.95	21.37	20.95	21.12
	Gain r.	6.41	5.77	5.29	5.86	6.41	5.80	5.39	5.90	10.71	10.49	10.60	10.56	22.10	21.75	20.71	21.70
	LASSO	6.05	5.68	5.44	5.68	6.12	5.79	5.49	5.80	10.63	10.56	10.56	10.60	20.83	20.73	20.56	20.75
	StepWise	7.38	6.36	6.02	6.24	7.31	6.44	6.08	6.34	11.93	11.18	11.95	10.75	24.31	23.28	22.43	22.73
	OneR	5.72	5.47	5.34	5.45	5.90	5.64	5.45	5.65	10.47	10.27	10.72	10.14	21.46	21.09	20.73	20.89
	RF imp.	6.81	5.88	5.20	5.47	7.22	6.17	5.25	5.97	11.32	10.65	10.40	10.57	23.81	22.55	20.29	22.09
		Ideological self-positioning scale (1–10)								Central gov. management "poor" or "very poor"							
Indicator Variable of	All vars.	2.86	2.75	2.41	2.59	2.19	2.18	1.89	2.11	1.06	1.52	2.08	2.28	5.85	5.92	4.97	6.36
	Boruta	2.92	2.78	2.71	2.67	2.24	2.23	2.15	2.20	0.85	1.63	1.60	2.20	5.62	6.27	4.75	6.64

Table 12 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
Inclusion in	CFS	2.74	2.65	2.41	2.64	2.13	2.09	1.87	2.08	1.91	2.35	2.04	2.44	5.81	6.66	5.02	6.76
s_v (z)	Chi-sq.	2.80	2.69	2.41	2.62	2.12	2.09	1.87	2.07	1.10	1.78	2.08	2.14	5.74	6.50	5.08	6.91
	Gain r.	2.72	2.63	2.40	2.64	2.15	2.10	1.87	2.10	2.35	2.71	2.05	2.67	5.84	6.67	5.00	6.62
	LASSO	2.82	2.71	2.47	2.62	2.14	2.12	1.91	2.09	0.95	1.67	2.06	2.10	5.61	6.38	5.22	6.84
	StepWise	2.80	2.68	2.53	2.61	2.11	2.10	1.95	2.07	1.07	1.81	1.95	2.23	5.75	6.60	5.13	7.02
	OneR	2.79	2.68	2.44	2.60	2.10	2.07	1.90	2.05	1.02	1.63	2.13	2.02	5.70	6.29	5.11	6.62
	RF imp.	2.76	2.66	2.43	2.59	2.11	2.08	1.89	2.05	1.33	1.90	2.04	2.21	5.71	6.46	5.01	6.79
Variable of	All vars.	2.86	2.76	2.41	2.58	2.19	2.20	1.89	2.11	1.06	1.47	2.07	2.27	5.85	5.87	4.96	6.19
interest (y)	Boruta	3.35	3.35	2.40	3.31	2.84	2.84	1.87	2.80	0.82	1.11	1.80	1.66	4.93	5.12	4.71	5.44
	CFS	2.54	2.53	2.40	2.53	1.96	1.97	1.86	1.97	0.29	0.17	1.37	0.42	3.52	3.81	4.37	4.02
	Chi-sq.	3.25	3.09	2.79	3.02	2.71	2.54	2.29	2.49	0.56	0.78	1.72	0.98	4.18	4.25	4.66	4.40
	Gain r.	3.35	3.34	2.43	3.32	2.85	2.83	1.90	2.81	0.40	0.62	1.66	0.66	3.78	3.94	4.63	4.00
	LASSO	3.17	3.01	2.75	2.89	2.63	2.48	2.20	2.37	0.89	1.13	1.84	1.41	4.35	4.51	4.84	4.71
	StepWise	3.04	2.88	2.53	2.71	2.47	2.34	1.96	2.21	0.60	0.98	1.67	1.48	4.60	4.87	4.72	5.26
	OneR	3.29	3.13	2.82	3.09	2.77	2.60	2.34	2.56	0.57	0.80	1.94	0.99	4.21	4.35	4.94	4.44
	RF imp.	3.22	3.11	2.65	3.03	2.67	2.57	2.15	2.52	1.27	1.53	2.02	2.10	5.54	5.50	4.90	5.86
		Preference for a state without autonomous comm.								Feels only Spanish							
Indicator	All vars.	12.40	11.74	10.17	11.04	8.85	8.49	7.58	8.41	10.74	11.10	10.39	10.83	9.24	9.88	9.27	9.73
variable of	Boruta	12.46	11.70	10.79	11.11	8.85	8.34	8.11	8.15	11.17	11.31	10.57	10.97	9.72	10.19	9.50	10.08
inclusion in	CFS	12.25	11.68	10.22	11.65	9.09	8.36	7.60	8.36	11.28	11.79	10.45	11.72	10.25	10.97	9.33	10.95
s_v (z)	Chi-sq.	12.41	11.69	10.20	11.30	8.83	8.22	7.56	8.07	11.35	11.57	10.41	11.38	9.89	10.49	9.28	10.52

