Computational methods for bias reduction in surveys

UNIVERSIDAD DE GRANADA

Doctoral Thesis

Luis Castro Martín
Thesis supervised by Prof. María del Mar Rueda García

Doctorate Program in Mathematical and Applied Statistics
University of Granada

Granada, May, 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Luis Castro Martín
ISBN: 978-84-1117-436-7
URI: http://hdl.handle.net/10481/75960

Contents

Agradecimientos

Si he logrado ver más lejos, ha sido
porque he subido a hombros de gigantes.

Isaac Newton (1675)

No puedo dejar de agradecer el entorno tan favorable que he tenido durante el desarrollo de esta tesis, sin el cual habría sido imposible llevarla a cabo.

En primer lugar y en especial a María del Mar por su incansable generosidad, paciencia, trabajo y sabiduría. Ha sido la mejor tutora que habría podido desear como ya sabrán todos los que la conocen.

También a toda la gente con la que he colaborado en el proceso (Andrés, Carmen, Ramón, Bea, Giovanna, Paqui...) que no han dudado en hacerme la vida más fácil siempre que han tenido la oportunidad. Así da gusto trabajar.

Y finalmente a mi familia y a mis amigos pero sobre todo a Laura, porque sin ella todo este esfuerzo no habría tenido sentido.

Muchas gracias a todos. Este logro es tan vuestro como mío.

Summary

Probability sampling has been a fundamental framework over time in order to carry out surveys from which reliable conclusions can be extracted and properly justified. However, the application of its basic principles is now being threatened by the surge of new technologies.

Online surveys are becoming a standard due to their ability to obtain big data in a simple, cheap and efficient manner. In contrast, the methodologies associated with these kinds of surveys are usually non-probabilistic. Often, a link with the questionnaire is publicly shared, following a snowball sampling design, implying the absence of representative design weights. This causes an important self-selection bias. Even when there is a sampling frame available, the reduced response rates associated with the lack of human interaction produce an important non-response bias. Finally, coverage biases are also common because part of the target population does not have access to some of the required mediums, whether it is an internet connection, a smartphone or some specific social network account.

Despite all these problems, their use is widely extended. Besides, the decrease over the last years in the response rates of traditional surveys has affected the viability of the alternatives. Therefore, great effort has been spent on developing techniques which allow us to reduce bias in non-probability surveys. The objective is proposing new methodologies in order to preserve the credibility of statistical studies while also making use of the advantages of new technologies.

The main proposals for this purpose are Propensity Score Adjustment, which estimates the inclusion probabilities in order to obtain some representative sample weights, and Statistical Matching, which is based on predicting and imputing the individual's responses. Both rely on an auxiliary probability sample containing some covariates in common with our non-probability sample, which includes the target variable of interest.

We contribute to the development of these techniques by proposing computational methods which significantly improve their efficacy. First, we consider their application with different advanced machine learning models, culminating in state-of-the-art techniques which optimize the results obtained. We also propose a novel method for combining Propensity Score Adjustment

and Statistical Matching, improving the bias reduction obtained with each method separately. We implement many of these methods along with other bias reduction alternatives for non-probability surveys in NonProbEst, an easy-to-use R package.

Additionally, we extend their application to more contexts. The Propensity Score Adjustment method, combined with calibration techniques, can be considered for overlapping panel surveys in order to obtain transversal as well as longitudinal estimates over time. This compensates the bias resulting from the non-response in successive measurements. In this way we propose several reliable estimators which are then applied to diverse parameters of interest in a research project about the evolution of COVID-19. We also consider a scenario in which the auxiliary probabilistic sample includes the target variable as well. An extensive comparative study is carried out with different possible strategies. The results show the benefits of the proposed methodologies.

Note: This thesis is presented as a compendium of six publications in relation with the contents of the thesis. The full version of the papers is included in Appendices A1 - A6.

Part I

Chapter 1

Introduction

The theoretical basis for probability sampling, initially established by Neyman (1934) and extended ever since by Horvitz and Thompson (1952) among others, lays the foundation for a reliable framework in order to carry out probabilistic surveys from which researchers from all fields can obtain information and draw unbiased conclusions. Over the years, this framework has been a fundamental reference for survey sampling, guaranteeing the credibility of the results obtained.

However, two factors have become extremely relevant recently affecting the viability of the classical methods. The first one is the increasing lack of response in surveys carried out with traditional methods (face-to-face or over the telephone), as analyzed by Díaz de Rada (2012), Kohut et al. (2012) and Marken (2018). Even though the low response rates greatly increase costs and jeopardize the ability of the institutions to extract enough data in order to carry out their research, this issue has been compensated by the second factor: the development of new technologies.

With the success of web technologies, researchers are able to carry out online surveys which are cheap, simple, and effective for obtaining big amounts of data. Berzofsky et al. (2018) and Brickman Bhutta (2012) propose methodologies based on the Twitter and Facebook, respectively, advertising applications which try to replicate a probability approach. Their applications excel at quickly obtaining responses but their probabilistic characteristics are very limited. Pötzschke and Braun (2017) makes use of Facebook in order to obtain data from a hard to reach population (Polish migrants in Europe). This approach has been proven to be very effective for such difficult scenarios as also analyzed by Iannelli et al. (2020). In fact, a comparison carried out by Gilligan et al. (2014) proves its advantages over traditional strategies such as social networking.

However, the major issue with these new methodologies are the inevitable changes in the reference theoretical framework which was used so far, as pointed out by Callegaro et al. (2015), Schonlau and Couper (2017) and

Díaz de Rada et al. (2019); causing important bias problems. First, the questionnaires are usually distributed with a snowball sampling design (Bethlehem, 2010). A link is publicly shared, often through social media, so there is not a sampling frame available. Therefore, a formal definition of "design weights" is difficult (Snijders, 1992) and an important self-selection bias is implied (Schonlau et al., 2009). Also, even in those cases where there is a sampling frame available, the response rates when using these procedures are even lower than those using more personal methods (Manfreda et al., 2008), in which the individuals feel more committed. This causes a significant non-response bias instead. Other difficulties associated with a lack of human assistance should also be considered, since an interviewer is able to explain concepts, clarify ambiguities and validate some responses (Anduiza and Galais, 2017; Gao et al., 2016). Finally, the use of these kinds of modern technologies is associated with coverage bias since not all the population has the same level of internet and smartphone access (Couper et al., 2018). The chosen distribution medium, such as paid advertisements in certain social networks, also implies an important coverage bias as the sampled population is then limited to their user base.

In this context, the use of online surveys is necessary even when their bias issues would invalidate the results obtained. Thus a lot of effort has been put by researchers in order to develop techniques which solve the mentioned bias problems. This work is needed if we want to preserve the confidence in statistical surveys. Beaumont (2020) carries out an extensive review on this matter.

Propensity Score Adjustment (PSA) (Lee, 2006; Lee and Valliant, 2009; Valliant, 2020) is one of the main methods proposed in order to reduce bias in such non-probabilistic surveys. It models the inclusion probability of each individual as its propensity to participate in the non-probability sample. The estimated propensity may then be used as a substitute of the inclusion probability in order to assign representative sample weights. Therefore, it is a reweighting technique which needs an auxiliary probabilistic sample with some covariates in common with our convenience sample. This auxiliary reference sample correctly represents the target population with some valid sample weights but it does not include the variable of interest. By applying PSA, it is possible to combine the information provided by both samples. This method has been thoroughly developed, including its application for the estimation of general parameters (Castro-Martín et al., 2020a) and different proposals in order to transform the propensities into weights (Valliant, 2020; Schonlau and Couper, 2017; Valliant and Dever, 2011; Lee, 2006; Lee and Valliant, 2009; Ferri-García et al., 2021a). Among them, Wang et al. (2020) proposes a novel Kernel Weighting approach which considers the distance among the estimated propensity scores, smoothed via a kernel function.

Statistical Matching is another alternative method which was proposed

by Rivers (2007) and further developed by Beaumont and Bissonnette (2011). It also uses an auxiliary probability sample with some covariates in common with our volunteer sample but without the target variable. However, instead of reweighting, it is an imputation technique. Thus the values for the target variable are imputed in the reference probability sample. A prediction model is obtained using the non-probability sample in order to carry out this imputation.