Table 12 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	Gain r.	12.22	11.70	10.17	11.81	9.28	8.47	7.57	8.59	11.08	11.78	10.42	11.76	10.33	11.09	9.31	11.06
	LASSO	12.44	11.71	10.44	11.33	8.89	8.28	7.76	8.18	11.42	11.56	10.63	11.43	9.90	10.47	9.48	10.60
	StepWise	12.38	11.68	10.57	11.28	8.78	8.20	7.82	8.05	11.29	11.54	10.66	11.31	9.84	10.46	9.54	10.49
	OneR	12.14	11.47	10.34	11.13	8.53	8.01	7.71	7.95	11.26	11.38	10.52	11.27	9.73	10.18	9.39	10.22
	RF imp.	12.31	11.62	10.27	11.32	8.81	8.17	7.65	8.10	11.31	11.62	10.46	11.44	9.99	10.58	9.35	10.62
Variable of interest (y)	All vars.	12.40	11.75	10.19	10.90	8.85	8.49	7.60	8.23	10.74	11.00	10.39	10.71	9.24	9.81	9.28	9.59
	Boruta	12.23	11.65	10.49	11.22	8.92	8.46	7.89	8.42	10.66	10.80	10.65	10.69	9.57	9.76	9.53	9.72
	CFS	11.04	10.91	10.49	10.88	8.30	8.16	7.86	8.16	10.38	10.36	10.51	10.29	9.41	9.39	9.40	9.29
	Chi-sq.	10.64	10.63	10.21	10.61	7.97	7.95	7.61	7.93	10.42	10.43	10.42	10.42	9.50	9.51	9.31	9.50
	Gain r.	10.79	10.75	10.25	10.71	8.10	8.04	7.63	8.01	10.30	10.28	10.50	10.22	9.34	9.32	9.44	9.22
	LASSO	11.34	10.93	10.61	10.84	8.53	8.14	7.99	8.16	10.64	10.56	10.80	10.48	9.27	9.33	9.70	9.25
	StepWise	12.53	11.75	11.38	11.69	9.41	8.71	8.66	8.92	10.90	10.84	11.06	10.55	9.54	9.62	9.98	9.27
	OneR	11.42	11.23	10.23	11.06	8.58	8.46	7.64	8.42	10.74	10.76	10.56	10.69	9.56	9.65	9.47	9.59
	RF imp.	10.78	10.68	10.20	10.61	8.04	8.00	7.61	7.94	10.42	10.45	10.41	10.46	9.51	9.53	9.30	9.55

The closer to zero a value, the less biased the mean estimate

Table 13 Effect of the estimates of population means in the real data simulation for each combination of methods, considering SRSWOR from the Internet population to obtain the nonprobability samples

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		Econ. situation in Spain "poor" or "very poor"								Personal econ. situation "poor" or "very poor"							
Indicator	Boruta	0.989	1	1.14	0.97	0.974	0.993	1.16	0.97	0.957	0.981	1.06	0.97	0.964	0.992	1.09	1.001
variable of	CFS	1.072	1.023	0.97	1.05	1.02	0.99	0.97	0.992	1.054	1.038	0.95	1.05	0.974	1.017	0.99	1.052
inclusion in	Chi-sq.	0.971	0.966	0.98	0.96	0.946	0.954	0.98	0.941	0.942	0.962	0.99	0.97	0.945	0.984	1.02	1.004
$s_v\ (z)$	Gain r.	1.112	1.04	0.96	1.09	1.054	0.999	0.96	1.02	1.103	1.074	0.94	1.09	0.99	1.034	0.98	1.074
	LASSO	0.989	0.98	1.04	0.97	0.96	0.957	1.05	0.934	0.953	0.982	1.01	0.98	0.964	1.005	1.05	1.025
	StepWise	1.013	1.001	1.13	1.00	0.978	0.975	1.14	0.97	0.956	0.987	1.06	0.97	0.968	1.008	1.10	1.02
	OneR	0.93	0.938	0.98	0.95	0.916	0.933	0.98	0.928	0.895	0.933	1.01	0.95	**0.903***	0.949	1.03	0.971
	RF imp.	0.998	0.977	0.98	0.99	0.966	0.954	0.98	0.947	0.975	0.994	0.96	1.00	0.944	0.991	1.00	1.008
Variable of	Boruta	0.994	0.996	1.03	1.02	0.981	0.975	1.02	0.994	1	0.994	1.04	0.98	1.002	1.006	1.02	1.002
interest (y)	CFS	0.935	1	1.08	1.05	0.905	0.968	1.07	0.977	0.956	0.982	1.07	0.99	**0.896***	0.93	1.07	0.94
	Chi-sq.	**0.861***	0.946	1.01	0.94	**0.854***	0.928	1.00	0.902	0.946	0.965	1.04	0.96	**0.876***	0.913	1.05	0.922
	Gain r.	0.906	0.968	0.97	1.00	0.889	0.945	0.98	0.948	0.945	0.993	0.98	0.97	**0.873***	0.932	1.01	0.943
	LASSO	**0.86***	0.953	1.01	0.97	**0.869***	0.935	1.00	0.928	0.9	0.94	0.96	0.93	**0.795***	**0.856***	0.99	**0.871***
	StepWise	1.045	1.052	1.16	1.06	1.001	1.036	1.15	1.007	1.038	1.033	1.11	1.01	0.999	1.018	1.15	1.008
	OneR	**0.83***	0.925	0.98	0.95	**0.822***	0.903	0.98	0.89	0.915	0.951	0.98	0.92	**0.837***	**0.886***	1.01	**0.885***
	RF imp.	0.983	1	1.04	1.00	0.989	1.005	1.04	0.988	1.003	0.994	1.04	1.00	0.983	0.981	1.02	0.986
		Ideological self-positioning scale (1–10)								Central gov. management "poor" or "very poor"							
Indicator	Boruta	1.017	1.009	1.21	1.04	1.01	1.024	1.21	1.049	0.959	0.983	0.97	1.01	0.945	1.02	0.96	1.036
variable of	CFS	**0.885***	**0.906***	0.97	1.00	**0.887***	**0.886***	0.94	0.927	0.994	0.987	0.91	0.98	0.927	1.021	0.91	0.993
inclusion in	Chi-sq.	0.927	0.937	0.98	0.99	**0.889***	**0.896***	0.96	0.927	0.947	0.965	0.94	0.98	0.934	1.028	0.93	1.04
$s_v\ (z)$	Gain r.	**0.871***	**0.897***	0.97	1.00	**0.894***	**0.889***	0.94	0.932	1.012	1.011	0.91	0.99	0.932	1.024	0.91	0.971

Table 13 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	LASSO	0.952	0.958	1.03	1.01	0.921	0.935	1.00	0.958	0.952	0.967	0.96	0.97	0.925	1.011	0.97	1.017
	StepWise	0.934	0.942	1.07	1.00	0.896	0.917	1.03	0.946	0.944	0.966	0.95	0.98	0.932	1.031	0.96	1.036
	OneR	0.928	0.935	1.00	0.98	0.898	0.902	0.99	0.913	0.951	0.969	0.94	0.98	0.933	1.004	0.94	1.01
	RF imp.	**0.907***	0.918	0.99	0.98	**0.881***	**0.886***	0.96	0.916	0.969	0.983	0.91	0.98	0.926	1.023	0.93	1.022
Variable of	Boruta	1.266	1.359	0.96	1.50	1.4	1.436	0.94	1.49	0.957	0.979	0.95	0.98	0.881	0.92	0.94	0.939
interest (y)	CFS	**0.783***	**0.837***	0.96	0.94	**0.796***	**0.813***	0.94	**0.861***	0.962	0.969	0.93	0.94	**0.803***	**0.829***	0.92	**0.809***
	Chi-sq.	1.22	1.204	1.23	1.31	1.336	1.247	1.25	1.301	0.975	0.955	0.93	0.94	**0.84***	**0.842***	0.91	0.86
	Gain r.	1.271	1.355	0.98	1.50	1.41	1.427	0.96	1.49	0.966	0.953	0.89	0.93	**0.799***	**0.811***	0.89	**0.798***
	LASSO	1.18	1.161	1.22	1.23	1.289	1.199	1.21	1.211	0.959	0.945	0.95	0.93	**0.828***	**0.842***	0.94	**0.844***
	StepWise	1.102	1.076	1.10	1.10	1.177	1.103	1.07	1.101	0.977	0.978	1.02	0.96	0.871	0.909	1.00	0.906
	OneR	1.243	1.237	1.26	1.36	1.371	1.292	1.29	1.344	0.968	0.953	0.93	0.92	**0.821***	**0.831***	0.93	**0.814***
	RF imp.	1.201	1.218	1.14	1.31	1.307	1.265	1.16	1.302	1.022	1.005	0.96	0.98	0.971	0.972	0.95	0.961
		Preference for a state without autonomous comm.								Feels only Spanish							
Indicator	Boruta	0.992	0.983	1.09	1.00	0.976	0.962	1.08	0.955	1.019	0.996	1.00	1.00	1.009	1.002	1.01	1.027
variable of	CFS	0.956	0.966	0.97	1.05	0.987	0.941	0.95	0.944	1.014	1.026	0.96	1.07	1.042	1.057	0.95	1.082
inclusion in	Chi-sq.	0.985	0.977	0.99	1.01	0.965	0.936	0.96	0.93	1.031	1.013	0.96	1.04	1.006	1.008	0.95	1.052

Table 13 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
s_v (z)	Gain r.	0.953	0.967	0.96	1.07	1.008	0.95	0.94	0.969	0.986	1.019	0.96	1.06	1.048	1.067	0.94	1.087
	LASSO	0.986	0.977	1.01	1.02	0.969	0.937	0.98	0.935	1.039	1.012	0.99	1.05	1.013	1.017	0.98	1.07
	StepWise	0.976	0.971	1.03	1.01	0.951	0.926	0.99	0.921	1.021	1.009	0.99	1.03	1	1.013	0.99	1.051
	OneR	0.956	0.955	1.01	0.99	0.935	0.914	0.99	0.903	1.024	0.999	0.97	1.02	0.998	0.993	0.95	1.023
	RF imp.	0.968	0.964	0.98	1.01	0.962	0.926	0.96	0.921	1.032	1.022	0.97	1.04	1.027	1.03	0.96	1.06
Variable of interest (y)	Boruta	0.98	0.991	1.03	1.03	1	0.991	1.02	0.989	0.991	0.981	1.01	1.00	1.025	1.006	1.00	1.02
	CFS	**0.847***	0.906	1.02	0.99	0.915	0.947	1.00	0.949	0.947	0.928	0.97	0.95	0.992	0.947	0.95	0.961
	Chi-sq.	**0.798***	**0.864***	0.97	0.95	**0.871***	0.909	0.95	0.909	0.95	0.929	0.96	0.96	1.001	0.95	0.94	0.975
	Gain r.	**0.814***	**0.879***	0.98	0.96	0.883	0.919	0.95	0.921	0.939	0.92	0.97	0.94	0.987	0.939	0.96	0.954
	LASSO	**0.863***	**0.894***	1.04	0.96	0.93	0.935	1.01	0.924	0.971	0.939	1.03	0.96	0.987	0.938	1.02	0.959
	StepWise	1.01	1	1.17	1.09	1.057	1.024	1.14	1.051	1.014	0.982	1.09	0.99	1.023	0.976	1.08	0.984
	OneR	**0.881***	0.928	0.99	1.00	0.949	0.971	0.97	0.973	0.966	0.948	1.00	0.96	0.977	0.941	0.99	0.96
	RF imp.	**0.815***	**0.873***	0.96	0.95	0.881	0.917	0.94	0.907	0.95	0.933	0.96	0.96	1.003	0.956	0.94	0.98