These methods can be combined with each other as well as with other classic techniques. Ferri-García and Rueda (2018) combines Propensity Score Adjustment with calibration weighting. Given some known population totals, calibration calculates new sample weights which are as close as possible to the original weights while respecting the calibration equation so the sample totals match said population totals. Therefore, the initial weights as returned by PSA can be further adjusted with this method. Chen et al. (2020) also proposes a Doubly Robust estimator which combines PSA with Statistical Matching. The potential estimation error resulting from the values imputed via Matching is weighted and corrected with the weights obtained via PSA. This proposal, which aims for a more consistent behavior of the final estimator, has been further developed by Liu et al. (2021) and Rafei et al. (2022a).

The proposed methods usually consider initially basic linear models for their application. The propensities estimated by PSA are obtained using logistic regression and the values are imputed by Matching using linear regression for numerical variables and logistic regression for categorical variables. However, some recent papers have considered the application of advanced machine learning models and algorithms with satisfactory results. Ferri-García and Rueda (2020) carries out a comparative study of some possible machine learning models for PSA. Chu and Beaumont (2019) proposes a modification of the decision tree algorithm specifically designed for estimating propensities. Kern et al. (2020) proposes Boosted Kernel Weighting, which considers boosting algorithms in order to improve the Kernel Weighting method. Integrating the predictive potential of state-of-the-art machine learning techniques into these methods, which are completely dependent on estimated values, has shown to be one of the main means in order to obtain optimal results in real use cases.

The aim of this thesis is to provide significant progress in this research field by computational methods which allow us to obtain reliable inferences, comparable to those obtained when applying classical probabilistic methodologies, while making use of the advantages offered by new technologies. In Appendix 1, the performance of Propensity Score Adjustment and Statistical Matching is tested in a complex simulation study considering several machine learning algorithms which can be used for their application. In Appendix 2, an easy-to-use R package is developed including the implementation of

a wide variety of bias reduction techniques in surveys. In Appendix 3, a novel method for combining PSA and Matching is proposed and its better performance over previous alternatives is proven in a comparative study. In Appendix 4, state-of-the-art machine learning techniques are applied to recently proposed methods in order to obtain optimal results. Their efficacy is justified in a simulation study and they are used to analyze a nonprobability survey sample on social effects of COVID-19. In Appendix 5, new weighting methods are developed in order to reduce the bias of overlapping panel surveys, allowing for accurate cross-sectional as well as longitudinal estimations in a research project on the evolution of COVID-19. In Appendix 6, several methodologies are considered for integrating probability and non-probability survey samples, both including the target variable(s), with advanced machine learning techniques.

Chapter 2

Objectives

2.1. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

The popularity of non-probabilistic surveys has led to the development of several techniques in order to compensate for the important bias problems that they imply. Many papers on Propensity Score Adjustment (Lee, 2006; Lee and Valliant, 2009; Valliant, 2020) and Statistical Matching (Rivers, 2007; Beaumont and Bissonnette, 2011) can be found in the literature. There are even some works combining them with other bias reduction techniques like calibration (Ferri-García and Rueda, 2018). However, these papers usually work with methodologies which apply simple linear models, like linear regression or logistic regression. Even though some works do introduce more advanced machine learning models (Ferri-García and Rueda, 2020), they are still limited to one main method (eg. Propensity Score Adjustment) without considering the alternatives (eg. Statistical Matching).

We carry out an exhaustive comparison at two levels. On one hand, a wide variety of possible machine learning techniques, beyond simple linear models, which can be applied with PSA as well as with Matching are considered. On the other hand, the performance improvement of PSA over Matching or vice versa is also measured.

Every possible combination is applied to a simulation with various real datasets, choosing different selection bias mechanisms for each one of them. The objective is aggregating the metrics obtained in order for the results to determine if one method is superior to the other, as well as the importance of introducing advanced machine learning models.

2.2. The R Package NonProbEst for Estimation in Non-probability Surveys

Techniques of all kinds have been proposed and developed in the literature in order to reduce the bias problems associated with non-probability surveys. Classic methods like calibration (Deville and Särndal, 1992), which considers known population totals for some auxiliary variables, are included in this category. More modern but well known methods like Propensity Score Adjustment, which takes advantage of auxiliary probabilistic samples with some covariates in common, should also be included. In the same scenario, Statistical Matching can also be applied in order to use said auxiliary probabilistic sample. If we have information on the whole population for some variables, superpopulation modeling techniques (Ferri-García et al., 2021b) like model based (Valliant et al., 2000), model assisted (Breidt et al., 2017) or model calibrated (Wu and Sitter, 2001) estimators are appropriate.

It should also be noted that those methods often have multiple variants which can be applied. For example, there are multiple formulas proposed in order to obtain the new adjusted weights from the propensities calculated in Propensity Score Adjustment (Valliant, 2020; Schonlau and Couper, 2017; Valliant and Dever, 2011; Lee and Valliant, 2009). We have also analyzed at this point (Castro-Martín et al., 2020b) the importance of working with different machine learning models, depending on the type of selection bias and the type of data we are dealing with.

Having such a wide and complex range of possibilities may be difficult for researchers trying to reduce the bias of a real non-probabilistic sample. The objective of the NonProbEst package is offering them easy-to-use implementations in a common programming language like R so estimations for key parameters like totals, means, variances or confidence intervals can be obtained in a simple manner. Therefore, state-of-the-art techniques can be applied without the need for excessively advanced knowledge or complex implementations.

2.3. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

The comparisons between Statistical Matching and Propensity Score Adjustment do not determine a clear winner at reducing bias from non-probability surveys estimations (Castro-Martín et al., 2020b). The conclusions rather indicate that the performance of each method depends on the kind of data and the selection mechanism for each specific case. Given the difficulty of estimating a priori which technique will perform better in a real

case scenario, where the metrics calculated at the simulations cannot be obtained, researchers look for a combination of PSA and Matching able to reduce the bias properly in every case.

A first approximation for this purpose would be shrinkage (Copas, 1983, 1993; Arcos et al., 2014), which applies a linear combination of both methods. The coefficients of this linear combination are determined by the estimated variance of each estimator. Chen et al. (2020) proposes a Doubly Robust estimator which corrects the error of the Matching estimator by using the weights obtained with PSA. Its results already show that this proposal is able to produce optimal estimations when PSA would be the best choice as well as when Matching would be the best choice.

However, previously proposed techniques apply PSA and Matching independently before combining the results. Our study proposes a new estimator which takes advantage of the very nature of the machine learning algorithms implied in the process in order to obtain a deeper combination of both methods. The objective is proving that using the weights produced by PSA when training the Matching model leads to a significantly more optimal estimator than the doubly robust estimator. Both alternatives are considered in a thorough comparative study which includes simulations with several real datasets and different selection mechanisms.

2.4. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

The development of bias reduction methods for non-probability surveys inevitably led to considering cutting edge machine learning techniques on their application (Kern et al., 2020). The objective of this paper is to test how much is XGBoost (Chen and Guestrin, 2016), proven to be one of the most effective machine learning algorithms for this kind of tabular data, combined with other advanced techniques like hyperparameter optimization via the Tree-structured Parzen Estimator algorithm (Bergstra et al., 2011, 2013), able to improve the results obtained so far.

To that end, an in-depth comparative study which shows its efficiency with respect to the most recently developed methods is carried out. It includes the TrIPW estimator (Chu and Beaumont, 2019) which applies PSA with a modification of the classical machine learning algorithm Decision Tree, specifically adapted for estimation with non-probability samples using an auxiliary probability sample. We also consider Kernel Boosting (Wang et al., 2020), which proposes an innovative way of obtaining adjusted weights from the propensities calculated via PSA. This improvement can also be combined with the use of gradient boosting algorithms like XGBoost. Lastly, techni-

ques which combine PSA with Statistical Matching like the Doubly Robust Estimator (Chen et al., 2020) or the Training Estimator (Castro-Martín et al., 2022) are also included in the study. Again, since these proposals can be applied with any machine learning model, its combination with XGBoost is also considered.

Said study consists of three parts. The first one replicates the simulation carried out by Chen et al. (2020), determining a reliable reference of the potential of the new techniques proposed. Once this potential has been proven, a more complex simulation with real datasets is carried out. Finally, we show their utility in a real application to the ESPACOV (Serrano del Rosal et al., 2020), a survey which provides information on the social effects of COVID-19 in Spain.

2.5. Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

Panel surveys are widely used when the target is seeing the evolution of certain characteristics over time (Kalton and Citro, 1995). The Health Care and Social Survey (ESSOC) (Sánchez-Cantalejo et al., 2021) research project is the perfect example. It arises from the need to provide data on the evolution of the COVID-19 impact that can be considered when making decisions to prepare and provide an effective Public Health response in the different affected populations, especially in the most vulnerable ones. The objective of this survey is to determine the magnitude, characteristics, and evolution of the impact of COVID-19 on overall health and its socioeconomic, psychosocial, behavioral, occupational, environmental, and clinical determinants in the general population and those with greater socioeconomic vulnerability.