Values greater than one indicate inefficiency, while values below one suggest that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used. Values in **bold** indicate that the hypothesis of the effect being equal or greater than 1 can be rejected ($\alpha = 0.05$) for that combination of methods

*Reject the null hypothesis that the effect is equal or greater than 1 for this combination of methods ($\alpha = 0.05$)

Table 14 Mean relative bias of the estimates of population means in the real data simulation for each combination of methods, considering inclusion probabilities proportional to age in the nonprobability sample

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		Econ. situation in Spain "poor" or "very poor"								Personal econ. situation "poor" or "very poor"							
Indicator variable of inclusion in s_v (z)	All vars.	8.90	8.11	7.16	8.44	9.56	8.28	7.73	9.82	5.20	4.80	6.17	5.81	26.33	24.48	22.53	26.63
	Boruta	9.04	8.37	8.13	8.38	9.32	8.38	8.89	9.15	5.18	4.85	6.82	5.46	26.19	24.99	24.39	26.42
	CFS	9.24	8.55	7.54	8.46	9.05	8.10	8.95	8.14	4.85	4.74	7.33	4.85	25.50	25.20	24.56	25.57
	Chi-sq.	8.00	7.91	7.26	7.86	7.82	7.77	7.88	7.77	3.36	3.45	6.50	3.38	22.79	22.77	22.83	22.77
	Gain r.	10.05	8.48	7.25	8.82	10.32	8.27	7.94	8.77	8.96	8.22	6.54	8.49	26.75	25.90	22.95	26.41
	LASSO	8.99	8.43	7.87	8.40	9.02	8.23	9.13	8.49	5.08	4.73	7.41	5.05	25.91	24.98	25.09	25.73
	StepWise	9.14	8.56	7.70	8.57	8.90	8.21	8.72	8.50	4.74	4.60	6.96	4.82	25.15	24.78	24.18	25.24
	OneR	7.95	7.91	7.24	7.85	7.78	7.80	7.90	7.79	3.36	3.39	6.44	3.30	22.87	22.82	22.71	22.76
	RF imp.	7.97	7.89	7.21	7.86	7.79	7.75	7.74	7.73	3.31	3.42	6.36	3.40	22.67	22.69	22.58	22.69
Variable of interest (y)	All vars.	8.90	8.13	7.16	8.38	9.56	8.35	7.73	9.78	5.20	4.68	6.19	5.83	26.33	24.34	22.57	26.68
	Boruta	8.63	8.01	7.54	8.25	8.86	8.00	8.18	8.85	5.69	5.45	6.43	6.25	26.24	24.88	22.99	26.61
	CFS	8.32	7.84	7.34	8.02	8.56	8.04	7.77	8.35	5.36	5.09	7.30	4.97	25.29	24.21	24.13	24.61
	Chi-sq.	7.75	7.45	7.29	7.60	8.26	7.84	7.78	8.37	5.95	5.77	6.55	5.61	24.74	24.05	23.15	24.29
	Gain r.	8.10	7.75	7.23	7.83	8.53	8.12	7.74	8.21	5.62	5.42	6.82	5.36	24.97	24.44	23.52	24.37
	LASSO	8.09	7.70	7.55	7.78	8.41	8.13	8.18	8.43	6.41	6.26	6.53	6.30	23.18	22.94	22.88	22.96
	StepWise	9.22	8.35	7.95	8.59	9.28	8.35	8.49	9.16	6.65	6.08	7.51	6.00	26.67	25.40	24.95	25.56

Table 14 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	OneR	7.66	7.47	7.22	7.48	8.06	7.78	7.76	8.02	5.18	5.07	6.29	5.29	23.37	22.96	22.79	23.20
	RF imp.	8.67	8.01	7.20	8.28	9.38	8.24	7.91	9.76	5.35	5.05	6.09	5.75	25.67	24.04	22.38	25.84
		Ideological self-positioning scale (1–10)								Central gov. management "poor" or "very poor"							
Indicator variable of inclusion in s_v (z)	All vars.	3.44	3.40	3.03	3.39	2.16	2.22	2.02	2.17	13.07	12.55	8.85	12.12	4.06	3.75	2.83	4.94
	Boruta	3.53	3.45	3.29	3.44	2.27	2.31	2.20	2.30	13.39	12.79	10.03	12.36	3.50	3.95	3.13	4.51
	CFS	3.44	3.36	3.00	3.35	2.14	2.12	1.87	2.14	13.32	12.77	8.77	12.54	3.56	4.70	3.37	5.06
	Chi-sq.	3.31	3.29	3.01	3.29	2.00	2.01	1.99	2.00	13.27	13.11	8.73	13.09	3.19	3.22	3.19	3.27
	Gain r.	3.37	3.25	3.01	3.29	2.26	2.20	1.97	2.24	9.27	8.90	8.74	8.88	4.00	5.37	3.25	5.30
	LASSO	3.50	3.43	3.14	3.42	2.25	2.26	1.99	2.31	13.39	12.96	9.27	12.70	3.24	3.68	3.62	4.02
	StepWise	3.43	3.36	3.12	3.35	2.14	2.13	2.03	2.12	13.26	12.81	9.16	12.58	3.70	4.48	3.46	4.94
	OneR	3.31	3.30	3.01	3.30	2.02	2.02	1.97	2.03	13.30	13.21	8.73	13.18	3.04	3.02	3.21	3.01
	RF imp.	3.30	3.29	3.01	3.29	2.00	2.00	2.00	2.00	13.23	13.10	8.71	13.09	3.21	3.26	3.18	3.25
Variable of interest (y)	All vars.	3.44	3.39	3.03	3.39	2.16	2.20	2.03	2.18	13.07	12.56	8.88	12.15	4.06	3.88	2.83	5.06
	Boruta	4.50	4.49	3.02	4.46	3.66	3.64	2.01	3.60	12.47	12.13	8.93	11.82	2.56	2.88	3.22	3.21
	CFS	3.10	3.10	2.99	3.11	2.04	2.05	1.96	2.05	13.88	13.28	9.60	13.32	0.39	0.85	2.33	0.91
	Chi-sq.	4.20	4.04	3.33	4.04	3.33	3.16	2.48	3.13	11.82	11.50	9.09	11.31	2.09	2.17	2.97	2.65
	Gain r.	4.49	4.46	3.05	4.44	3.65	3.62	2.06	3.59	12.11	11.87	9.04	11.87	0.77	1.00	2.80	1.01
	LASSO	4.11	3.93	3.31	3.90	3.25	3.05	2.38	3.01	10.90	10.69	9.06	10.76	2.01	2.16	2.85	2.24
	StepWise	3.84	3.67	3.17	3.62	2.93	2.77	2.16	2.71	13.41	12.76	9.49	12.82	1.40	1.81	2.73	2.06
	OneR	4.28	4.11	3.37	4.10	3.45	3.24	2.53	3.22	12.05	11.86	8.76	11.82	0.75	0.95	3.16	0.99

Table 14 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	RF imp.	4.13	3.99	3.29	4.03	3.20	3.08	2.42	3.11	11.39	11.11	8.86	10.82	3.84	3.69	2.76	4.16
		Preference for a state without autonomous comm.								Feels only Spanish							
Indicator variable of inclusion in s_v (z)	All vars.	17.69	17.46	15.17	17.43	9.45	9.17	8.92	9.56	20.11	20.15	17.12	19.60	9.17	9.96	9.29	8.99
	Boruta	17.95	17.68	15.87	17.55	9.59	9.24	9.01	9.48	20.54	20.45	18.00	19.84	9.68	10.18	9.54	9.67
	CFS	17.97	17.62	15.09	17.53	9.47	8.82	8.56	8.72	20.88	20.65	17.57	20.42	10.20	10.80	9.75	10.88
	Chi-sq.	17.24	17.17	14.98	17.16	8.57	8.53	8.51	8.56	20.50	20.43	17.19	20.41	9.55	9.60	9.50	9.65
	Gain r.	17.07	16.94	14.98	17.08	10.14	9.23	8.49	9.38	18.69	19.11	17.25	19.09	10.75	11.48	9.57	11.51
	LASSO	17.87	17.58	15.54	17.57	9.68	9.25	8.83	9.38	20.62	20.46	17.96	20.18	9.83	10.29	9.99	10.26
	StepWise	17.85	17.59	15.42	17.52	9.27	8.81	8.75	8.87	20.68	20.55	17.67	20.22	9.97	10.53	9.70	10.39
	OneR	17.22	17.19	14.96	17.19	8.66	8.68	8.51	8.71	20.50	20.44	17.19	20.43	9.55	9.56	9.53	9.54
	RF imp.	17.22	17.15	14.96	17.16	8.54	8.51	8.52	8.51	20.49	20.41	17.17	20.43	9.55	9.58	9.53	9.56
Variable of interest (y)	All vars.	17.69	17.42	15.18	17.46	9.45	9.19	8.90	9.60	20.11	20.21	17.09	19.61	9.17	10.01	9.24	9.12
	Boruta	17.33	17.01	15.39	16.97	9.65	9.17	9.10	9.42	19.18	19.07	17.69	18.71	10.23	10.61	9.92	9.95
	CFS	16.08	16.00	15.16	15.96	9.31	9.19	8.75	9.15	18.28	18.25	17.34	18.12	10.33	10.32	9.74	10.14
	Chi-sq.	15.56	15.54	14.96	15.55	8.84	8.82	8.52	8.84	18.22	18.21	17.19	18.21	10.40	10.40	9.55	10.38
	Gain r.	15.84	15.81	15.01	15.80	9.10	9.03	8.56	9.00	17.97	17.96	17.25	17.84	10.11	10.09	9.64	9.98
	LASSO	16.49	16.17	15.32	16.02	9.72	9.32	8.76	9.22	19.50	19.35	17.74	18.97	9.81	9.85	9.62	9.44
	StepWise	17.84	17.19	16.09	16.96	10.61	9.80	9.34	9.98	19.92	19.67	18.05	19.29	10.04	10.19	10.01	9.77
	OneR	16.83	16.67	15.02	16.64	9.77	9.62	8.63	9.69	18.57	18.48	17.35	18.37	10.17	10.16	9.67	10.08
	RF imp.	15.61	15.61	14.95	15.62	8.85	8.82	8.52	8.89	18.22	18.26	17.17	18.20	10.39	10.43	9.52	10.37