The ESSOC has an overlapping panel design so the same individuals are sampled at every measurement. While this design is very useful for obtaining cross-sectional as well as longitudinal estimations, the fatigue of the surveyed population when being repeatedly sampled over time causes an increasing non-response problem (Kalton et al., 1985; Lepkowski, 1989; Kalton and Brick, 1995). This lack of response introduces a bias in the results which may invalidate them. Therefore, some reliable techniques in order to correct this bias are needed.

The objective of this work is developing weighting methods for estimating totals, proportions and change or differences of a population characteristic, from overlapping panel survey data, using various combined methods such as Propensity Score Matching, machine learning and calibration. The reweighting methods are formulated based on the ESSOC structure but can be adapted to any other type of overlapping panel design.

2.6. Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain

So far we have worked with non-probability surveys which make use of auxiliary probabilistic surveys with some covariates in common in order to reduce the bias of the estimators for the target variables. However, there is another typical scenario in which a probability survey with a small sample is complemented with a non-probability sample (DiSogra et al., 2011; Robbins et al., 2021; Wiśniowski et al., 2020). This second sample benefits from popular technologies like online surveys which make it easier and cheaper to obtain big amounts of data. On the other hand, we have already established the important bias problem they imply. Since both samples contain the same variables, they should be combined in order to obtain estimations with small variance while also being reliable.

Such is the case in the Survey on the impact of the COVID-19 pandemic in Spain (ESPACOV) (Serrano del Rosal et al., 2020) which used a mixed multiphase sampling design inspired by the responsive approach (Groves and Heeringa, 2006). There were two editions of the survey addressing the opinions and attitudes of the Spanish population regarding the COVID-19 crisis, as well as the assessments of its management and its consequences, either anticipated (ESPACOV I) or endured (ESPACOV II). Both editions of the ESPACOV Survey were web based and followed a sampling design that combined the use of SMS invitations to take part in the survey, sent to a list of randomly generated mobile phone numbers (probability-based sample), with the publication of Facebook, Instagram and Google Ads segmented to purposely oversample the socio-demographic profiles that were underrepresented in the probability-based sample (non-probability sample). An in-depth explanation and justification of this methodology is provided in Rinken et al. (2020).

Our objective is proposing a methodology which, with the help of machine learning techniques, integrates probability and non-probability samples. The efficacy of the estimators is evaluated in a simulation study which includes a comparison with other alternative proposals (DiSogra et al., 2011; Elliott and Haviland, 2007; Robbins et al., 2021). Lastly, they are applied to the second wave of the Survey on the impact of the COVID-19 pandemic in Spain, allowing us to conclude that the estimation method we propose is the best option for reducing observed biases in the data.

Chapter 3

Methodology

3.1. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. Let s_V be a convenience (or volunteer) nonprobability sample of size n_V. Let y be the variable of interest in the survey estimation.

A first estimator for the population mean, $\overline{Y}$, can be the naive estimator based on the sample mean of y in s_V:

$$\hat{\overline{Y}} = \sum_{i \in s_V} \frac{y_i}{n_V} \tag{3.1}$$

In order to correct the bias implied in s_V, let s_R be a reference auxiliary sample of size n_R selected from U under a probability sampling design (s_R, p_R) with $\pi_i = \sum_{s_R \ni i} p_R(s_R)$ (where $s_R \ni i$ denotes the samples which contain the unit i) the first order inclusion probability for individual i, we denote by $d_i = 1/\pi_i$ the design weights for the units in the reference sample. Let $\mathbf{x}_i$ be the values presented by individual i for a vector of covariates $\mathbf{x}$. Those covariates are common to both samples, while we only have measurements of the variable of interest y for the individuals in the convenience sample.

The popular Propensity Score Adjustment (PSA) technique (Lee, 2006; Lee and Valliant, 2009; Valliant, 2020) is applied assuming some conditions. We assume that the selection mechanism of s_V is ignorable, this is:

$$P(\delta_i = 1 | y_i, \mathbf{x}_i) = P(\delta_i = 1 | \mathbf{x}_i), i \in s_V \tag{3.2}$$

where δ_i is the following indicator variable:

$$\delta_i = \begin{cases} 1 & i \in s_V \\ 0 & i \notin s_V \end{cases}, i = 1, 2, ..., N$$

We also assume that the selection mechanism follows a parametric model:

$$P(\delta_i = 1|\mathbf{x}_i) = p(\mathbf{x_i}) \tag{3.3}$$

where the function $p(\mathbf{x_i})$ is determined by the chosen machine learning model.

This machine learning model is trained with $\mathbf{x_i}, i \in s_V \cup s_R$ on the variable δ_i, obtaining the model $\hat{p}$. The predicted propensities to participate in the volunteer sample can be calculated as $\hat{p}_i = \hat{p}(\mathbf{x}_i)$. These are transformed into weights by inverting them: $w_i = 1/\hat{p}_i$. Thus the inverse propensity score weighting estimator (IPSW) is:

$$\hat{\overline{Y}}_{IPSW} = \frac{1}{\sum_{i \in s_V} w_i} \sum_{i \in s_V} y_i w_i \tag{3.4}$$

The propensities can also be stratified (Valliant and Dever, 2011) in a fixed number of groups, c, with the idea of grouping individuals with similar volunteering propensities. Then the average propensity within each group is calculated:

$$\overline{\pi}_c = \sum_{i \in s_R^c \cup s_V^c} \hat{p}_i(\mathbf{x})/(n_R^c + n_V^c) \tag{3.5}$$

where n_R^c and n_V^c are the number of individuals from the reference and the volunteer sample respectively that belong to the c-th group. Thus the stratified PSA estimator would be:

$$\hat{\overline{Y}}_{SPSA} = \frac{1}{\sum_{i \in s_V} w_i^{strat}} \sum_{i \in s_V} y_i w_i^{strat} \tag{3.6}$$

where $w_i^{strat} = 1/\overline{\pi}_c$ with $i \in s_V^c$.

Statistical matching (SM) (Rivers, 2007; Beaumont and Bissonnette, 2011) is an alternative model-based approach. The idea in this context is to model the relationship between the target variable y_i and the covariates $\mathbf{x}_i$ using the volunteer sample s_V in order to predict y_i for the reference sample s_R (where y has not been measured). SM assumes that y is a realization of a superpopulation random variable Y, which follows a functional relationship with the set of covariates $\mathbf{x}$ such that:

$$y_i = m(\mathbf{x}_i) + e_i,\ i = 1, 2, ..., N, \tag{3.7}$$

where the function $m(\mathbf{x_i})$ is determined by the chosen machine learning model.

In this case, the model is trained with $\mathbf{x_i}, i \in s_V$ on the variable y_i and applied to s_R, obtaining the imputed values $\hat{y}_i = \hat{m}(\mathbf{x}_i)$. The matching estimator is then given by:

$$\hat{\overline{Y}}_{SM} = \frac{1}{\sum_{i \in s_R} d_i} \sum_{s_R} \hat{y}_i d_i \tag{3.8}$$

Both PSA and SM are applied in a comparative study with the machine learning models described below.

- Generalized linear models (GLM): The most basic model applies a linear combination to the input variables with the coefficients $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in order to obtain the prediction. When covariates suffer from multicollinearity, ridge regression (McDonald, 2009) proposes an identity term to control instability: $\beta = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$. The Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996) proposes instead using a penalty parameter, α, according to the following optimization problem:

$$\begin{array}{l} argmin \sum_{i=1}^{N}(y_i - \alpha - \sum_j \beta_j \mathbf{x}_{ij})^2 \\ subject\ to \sum_j |\beta_j| \leq t \end{array} \tag{3.9}$$

  Ridge and LASSO are both considered standard penalized regression models Van Houwelingen (2001). For classification problems, like in PSA, a logistic function is applied to the output in order to transform it into a probability.