The closer to zero a value, the less biased the mean estimate

Table 15 Effect of the estimates of population means in the real data simulation for each combination of methods, considering inclusion probabilities proportional to age in the nonprobability sample

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
		Econ. situation in Spain "poor" or "very poor"								Personal econ. situation "poor" or "very poor"							
Indicator	Boruta	0.986	1.02	1.12	0.981	0.958	1.008	1.117	0.922	1.004	0.994	1.00	1.008	0.989	1.019	1.08	0.993
variable of	CFS	1.002	1.039	0.98	0.991	0.931	1.001	1.071	**0.88***	0.994	1.03	0.93	0.992	0.96	1.044	1.066	0.965
inclusion in	Chi-sq.	**0.848***	0.945	0.94	0.895	**0.817***	0.924	0.965	**0.773***	0.978	0.995	0.90	0.96	**0.841***	0.918	0.984	**0.82***
s_v (z)	Gain r.	1.102	1.027	0.94	1.017	1.046	1.034	0.971	0.901	1.038	1.04	0.90	1.023	1.024	1.095	0.979	0.998
	LASSO	0.984	1.033	1.05	0.982	0.94	1.007	1.102	0.884	0.994	1.012	0.98	1.003	0.976	1.026	1.118	0.963
	StepWise	0.999	1.048	1.01	1	0.924	1.013	1.067	0.885	0.98	1.012	0.95	0.986	0.945	1.019	1.071	0.931
	OneR	**0.844***	0.944	0.94	0.898	**0.814***	0.925	0.977	**0.778***	0.975	0.988	0.89	0.964	**0.841***	0.915	0.973	**0.814***
	RF imp.	**0.844***	0.941	0.94	0.895	**0.814***	0.922	0.955	**0.771***	0.976	0.994	0.89	0.967	**0.835***	0.914	0.967	**0.815***
Variable of	Boruta	0.967	1.005	1.03	0.99	0.94	0.992	1.034	0.918	0.986	0.995	1.00	0.992	0.997	1.024	1.024	0.997
interest (y)	CFS	0.928	0.976	1.04	0.965	0.902	0.966	1.03	**0.856***	0.985	1.001	1.03	1.003	0.935	0.971	1.073	**0.891***
	Chi-sq.	**0.874***	0.928	1.00	0.921	0.888	0.962	1.001	**0.881***	0.989	1.01	0.94	0.988	**0.914***	0.973	1.003	**0.87***
	Gain r.	0.884	0.943	0.96	0.921	0.891	0.965	0.966	**0.83***	0.982	1.01	0.94	0.988	0.923	0.985	1.026	**0.883***
	LASSO	**0.878***	0.938	1.02	0.91	**0.881***	0.97	1.026	**0.845***	0.945	0.982	0.90	0.943	**0.85***	0.922	0.977	**0.822***
	StepWise	1.041	1.053	1.12	1.042	0.969	1.025	1.096	0.938	1.053	1.064	1.01	1.032	1.003	1.048	1.109	0.941
	OneR	**0.815***	0.892	0.94	**0.864***	**0.837***	0.925	0.962	**0.801***	0.941	0.972	0.92	0.944	**0.863***	0.923	0.99	**0.838***
	RF imp.	0.968	0.984	1.00	0.979	0.986	0.991	0.992	0.991	0.992	1.015	1.00	0.987	0.969	0.985	0.993	0.959
		Ideological self-positioning scale (1–10)								Central gov. management "poor" or "very poor"							
Indicator	Boruta	1.031	1.022	1.17	1.017	1.029	1.04	1.166	1.035	1.025	1.019	1.16	1.02	0.945	1.027	1.126	0.956
variable of	CFS	0.967	0.961	0.97	0.957	0.892	0.915	**0.872***	**0.864***	1.007	1.012	0.95	1.038	0.896	1.088	0.979	0.96
inclusion in	Chi-sq.	**0.902***	0.928	0.97	0.928	**0.82***	**0.834***	0.924	**0.769***	1.002	1.052	0.95	1.099	0.854	0.904	0.949	**0.773***
s_v (z)	Gain r.	0.929	**0.905***	0.97	0.925	0.946	0.936	0.922	0.891	**0.632***	**0.646***	0.95	**0.675***	0.924	1.119	0.967	0.944