- Discriminant Analysis: It can only be used for discrete variables (classification problems). Let y be the dependent variable with K classes, π_k the probability of an individual of belonging to the kth class, $\mathbf{X}$ the matrix of covariates $n \times p$, and $f_k(\mathbf{x})$ the joint distribution of $\mathbf{x}$ conditioned to y taking the kth class. As described in James et al. (2013) Linear Discriminant Analysis (LDA) assigns an individual the class that maximizes the probability:

$$P(y_i = k|\mathbf{x} = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_i)}, k = 1, ..., K \tag{3.10}$$

- Decision trees, Bagged trees and Random Forests: Decision trees split sequentially the input data via conditional clauses until they reach a terminal node, which assigns an specific class or value. This process results in the following estimation for the expectance $E_m(y_i|\mathbf{x}_i)$:

$$E_m(y_i|\mathbf{x}_i) = \begin{cases} \overline{y(s^{J_1})} & \{i \in s/\mathbf{x}_i \in J_1\} \\ ... & ... \\ \overline{y(s^{J_k})} & \{i \in s/\mathbf{x}_i \in J_k\} \end{cases} \tag{3.11}$$

  where $\overline{y(s^{J_i})}$ denote the mean of y among the members of the sampled population, s, meeting the criteria of the i-th terminal node. Bagged trees combine this approach with bagging (Breiman, 1996), which averages the predictions of multiple decision trees trained with boostrapped subsamples of the complete dataset. The Random Forests variant (Breiman, 2001) also selects a random subset of the covariates for each decision tree forming the ensemble.

- Gradient Boosting Machine (GBM): It also works as an ensemble of *weak classifiers*. Boosting is an iterative process that trains subsequent models giving more importance to the data for which previous models failed. This idea can be interpreted as an optimization problem (Breiman, 1997) and, therefore, it is suitable for the gradient descent algorithm (Friedman, 2001). Then the estimates for y are:

$$E_m(y_i|\mathbf{x}_i) = v^T J(\mathbf{x}_i) \tag{3.12}$$

  where $J(\mathbf{x}_i)$ stands for a matrix of terminal nodes of m decision trees and v is a vector representing the weight of each tree. GBM has improved previous state-of-the-art models for some cases (Touzani et al., 2018).

- k-Nearest Neighbors: The algorithm simply averages the value of the target variable for the k individuals closer to the estimated individual (its k nearest neighbors), given a certain distance dependent on the covariates. This is:

$$E_m(y_i|\mathbf{x}_i) = \frac{\sum_{j \in s/d(\mathbf{x}_i,\mathbf{x}_j) \leq d(\mathbf{x}_i,\mathbf{x}_{(k)})} y_j}{k} \tag{3.13}$$

  where $\mathbf{x}_{(1)},\ldots,\mathbf{x}_{(n-1)}$ are, respectively, the closest and the furthest individual to $\mathbf{x}_i$.

- Naive Bayes: It can only be used to predict discrete variables. For PSA, it uses the Bayes theorem to predict the propensities as follows:

$$\hat{p}_i(\mathbf{x}_i) = \frac{P(\delta_i = 1)P(\mathbf{X} = \mathbf{x}_i|\delta_i = 1)}{P(\mathbf{X} = \mathbf{x}_i)} \tag{3.14}$$

  Its simplicity has proven to be effective under certain conditions (Ferri-García and Rueda, 2020).

- Neural networks with Bayesian Regularization: Neural networks work as universal approximators (Csáji et al., 2001) combining linear and non-linear functions. The inputs follow an iterative process through one or more hidden layers until reaching the last layer, which produces the final output. The weights are initialized randomly and then optimized via gradient descent with the backpropagation algorithm Rumelhart et al. (1986). Since overfitting is usually an important problem, prior distributions can be imposed to the weights as a regularization method (Burden and Winkler, 2008). Another option is bagging, as it is applied to decision trees. This approach is known as Model Averaged Neural Networks (Ripley, 2007).

All the methods and models are applied to the following three different populations, each considering two different sampling strategies.

- Spanish Life Conditions Survey (2012 edition) (Spanish National Institute of Statistics, 2012): It collects the data for 28,610 adults living in Spain about economic and life conditions variables. The target variable is the self-reported health on a scale from 1 to 5. The algorithms were trained using 56 variables. The first sampling strategy was a simple random sampling excluding the individuals without internet access. The second sampling strategy also included a non-linear propensity to participate in the sample using the formula $Pr(yr) = \frac{yr^2 - 1900^2}{1996^2 - 1900^2}$, where yr is the year the individual was born.

- BigLucy (Gutiérrez, 2009): It consists of various financial variables of 85,396 industrial companies of a city for a particular fiscal year. The target variable is the annual income in the previous fiscal year. The algorithms were trained using 4 variables. The first sampling strategy was simple random sampling among the companies with SPAM options that are not small companies, simulating a significant coverage bias. The second sampling strategy was simple random sampling among the companies with SPAM options including a propensity to participate calculated as $Pr(taxes) = min(taxes^2/30, 1)$ where $taxes$ is the company's income tax.

- Bank Marketing Data Set Moro et al. (2014): It includes information about 41,188 phone calls related to direct marketing campaigns of a Portuguese banking institution. The target variable is the mean contact duration. 18 variables were used for training. For the first sampling strategy, we applied simple random sampling among the clients contacted more than 3 times. For the second sampling strategy, we applied simple random sampling among the clients contacted more than twice.

Each population and sampling strategy was simulated using various sample sizes: 1000, 2000 and 5000. The same size is taken for the convenience sample and for the reference sample. For each scenario (the specific method, model, population, sampling strategy and sample size combination), 500 simulations were executed.

3.2. The R Package NonProbEst for Estimation in Non-probability Surveys

In the same context as Section 3.1, several estimators are implemented in order to reduce the bias of the naive estimator 3.1, depending on the available auxiliary information. We distinguish three different cases:

- InfoTP: Only the population totals of the auxiliary variables are known (often called control totals).

- InfoES: The values of the auxiliary variables are available for every element in a probability sample.
- InfoEP: The values auxiliary variables are available for every element in the whole population.

For the first case, InfoTP, the calibration method is implemented. Let $\mathbf{x}_i$ be the value taken on unit i by a vector of auxiliary variables which population total is assumed to be known, $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$. The calibration estimation of Y consists in the computation of a new vector of weights w_i for $i \in s_V$ which modifies as little as possible the original sample weights, $w_{Vi} = N/n_V$ in our case since the sample is non-probabilistic, respecting at the same time the calibration equation:

$$\sum_{i \in s_V} w_i \mathbf{x}_i = \mathbf{X} \tag{3.15}$$

Given a pseudo-distance $G(w_i, w_{Vi})$, the calibration process consists in finding the solution to the minimization problem

$$\min_{w_i} \{ \sum_{i \in s_V} G(w_i, w_{Vi}) \} \tag{3.16}$$

while respecting the calibration equation. Several distances as defined in Deville and Särndal (1992) are included in the package.

For the second case, InfoES, several techniques can be applied as described in Section 3.1. The Inverse Propensity Score Weighting estimator is implemented as in 3.4. In addition, other variants of Propensity Score Adjustment are included. These change the way the propensities are transformed into weights. The estimator proposed in Schonlau and Couper (2017) uses the following formula instead:

$$w_i^{PSA2} = \frac{1 - \hat{p}_i}{\hat{p}_i} \tag{3.17}$$

The stratified version 3.6 is also implemented along with the alternative proposed in Lee and Valliant (2009), which calculates an adjustment factor as

$$f_c = \frac{\sum_{i \in s_R^c} w_i / \sum_{i \in s_R} w_i}{\sum_{i \in s_V^c} w_i / \sum_{i \in s_V} w_i} \tag{3.18}$$

and the estimator is given by

$$\hat{\overline{Y}}_{SPSA2} = \frac{1}{\sum_{i \in s_V} w_i^{strat2}} \sum_{i \in s_V} y_i w_i^{strat2} \tag{3.19}$$

where $w_i^{strat2} = w_i f_c$ with $i \in s_V^c$. The variance of all these variants can be obtained with Jackknife's variance estimator (Quenouille, 1956). Let $\hat{\bar{y}} =$

$\frac{1}{N}\sum_{i\in s_V} w_i y_i$ be the estimator of the mean of y, his Leave-One-Out Jackknife estimator of the variance is given by:

$$\hat{V}(\hat{\overline{y}}) = \frac{n-1}{n}\sum_{j=1}^{n}(\overline{y}_{(j)} - \overline{y})^2 \tag{3.20}$$

where $\overline{y}_{(j)}$ is the value of the estimator $\hat{\overline{y}}$ after dropping unit j from s_V and where $\overline{y}$ is the mean of values $\overline{y}_{(j)}$. Finally, the Statistical Matching method 3.8 is also available in the package.