Table 15 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	LASSO	1.01	1.004	1.07	1.003	0.979	1.014	0.996	0.992	1.022	1.038	1.04	1.062	0.893	0.996	1.102	0.918
	StepWise	0.964	0.962	1.04	0.958	0.908	0.925	0.995	**0.871***	1.006	1.019	1.00	1.043	0.931	1.075	1.069	0.952
	OneR	**0.903***	0.932	0.97	0.934	**0.832***	**0.848***	0.912	**0.792***	1.007	1.063	0.95	1.11	**0.843***	0.892	0.961	**0.76***
	RF imp.	**0.901***	0.926	0.98	0.929	**0.823***	**0.836***	0.932	**0.771***	0.999	1.052	0.95	1.098	**0.854***	0.91	0.951	**0.773***
Variable of	Boruta	1.587	1.635	0.97	1.611	1.893	1.911	0.933	1.715	0.927	0.951	0.99	0.937	0.916	0.958	0.983	0.85
interest (y)	CFS	**0.81***	**0.84***	0.96	**0.839***	**0.845***	**0.869***	0.917	**0.798***	1.081	1.086	1.08	1.132	0.862	0.891	1.015	**0.8***
	Chi-sq.	1.424	1.375	1.17	1.378	1.68	1.578	1.24	1.437	**0.884***	0.908	1.00	0.926	0.9	0.907	0.974	**0.833***
	Gain r.	1.581	1.624	1.00	1.605	1.883	1.903	0.965	1.714	0.909	0.943	0.99	0.974	**0.843***	0.874	0.975	**0.738***
	LASSO	1.369	1.312	1.17	1.287	1.624	1.514	1.19	1.355	**0.762***	**0.788***	0.99	**0.829***	**0.838***	0.862	0.984	**0.746***
	StepWise	1.214	1.158	1.10	1.129	1.4	1.308	1.1	1.201	1.038	1.038	1.08	1.081	0.886	0.923	1.029	**0.817***
	OneR	1.464	1.418	1.19	1.413	1.754	1.656	1.275	1.501	**0.887***	0.93	0.96	0.954	**0.835***	0.886	0.994	**0.726***
	RF imp.	1.379	1.348	1.15	1.371	1.599	1.525	1.197	1.431	**0.849***	**0.871***	0.99	**0.871***	0.952	0.966	0.98	0.866
		Preference for a state without autonomous comm.								Feels only Spanish							
Indicator	Boruta	1.015	1.015	1.08	1.009	0.968	0.989	1.012	0.948	1.032	1.022	1.10	1.02	1.017	1.017	1.066	1.042
variable of	CFS	1.015	1.01	0.97	1.005	0.917	0.947	0.897	**0.848***	1.056	1.038	1.03	1.065	1.021	1.054	0.984	1.069
inclusion in	Chi-sq.	0.945	0.963	0.96	0.968	**0.822***	0.876	0.883	**0.795***	1.023	1.018	0.99	1.063	0.958	0.932	0.937	0.935
s_v (z)	Gain r.	**0.921***	0.937	0.96	0.957	0.987	0.996	0.886	0.92	**0.867***	**0.902***	0.99	0.943	1.071	1.131	0.941	1.136

Table 15 continued

		Raking = No				Raking = Yes				Raking = No				Raking = Yes			
		LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN	LR	GBM	kNN	NN
	LASSO	1.01	1.009	1.03	1.012	0.97	0.993	0.96	0.939	1.036	1.024	1.08	1.045	1.015	1.015	1.049	1.043
	StepWise	1.004	1.005	1.01	1.004	0.907	0.933	0.955	0.885	1.039	1.03	1.04	1.047	1.004	1.035	0.996	1.047
	OneR	0.943	0.965	0.96	0.971	**0.826***	0.882	0.88	**0.8***	1.023	1.019	0.99	1.066	0.961	0.934	0.937	0.933
	RF imp.	0.943	0.961	0.96	0.968	**0.819***	**0.868***	0.882	**0.782***	1.022	1.016	0.98	1.066	0.956	0.93	0.935	0.928
Variable of interest (y)	Boruta	0.964	0.962	1.01	0.958	0.995	1.006	0.992	0.956	**0.924***	**0.912***	1.06	0.929	1.057	1.047	1.044	1.006
	CFS	**0.851***	**0.869***	0.98	**0.864***	0.924	0.98	0.906	**0.861***	**0.847***	**0.842***	1.01	**0.876***	1.054	1.02	0.973	1.002
	Chi-sq.	**0.802***	**0.824***	0.96	**0.824***	0.879	0.944	0.884	**0.834***	**0.837***	**0.831***	0.99	**0.877***	1.062	1.02	0.944	1.017
	Gain r.	**0.826***	**0.849***	0.96	**0.845***	0.902	0.964	0.887	**0.847***	**0.822***	**0.817***	1.00	**0.851***	1.036	0.997	0.952	0.986
	LASSO	**0.877***	**0.875***	1.00	**0.861***	0.949	0.986	0.938	**0.869***	0.945	0.932	1.06	0.954	1.032	0.988	1.03	1.004
	StepWise	1.013	0.98	1.10	0.959	1.068	1.066	1.077	0.966	0.988	0.962	1.11	0.983	1.052	1.014	1.134	1.007
	OneR	**0.908***	**0.92***	0.97	**0.913***	0.976	1.028	0.923	0.929	**0.867***	**0.856***	1.01	**0.893***	1.041	1.009	0.969	1.004
	RF imp.	**0.81***	**0.833***	0.96	**0.832***	0.883	0.947	0.89	**0.851***	**0.838***	**0.837***	0.99	**0.877***	1.058	1.023	0.942	1.013

Values greater than one indicate inefficiency, while values below one suggest that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used. Values in **bold** indicate that the hypothesis of the effect being equal or greater than 1 can be rejected ($\alpha = 0.05$) for that combination of methods

*Reject the null hypothesis that the effect is equal or greater than 1 for this combination of methods ($\alpha = 0.05$)

References

Austin PC (2008) A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. Stat Med 27(12):2037–2049

Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 46(3):399–424

Austin PC, Stuart EA (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 34(28):3661–3679

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

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519

Borodovsky JT, Marsch LA, Budney AJ (2018) Studying cannabis use behaviors with Facebook and web surveys: methods and insights. JMIR Public Health Surv 4(2):e48