For the third case, InfoEP, superpopulation models are applied. These assume that the population under study $\mathbf{y} = (y_1, ..., y_N)'$ is a realization of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)'$ having a superpopulation model ξ. To incorporate auxiliary information $\mathbf{x}_i$ available for all $i \in U$, we assume that y follows a parametric model:

$$Y_k = m(\mathbf{x}_i) + e_i, \quad i = 1, ..., N. \tag{3.21}$$

where m is the chosen machine learning model and the random vector $e = (e_1, ..., e_N)'$ is assumed to have zero mean and a positive definite covariance matrix which is diagonal (Y_i are mutually independent). Let $\overline{s}_V = U - s_V$. The model is trained with $\mathbf{x_i}, i \in s_V$ on the variable y_i and applied to $\overline{s}_V$, obtaining the imputed values $\hat{y}_i = \hat{m}(\mathbf{x}_i)$. Then the following estimators are considered:

- the model-based estimator:

$$\hat{Y}_m = \sum_{i\in s_V} y_i + \sum_{i\in \overline{s}_V} \hat{y}_i \tag{3.22}$$

- the model-assisted estimator:

$$\hat{Y}_{ma} = \sum_{i\in U} \hat{y}_i + \sum_{i\in s_V} (y_i - \hat{y}_i) w_{Vi} \tag{3.23}$$

- the model-calibrated estimator:

$$\hat{Y}_{mcal} = \sum_{i\in s_V} y_i w_i^{CAL} \tag{3.24}$$

  where w_i^{CAL} are such that they minimize $\sum_{i\in s_V} G\left(w_i^{CAL}, w_{Vi}\right)$, where $G(\cdot,\cdot)$ is a particular distance function, subject to

$$\sum_{i\in s_V} w_i^{CAL}\hat{y}_i = \sum_{i\in U} \hat{y}_i.$$

For all the methods implemented, NonProbEst allows the use of a wide variety of classification and regression models by relying on caret (Kuhn, 2015), a well known machine learning package.

3.3. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

In the same context as Section 3.1, multiple techniques are proposed in order to combine the Propensity Score Adjustment estimator 3.4, which total will be referred to as $\hat{Y}_{PSA}$, and the Statistical Matching estimator 3.8, which total will be referred to as $\hat{Y}_{SM}$.

- Shrinkage:

$$\hat{Y}_{srk} = K\hat{Y}_{SM} + (1-K)\hat{Y}_{PSA} \tag{3.25}$$

  where K is a constant satisfying $0 < K < 1$. In particular, optimum value for k if s_V and s_R are independent is

$$k_{opt} = \frac{V(\hat{Y}_{PSA})}{V(\hat{Y}_{SM}) + V(\hat{Y}_{PSA})}. \tag{3.26}$$

  Since those variances are unknown, the alternatives considered are $K_1 = n_R/(n_R + n_V)$ and $K_2 = V(\hat{\theta}_{PSA})/(V(\hat{\theta}_{PSA}) + V(\hat{\theta}_{SM}))$ where $V(\hat{\theta}_{PSA})$ and $V(\hat{\theta}_{SM})$ are the variances of $\hat{Y}_{PSA}$ and $\hat{Y}_{SM}$, respectively, observed at the simulations.

- Doubly Robust estimator (Chen et al., 2020):

$$\hat{Y}_{DR} = \sum_{s_R} \hat{y}_i d_i + \sum_{s_V} w_i(y_i - \hat{y}_i), \tag{3.27}$$

  based on the idea of the difference estimator (Särndal et al., 2003), which considers the following decomposition:

$$Y = \sum_{U} \hat{y}_i + \sum_{U} (y_i - \hat{y}_i) \tag{3.28}$$

- Training data with PSA weights. Since most machine learning models allow considering weights for the training data, we also propose an estimator obtained with the following algorithm:

  - Calculate w_i for $i \in s_V$ by using some machine learning classification algorithm.
  - Train a model $E_m(y_i|\mathbf{x}_i)$ using x_i for $i \in s_V$ weighted with w_i for $i \in s_V$. Each machine learning model may have its own weighting mechanism.
  - Obtain $\hat{y_i}^{tr}$ for $i \in s_R$ by using the model trained in the previous step.

- Estimate the total as

$$\hat{Y}_{tr} = \sum_{s_R} \hat{y_i}^{tr} d_i \tag{3.29}$$

All the alternatives, along with the naive estimator 3.1, PSA 3.4 and Matching 3.8, are considered in a comparative study similar to the one conducted in Section 3.1.

3.4. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

In the same context as Section 3.1, newly proposed estimators are considered:

- The TrIPW estimator (Chu and Beaumont, 2019): A modification of PSA 3.4 that uses a modified version of the Classification And Regression Trees (CART) algorithm (Breiman et al., 1984). The propensity for each individual i $\in s_V$ is calculated as:

$$\tilde{p_i}^{CART} = \frac{\#\left(l(i) \cap s_V\right)}{\#\left(l(i)\right)} \tag{3.30}$$

  where $l(i)$ represents the terminal node of the CART algorithm trained on s_V in which the i-th individual of U lies and $\#\left(l(i)\right)$ the number of individuals it includes. Given that $U - s_V$ is not available, that propensity has to be estimated as:

$$\hat{p_i}^{CART} = \frac{\#\left(l(i) \cap s_V\right)}{\hat{\#}\left(l(i)\right)} = \frac{\#\left(l(i) \cap s_V\right)}{\sum_{j \in l(i) \cap s_R} d_j} \tag{3.31}$$

- Kernel Weighting (Wang et al., 2020): A modification of PSA 3.4. In this case, the propensity estimation process of $\hat{p_i}$ for $i \in s_V$ is not affected. The authors consider initially the use of logistic regression as the chosen machine learning model. For $j \in s_R$ we compute the distance of its estimated propensity score from each i in the non-probability sample as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \hat{p_i}(\mathbf{x}_i) - \hat{p_j}(\mathbf{x}_j) \tag{3.32}$$

  Then, a zero-centred Kernel function is applied to smooth distances. Thus, the pseudoweights can be calculated:

$$k_{ij} = \frac{K\left\{d(\mathbf{x}_i, \mathbf{x}_j)/h\right\}}{\sum_{j \in s_V} K\left\{d(\mathbf{x}_i, \mathbf{x}_j)/h\right\}} \tag{3.33}$$

where $K(\cdot)$ is the applied Kernel function (ie. Gaussian) and h is the bandwidth. The final weights for $i \in s_V$ are obtained as:

$$w_i^{KW} = \sum_{j \in s_R} k_{ij} d_j \tag{3.34}$$

The previously described Doubly Robust estimator 3.27 and the Training estimator 3.29 are also considered.

A combination of these methods with a well-tested powerful machine learning technique like XGBoost is proposed. XGBoost (Chen and Guestrin, 2016) is an improved variant of the gradient boosting algorithm (Friedman, 2001) which works as a decision tree ensemble. The final prediction is defined as follows:

$$\hat{y}_{xgi} = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \tag{3.35}$$

where K is the number of trees forming the ensemble and $\mathcal{F} = \{f(\mathbf{x}) = \omega_{q(\mathbf{x})}\}$; with $q : \mathbb{R}^m \to T$ representing the structure of each tree which, given $\mathbf{x}_i$, returns its corresponding final node and ω_i the score on the i-th final node. The trees f_k, $k = 1, ..., K$, are built aiming to minimize the following regularized objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_{xgi}, y_i) + \sum_k \Omega(f_k) \tag{3.36}$$

where the first term l is a differentiable convex function which measures the error of the estimations and the second term regularizes the function penalizing complex trees. The objective function is minimized iteratively with the Gradient Tree Boosting method (Friedman, 2001), including several modifications which improve its efficiency and efficacy.

XGBoost is applied as the chosen machine learning algorithm for the methods described above in an extensive comparative study. This study includes a replica of the experiments with simulated populations conducted in Chen et al. (2020) and another set of similar experiments considering the following real datasets:

- Hotel Booking Demand Dataset: As described in 3.1.
- Adult Dataset (Dua and Graff, 2017): It includes census income information for 32,561 adult individuals from the 1994 Census database of the United States. The target is estimating the proportion of individuals who make over $50K a year. 14 covariates are used for training. For the first nonprobability sampling strategy, individuals who make over $50K a year have double probability of being chosen. For the

second nonprobability sampling strategy, individuals who make over \$50K/yr have a propensity to participate multiplied by $Pr(a) = 2a^2$, where a is the individual's age. The probabilistic samples are obtained via simple random sampling.

Even though default XGBoost hyperparameters are used for an initial simulation, the importance of hyperparameter optimization in bias reduction methods for non-probability samples is evaluated with another set of experiments applying the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011, 2013) in order to find the optimal values.

All the described methods are applied to the ESPACOV (Serrano del Rosal et al., 2020), a survey conducted in Spain in the fourth week of the strict lockdown imposed on March 14th, 2020, which provides information on the living conditions of the population, acquired habits, health and consequences of the state of alarm and home confinement. The variance of each estimator is estimated via bootstraping (Wolter, 2007).

3.5. Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. We want to estimate a population parameter of a variable of interest, y. On the first measurement (t = 1), a sample $s^{(1)}$ of size $n^{(1)}$ is selected from the population U by random stratified sampling. Let h be the stratum to which unit i belongs, $(h = 1, ...L)$ and $s_h^{(1)}$ be the sample corresponding to stratum h on occasion 1. The lack of response in the sample $s^{(1)}$ divides it into

$$s_{rh}^{(1)} = \{i \in s^{(1)}/\text{respond in stratum h }\}$$
$$s_{fh}^{(1)} = \{i \in s^{(1)}/\text{missing in stratum h }\},$$

Let $m_h^{(1)}$ denote the number of the observations obtained from the $n_h^{(1)}$ sampled units, so $\sum_h m_h^{(1)}$ is the size of $s_r^{(1)}$.

In following measurements $t = 2, 3, ..., k$ we denote by $s_{rh}^{(t)}$ the sample of respondents in measurement t in stratum h of the original sample $s^{(1)}$, the size of which we denote by $m_h^{(t)}$. To complete the sample, a new sample $s_{new}^{(t)}$ is selected from the population U by stratified sampling independently of the sample $s^{(1)}$. Let $n_{hnew}^{(t)}$ be the size of the sample $s_{new}^{(t)}$ in stratum h and denote by $m_{hnew}^{(t)}$ the size of the sample of respondents in this stratum, $s_{rhnew}^{(t)}$.

Let $y_i^{(t)}$ be the value of the target variable associated to the i-th unit in measurement t. An initial estimation of the total of Y in the first occasion

is given by the Horvitz-Thompson estimator:

$$\hat{Y}_{ht}^{(1)} = \sum_{h} \sum_{i \in s_{rh}^{(1)}} d_{ih} y_{ih}^{(1)} \tag{3.37}$$

where $d_{ih} = \frac{N_h}{n_h^{(1)}}$.

In order to take the response rate in stratum h, $r_h = \frac{m_h^{(1)}}{n_h^{(1)}}$, into account, the initial weights are replaced with $d_{ih}^{(1)} = \frac{d_{ih}}{r_h}$ and the estimator is given by:

$$\hat{Y}^{(1)} = \sum_{h} \sum_{i \in s_{rh}^{(1)}} d_{ih}^{(1)} y_{ih}^{(1)} \tag{3.38}$$

For the followings measurements, since both $s_{new}^{(t)}$ and $s_r^{(t)}$ are available, two different estimators have to be considered:

$$\hat{Y}_n^{(t)} = \sum_{h} \sum_{s_{rhnew}^{(t)}} \frac{N_h}{n_{hnew}^{(t)}} \frac{n_{hnew}^{(t)}}{m_{hnew}^{(t)}} y_{ih}^{(t)} = \sum_{h} \sum_{s_{rhnew}^{(t)}} d_{ihn}^{(t)} y_{ih}^{(t)} \tag{3.39}$$

$$\hat{Y}_r^{(t)} = \sum_{h} \sum_{i \in s_{rh}^{(t)}} \frac{N_h}{n_h^{(1)}} \frac{n_h^{(1)}}{m_h^{(t)}} y_{ih}^{(t)} = \sum_{h} \sum_{i \in s_{rh}^{(t)}} d_{ihr}^{(t)} y_{ih}^{(t)} \tag{3.40}$$

They can be combined with the following estimator:

$$\hat{Y}^{(t)} = \alpha_1 \hat{Y}_r^{(t)} + \alpha_2 \hat{Y}_n^{(t)} \tag{3.41}$$

where α_1 and α_2 are nonnegative constants such that $\alpha_1 + \alpha_2 = 1$. The optimal coefficients would be $\alpha_1 = 1 - \alpha_2 = \frac{V(\hat{Y}_n^{(t)})}{V(\hat{Y}_r^{(t)}) + V(\hat{Y}_n^{(t)})}$ but, since those variances are unknown, a simple solution is to weight each estimator by the weight that sample has in the total sample. In this way we consider the self-weighted total estimator:

$$\hat{Y}_{sw}^{(t)} = \sum_{h} \frac{N_h}{m_h^{(t)} + m_{hnew}^{(t)}} \Big(\sum_{i \in s_{rh}^{(t)}} y_{ih}^{(t)} + \sum_{i \in s_{rhnew}^{(t)}} y_{ih}^{(t)} \Big) = \sum_{h} \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} d_{ich}^{(t)} y_{ih}^{(t)} \tag{3.42}$$

Calibration, as described in formula 3.16, should also be applied with known population totals. Thus the calibration total estimator is obtained as:

$$\hat{Y}_{\text{CAL}}^{(t)} = \sum_{h} \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} y_{ih}^{(t)} \tag{3.43}$$

where $w_{ih}^{(t)}$ are the calibrated weights from $d_{ich}^{(t)}$ for each stratum h.

Therefore, the absolute change from one measurement to the first measurement of the variable, denoted by $\theta^{(t)} = Y^{(t)} - Y^{(1)}$, is estimated as:

$$\hat{\theta}^{(t)}_{abs} = \hat{Y}^{(t)}_{\text{CAL}} - \hat{Y}^{(1)}_{\text{CAL}} \tag{3.44}$$

Similarly, the relative change $\theta^{(t)}_{rel} = \frac{Y^{(t)} - Y^{(1)}}{Y^{(1)}}$ can be estimated with the ratio estimator as:

$$\hat{\theta}^{(t)}_{rel} = \frac{\hat{\theta}^{(t)}_{abs}}{\hat{Y}^{(1)}_{\text{CAL}}} \tag{3.45}$$

Another parameter of interest is the gender gap. Let $Gen = \{M, W\}$ be the variable measured in $s^{(t)}, t = 1, 2, 3, ..., k$ which reflects whether a respondent is a man (M) or a woman (W). We define the two indicator variables: $I^M_{ih} = 1$ if the unit i in stratum h is a man and 0 elsewhere, and I^W_{ih} in a similar way. The absolute gender gap estimator in the absolute change is defined as follows:

$$\begin{gathered} \hat{GGabs}^{(t)}_{abs} = \hat{\theta}^{(t)}_W - \hat{\theta}^{(t)}_M = \\ = (\sum_h \sum_{i \in s^{(t)}_{rh} \cup s^{(t)}_{rhnew}} w^{(t)}_{ih} y^{(t)}_{ih} I^W_{ih} - \sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^W_{ih}) - \\ (\sum_h \sum_{i \in s^{(t)}_{rh} \cup s^{(t)}_{rhnew}} w^{(t)}_{ih} y^{(t)}_{ih} I^M_{ih} - \sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^M_{ih}) \end{gathered} \tag{3.46}$$

Alternatively, we define the relative gender gap estimator in the absolute change as follows:

$$\hat{GGabs}^{(t)}_{rel} = \frac{\hat{GGabs}^{(t)}_{abs}}{\hat{\theta}^{(t)}_M} = \frac{\hat{\theta}^{(t)}_W - \hat{\theta}^{(t)}_M}{\hat{\theta}^{(t)}_M} \tag{3.47}$$

We also define the absolute gender gap in the relative change as follows:

$$\hat{GGrel}^{(t)}_{abs} = \hat{\theta}^{(t)}_{relW} - \hat{\theta}^{(t)}_{relM} = \frac{\hat{\theta}^{(t)}_W}{\sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^W_{ih}} - \frac{\hat{\theta}^{(t)}_M}{\sum_h \sum_{i \in s^{(1)}_{rh}} w^{(1)}_{ih} y^{(1)}_{ih} I^M_{ih}} \tag{3.48}$$

and the relative gender gap in the relative change as follows:

$$\hat{GGrel}^{(t)}_{rel} = \frac{\hat{GGrel}^{(t)}_{abs}}{\hat{\theta}^{(t)}_{relM}} \tag{3.49}$$

In order to produce longitudinal estimations, the weights after the first measurement should take attrition into account. Therefore, the weights are adjusted for non-response using the Propensity Score Adjustment method to model the probability that a unit of the sample $s^{(1)}_r$ responds on occasion t. For each sample unit $s^{(1)}_r$ let be $\delta^{(t)}_i = 1$ if $i \in s^{(t)}_r$ and $\delta^{(t)}_i = 0$ if $i \in s^{(1)}_r - s^{(t)}_r$. We assume that the selection mechanism of response is ignorable

and follows a parametric model. In order to obtain the estimated propensities $\hat{\pi}_i^{(t)}$, we train a model with $s_r^{(1)}$ where $\mathbf{x}_i$ includes every available variable observed in the sample. Said model minimizes the logistic loss for $\delta_i^{(t)}; i \in s_r^{(1)}$. The chosen machine learning model for this purpose is XGBoost, including hyperparameter optimization as described in Section 3.4.