Breidt FJ, Opsomer JD (2017) Model-assisted survey estimation with modern prediction techniques. Stat Sci 32(2):190–205

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth, Belmont

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

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

Buskirk TD, Kolenikov S (2015) Finding respondents in the forest: a comparison of logistic regression and random forest models for response propensity weighting and stratification. Survey nsights: methods from the field, weighting: practical issues and 'how to' approach

Castro-Martín L, Rueda MM, Ferri-García R (2020) Estimating general parameters from non-probability surveys using propensity score adjustment. Mathematics 8(11):1–14

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

Chen JKT, Valliant RL, Elliott MR (2019) Calibrating non probability surveys to estimated control totals using LASSO, with an application to political polling. J R Stat Soc Ser C Appl Stat 68(3):657–681

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

Couper M (2000) Web surveys: a review of issues and approaches. Public Opin Quart 64(4):464–494

Couper M, Kapteyn A, Schonlau M, Winter J (2007) Noncoverage and non-response in an internet survey. Soc Sci Res 36:131–148

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

Deville JC, Särndal CE, Sautory O (1993) Generalized raking procedures in survey sampling. J Am Stat Assoc 88(423):1013–1020

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

Ferri-Garca R, Castro-Martín L, Rueda MM (2020) Evaluating machine learning methods for estimation in online surveys with superpopulation modeling. Math Comput Simulat 186:19–28

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

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

Gossop M, Darke S, Griffiths P, Hando J, Powis B, Hall W, Strang J (1995) The Severity of Dependence Scale (SDS): psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. Addiction 90(5):607–614

Hall MA (1999) Correlation-based feature selection for machine learning. Dissertation, University of Waikato, Department of Computer Science

Hesterberg T (2015) Resample: resampling functions. R package version 0.4. https://CRAN.R-project.org/package=resample

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

Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11(1):63–90

Kuhn M (2018) Caret: classification and regression training. R package version 6.0-81. https://CRAN.R-project.org/package=caret

Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Stat Softw 36(11):1–13

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

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

Legleye S, Karila L, Beck F, Reynaud M (2007) Validation of the CAST, a general population Cannabis Abuse Screening Test. J Subst Abuse 12(4):233–242

Marken S (2018) Still listening: the state of telephone surveys. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx. Accessed 21 Jan 2020

Meng XL (2018) Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and The 2016 US Election. Ann Appl Stat 2:685–726

Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ (2011) Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 174(11):1213–1222

National Institute of Statistics of Spain (2018) Survey on equipment and use of information and communication technologies in households. http://www.ine.es/prensa/tich_2018.pdf. Accessed 19 Jan 2020

Nicodemus KK, Malley JD, Strobl C, Ziegler A (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinform 11(1):1–13

Olivencia-Carrión MA, Ramírez-Uclés I, Holgado-Tello F, López-Torrecillas F (2018) Validation of a Spanish questionnaire on mobile phone abuse. Front Psychol 9:621

Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, Stürmer T (2011) The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiol Drug Saf 20(6):551–559

Pedrero-Pérez E, Rodríguez-Monje MT, Gallardo-Alonso F, Fernández-Girán M, Pérez-López M, Chicharro-Romero J (2007) Validación de un instrumento para la detección de trastornos de control de impulsos y adicciones: el MULTICAGE CAD-4. Trastor Adict 9:269–278

Phipps P, Toth D (2012) Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. Ann Appl Stat 6(2):772–794

Quenouille MH (1956) Notes on bias in estimation. Biometrika 43(3/4):353–360

Quinlan JR (1993) C 4.5: Programs for machine learning. The Morgan Kaufmann Series in Machine Learning, San Mateo, CA

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106

Ranalli MG, Arcos A, Rueda MM, Teodoro A (2016) Calibration estimation in dual-frame surveys. Stat Method Appl 25(3):321–349

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

Rubin DB, Thomas N (1996) Matching using estimated propensity scores: relating theory to practice. Biometrics 52(1):249–264

Rueda MM (2019) Comments on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. Test 28(4):1077–1081

Rueda MM, Martínez S, Martínez H, Arcos A (2006) Mean estimation with calibration techniques in presence of missing data. Comput Stat Data Anal 50(11):3263–3277

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA (2009) High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 20(4):512

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

Schonlau M, van Soest A, Kapteyn A (2007) Are "Webographic" or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? Surv Res Methods 1(3):155–163

Spanish Center for Sociological Research (2019) January Barometer (study number 3238). http://www.cis.es/cis/opencm/EN/1_encuestas/estudios/ver.jsp?estudio=14442. Accessed 18 Jan 2021

Taylor H (2000) Does Internet research work? Int J Market Res 42(1):51–63

Taylor H, Bremer J, Overmeyer C, Siegel JW, Terhanian G (2001) The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 US elections. Int J Market Res 43(2):127–135

Thornton L, Batterham PJ, Fassnacht DB, Kay-Lambkin F, Calear AL, Hunt S (2016) Recruiting for health, medical or psychosocial research using Facebook: systematic review. Internet Interv 4:72–81

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58(1):267–288

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

Valliant R, Dever JA (2011) Estimating Propensity Adjustments for Volunteer Web Surveys. Sociol Method Res 40(1):105–137

Xue B, Zhang M, Browne WN (2015) A comprehensive comparison on evolutionary feature selection approaches to classification. Int J Comput Intell Appl 14(2):1550008

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856–863)