The estimated propensities for each unit i of sample $s_{rh}^{(t)}$, $\hat{\pi}_{ih}^{(t)}$, are used to reweighting for nonresponse, and we define an estimator for $\theta^{(t)}$ from the sample of respondents on occasion t by:

$$\hat{\theta}_l^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ih}^{(1)} \frac{1}{\hat{\pi}_{ih}^{(t)}} (y_{ih}^{(t)} - y_{ih}^{(1)}) = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ihPSA}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \quad (3.50)$$

This weights $d_{ihPSA}^{(t)}$ are also calibrated as described in formula 3.16 with known population totals, obtaining the final weights $v_{ih}^{(t)}$. The final calibrate estimator is given by

$$\hat{\theta}_c^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \quad (3.51)$$

This process is repeated with $\theta^{(t,t-1)} = Y^{(t)} - Y^{(t-1)}$ in order to model the non-response with respect to the sample obtained in the previous occasion. Thus the longitudinal estimator of $\theta^{(t,t-1)} = Y^{(t)} - Y^{(t-1)}$ can be defined as follows:

$$\hat{\theta}_c^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) \quad (3.52)$$

We may be also interested in the number of population individuals whose value of y increases, decreases or remains the same between $t-1$ and t. Let A be a subset of interest ($\mathbb{R}^+$, $\mathbb{R}^-$ or 0); the estimator of the number of population individuals for which $y^{(t)} - y^{(t-1)} \in A$ can be estimated as follows:

$$\hat{\theta}_{cA}^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A,\ I_A = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \in A \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \notin A \end{cases} \quad (3.53)$$

We can also obtain the estimator of the rate of people whose value in y has decreased between $t-1$ and t, in reference to the people whose value in y has increased between $t-1$ and t:

$$\widehat{DIRate}_c^{(t,t-1)} = \frac{\hat{\theta}_{cA_{R^-}}^{(t,t-1)} - \hat{\theta}_{cA_{R^+}}^{(t,t-1)}}{\hat{\theta}_{cA_{R^+}}^{(t,t-1)}} \quad (3.54)$$

Finally, estimators of the gender gap of the change between $t-1$ and t can be defined as follows:

$$\hat{GGlong}_{abs}^{(t)} = \hat{\theta}_{cW}^{(t,t-1)} - \hat{\theta}_{cM}^{(t,t-1)} \quad (3.55)$$

$$\hat{GGlong}^{(t)}_{absA} = \hat{\theta}^{(t,t-1)}_{cAW} - \hat{\theta}^{(t,t-1)}_{cAM} \tag{3.56}$$

$$\hat{GGlong}^{(t)}_{rel} = \frac{\hat{GGlong}^{(t)}_{abs}}{\hat{\theta}^{(t,t-1)}_{cM}} \tag{3.57}$$

$$\hat{GGlong}^{(t)}_{relA} = \frac{\hat{GGlong}^{(t)}_{absA}}{\hat{\theta}^{(t,t-1)}_{cAM}} \tag{3.58}$$

3.6. Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain

In the same context as Section 3.1, now the values of the variable of interest y are also available in the probabilistic sample s_R. Therefore, the population total Y can be estimated via the Horvitz-Thompson estimator:

$$\hat{Y}_R = \sum_{i \in s_R} \frac{y_i}{\pi_i} = \sum_{i \in s_R} d_i y_i \tag{3.59}$$

which is design-unbiased of the population total if there is not lack of response. The design-based variance of this estimator is given by:

$$V_p(\hat{Y}_R) = \sum_{i,j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \left(\pi_{ij} - \pi_i \pi_j\right). \tag{3.60}$$

where π_{ij} are the second order probabilities of the sampling design p_R. If $\pi_{ij} > 0 \ \forall (i,j)$, an unbiased estimator is given by:

$$\widehat{V}_p(\hat{Y}_R) = \sum_{i,j \in s_R} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{3.61}$$

In order to incorporate the non-probability sample data, the Inverse Probability Weighted estimator $\hat{Y}_{IPW}$ is also calculated following the methodology described in formula 3.4. Chen et al. (2020) determines an expression for its variance as

$$V(\hat{Y}_{IPW}) = \sum_{i=1}^{N} (y_i/\hat{p}_i - \mathbf{b}_1^T \mathbf{x}_i)^2 (1-\hat{p}_i)\hat{p}_i + \mathbf{b}_1^T D \mathbf{b}_1 \tag{3.62}$$

where $\mathbf{b}_1^T = \sum_{i=1}^{N} (1-\hat{p}_i) y_i \mathbf{x}_i^T / \sum_{i=1}^{N} \hat{p}_i (1-\hat{p}_i) \mathbf{x}_i \mathbf{x}_i^T$, and $D = V_p(\sum_{i \in s_R} d_i \hat{p}_i \mathbf{x}_i)$ where V_p denotes the design-based variance under the sampling design p. However, from a practical viewpoint, it is better to use jackknife or bootstrap techniques Wolter (2007) for its estimation.

A simple estimator is calculated by weighting the estimators obtained from each sample:

$$\hat{Y}_{com} = \alpha\hat{Y}_R + (1-\alpha)\hat{Y}_{IPW} \tag{3.63}$$

where α is a nonnegative constant such that $0 \leq \alpha \leq 1$. Different values can be considered, such as: $\alpha_0 = \frac{\hat{V}(\hat{Y}_{IPW})}{\hat{V}_p(\hat{Y}_R)+\hat{V}(\hat{Y}_{IPW})}$, $\alpha_n = \frac{n_R}{n_R+n_V}$ or $\alpha_e = 0{,}5$. The resulting estimator can be rewritten as:

$$\hat{Y}_{com} = \sum_{i\in s} y_i d_i^\star \tag{3.64}$$

being $s = s_R \bigcup s_V$ and

$$d_i^\star = \begin{cases} \alpha d_i & \text{if } i \in s_R \\ (1-\alpha)w_i & \text{if } i \in s_V \end{cases} \tag{3.65}$$

Calibration can also be applied to these $d_i^\star$ weights with some known population totals, as described in formula 3.16, to obtain new calibrated weights $w_i^\star$. Thus the calibration estimator can be calculated as:

$$\hat{Y}_{\text{CAL}} = \sum_{i\in s} w_i^\star y_i. \tag{3.66}$$

The following estimators are evaluated in a comparative study:

- Reference estimator ($\hat{Y}_{REF}$): the two samples are joined and calibration is performed to obtain the final estimator.
- Elliott and Haviland estimator ($\hat{Y}_{EH}$): we join the probabilistic and non-probabilistic sample and obtain the final estimator using the formulas proposed in Elliott and Haviland (2007).
- Based on Robbins et al. (2021) we calculate four estimators:
  - the disjoint propensity score (DPS) weights estimator (section 2.1.1. of Robbins et al. (2021)): $\hat{Y}_{RDR1}$
  - the simultaneous weights estimator (section 2.1.2. of Robbins et al. (2021)): $\hat{Y}_{RDR2}$
  - the disjoint calibration (DC) weights estimators (section 2.2 of Robbins et al. (2021)): $\hat{Y}_{RDR3}$
  - the combined calibration estimator (section 2.2 of Robbins et al. (2021)): $\hat{Y}_{RDR4}$
- Propensities estimator ($\hat{Y}_{PPSA}$): the probability and non-probability sample propensities are obtained, both samples are merged and calibration is performed to obtain the final estimator using the inverse of propensities as initial weights.

- Calibration - PSA estimator ($\hat{Y}_{CPSA}$): calibration is performed in the probability sample and in the non-probability sample we calculate the propensities. To obtain the final estimator, we combine them in several ways, considering $\alpha_{0,5}$, α_n and α_0, and then we do the calibration. We will denote these estimators $\hat{Y}_{CPSA-0,5}$, $\hat{Y}_{CPSA-n}$ and $\hat{Y}_{CPSA-\alpha_0}$.

In all the estimators in which the propensities are calculated we use as the chosen machine learning model both logistic regression and XGBoost, including hyperparameter optimization for the later as described in Section 3.4.

We simulate a population of size 500,000 in which we have two target variables y_1 and y_2, and eight auxiliary variables to perform the PSA algorithms and the calibration, $x_1, \ldots, x_8$, defined as follows:

$$x_{1i}, x_{3i}, x_{5i}, x_{7i} \sim B(0{,}5), \quad i \in U \tag{3.67}$$

$$x_{ji} \sim N(\mu_{ji}, 1), \quad i \in U, j = 2, 4, 6, 8 \tag{3.68}$$

with

$$\mu_{ji} = \begin{cases} 2, & \text{if } x_{(j-1)i} = 1 \\ 0, & \text{if } x_{(j-1)i} = 0 \end{cases}, \quad i \in U, j = 2, 4, 6, 8$$

The target variables were simulated as follows:

$$y_{1i} = N(10, 4) + 5\pi_i, \quad i \in U \tag{3.69}$$

$$y_{2i} = N(10, 4) + 2(x_{7i} = 1) - 2(x_{7i} = 0) + x_{8i} + 5\pi_i, \quad i \in U \tag{3.70}$$

500 iterations are carried out in order to obtain the Relative Bias and Root Mean Square Relative Error for each method. In each iteration, we draw a probability sample of size $n_P = 250$ and a non-probability sample of sizes $n_{NP} = 500; 1,000; 2,000$. The probability sample is drawn by simple random sampling without replacement (SRSWOR) from the full population. The non-probability sample is drawn according to the following probability sampling design:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = -0{,}5 + 2{,}5(x_{5i} = 1) + \sqrt{2\pi}x_{6i}x_{8i} - 2{,}5(x_{7i} = 1), \quad i \in U. \tag{3.71}$$

Chapter 4

Results

4.1. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

The results of the simulation study are summarized in the following points:

- In general, it can be observed that Statistical Matching outperforms PSA, which outperforms PSA with stratification.
- In the case of standard deviation, there is also evidence that PSA without propensity stratification provides higher values than Matching.
- Nevertheless, all three methods consistently reduce the sample bias.
- In terms of machine learning algorithms, basic linear models seem the most robust. Others, like Naive Bayes or k-Nearest Neighbors, can achieve outstanding results but only for some cases.
- Basic decision trees can lead to worse estimations than the naive estimator for some cases. Bagged trees are also outperformed by the other models.
- In order to avoid overfitting, presumably one of the main problems of matching, Ridge regression should be prefered if the data suffer from multicollinearity. Otherwise, linear regression alone is fast and effective.

4.2. The R Package NonProbEst for Estimation in Non-probability Surveys

The resulting package includes the following functions:

- *propensities*: It computes the estimated propensities given a convenience sample and a reference sample.
- *matching*: It predicts unknown responses by applying Statistical Matching.
- *model_based*: It calculates a model based mean or total estimation.
- *model_assisted*: It calculates a model assisted mean or total estimation.
- *model_calibrated*: It calculates a model calibrated mean or total estimation.
- *valliant_weights*: It computes the sample weights given the estimated propensities using the formula introduced by Valliant (2020).
- *sc_weights*: It computes the sample weights given the estimated propensities using the formula introduced by Schonlau and Couper (2017).
- *vd_weights*: It computes the sample weights given the estimated propensities using the formula introduced by Valliant and Dever (2011).
- *lee_weights*: It computes the sample weights given the estimated propensities using the formula introduced by Lee (2006) and Lee and Valliant (2009).
- *calib_weights*: It applies calibration to some sample weights.
- *mean_estimation*: It estimates the population mean given some sample weights.
- *total_estimation*: It estimates the population total given some sample weights.
- *prop_estimation*: It estimates the proportion of a category in the population given some sample weights.
- *jackknife_variance*: It calculates the Jackknife variance with reweighting for PSA.
- *generic_jackknife_variance*: It calculates the Jackknife variance with reweighting for an arbitrary estimator.
- *fast_jackknife_variance*: It calculates the Jackknife variance without reweighting.
- *confidence_interval*: It calculates the confidence interval for a given estimator.

A wide variety of machine learning models can be specified for the following functions: *propensities*, *matching*, *model_based*, *model_assisted* and *model_calibrated*.

4.3. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

The results of the simulation study are summarized in the following points:

- The proposed Training estimator always obtains the best estimations.
- Even in those cases for which Propensity Score Adjustment outperforms Statistical Matching, Training is still a better choice.
- The Doubly Robust estimator offers very similar results, although slightly worse than the Training estimator.
- Shrinkage is not as optimal because its efficacy is simply an average between Statistical Matching and Propensity Score Adjustment.

4.4. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

The results of the comparative study with simulated populations are summarized in the following points:

- Even though linear/logistic regression is theoretically unbeatable for the linear models, it is observed that XGBoost can also effectively remove the bias in those cases.
- XGBoost is still able to correctly learn the non-linear models, where linear/logistic regression suffers.
- The TrIPW and the XGBoosted Kernel Weighting estimators suffer from overfitting without hyperparameter optimization.
- Doubly Robust estimators are not required for such simple models.

The results of the comparative study with real populations are summarized in the following points:

- A significant improvement in the estimations can be observed when using XGBoost instead of linear or single tree regressors.
- XGBoost is more useful as more data is available.

- In the majority of cases, the Matching based variants obtain the best results. However, for some specific cases, XGBoosted Kernel Weighting is better.

- Doubly Robust estimators may yield slightly more accurate estimations in these cases with XGBoost as well, specially when using the Training estimator.

- Applying hyperparameter optimization considerably improves the estimations. In some cases, this improvement is so significant that the method which was the worst one without optimization is now the best alternative.

The results of the application to the ESPACOV are summarized in the following points:

- The results generally show that the application of bias correction techniques provides an important shift with respect to the unweighted estimate.

- There is a small and expectable increase in variance from the unweighted case.

- The differences are not as significant when the target variables are not related to the covariates used.

4.5. Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

We have developed novel estimators for several parameters in overlapping panel surveys. The applied weighting methods reduce the bias in the estimation of the total, the absolute and relative difference of the total, transversal gender gaps in absolute and relative terms, longitudinal differences, deterioration/improvement rates and longitudinal gender gaps in relative and absolute terms.

The results from applying the proposed methods to the ESSOC are summarized in the following points:

- The self-perceived general health worsens for the population older than 35 years as the pandemic advances.

- This deteriorating of general health can be observed more as age increases.

- Women's self-perceived health is more affected by the pandemic than men's in the long term.
- However, men's self-perceived health is more affected by the initial lockdown.
- For the population between 16 and 34 years old, the evolution has been stable throughout the pandemic since the lockdown, for men as well as for women.

4.6. Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain

The results of the simulation study are summarized in the following points:

- Calibration in both samples is not enough to completely remove selection bias, although this approach provides smaller RB and RMSRE than other methods.
- The method proposed by Elliott and Haviland (2007) is vastly efficient at removing part of the selection bias.
- The combination of calibration and PSA reduces bias and RMSRE, particularly when the algorithm used in PSA is XGBoost, although the advantage of this algorithm vanishes as the non-probability sample size increases.
- The behavior of the estimators considered in Robbins et al. (2021) is very diverse. In some cases, the Relative Bias is even larger than the case where only calibration is used. Others are able to considerably reduce RB and RMSRE as long as XGBoost is used.
- The proposed estimator $\hat{Y}_{CPSA}$, particularly when the coefficients applied consider the estimated variances ($\hat{Y}_{CPSA-\alpha_0}$), is the best alternative.

This new estimation method, which has shown to be very efficient at reducing biases, is applied to several variables in the ESPACOV II in order to assess the impact of the COVID-19 in Spain. The results solve the discrepancies between the conclusions obtained from the probability sample and those obtained from the non-probability sample.

Chapter 5

Conclusions

5.1. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

Two of the most popular proposed methods to reduce biases produced by nonprobability sampling, PSA and Matching, are studied through a complex simulation study considering several datasets and different self-selection scenarios. Results show that Statistical Matching provides, in general, better results than PSA on bias reduction and RMSE, regardless of the dataset and selection mechanism.

In addition, linear models and k-Nearest Neighbors provided, on average, better predictions in terms of bias reduction than more complex models, such as GBM and Bagged Trees. Complex models are often dependent on many factors like hyperparameter optimization or data preprocessing. This causes high variability for the results obtained. Therefore, basic models are more robust and would be preferred in order to obtain reliable results.

These conclusions are obtained after working with real-life examples. Thus, we cannot ensure whether these selection bias mechanisms are Missing At Random (MAR) or Missing Not At Random (MNAR), as the causality relationships are not known. It is known that the selection mechanism makes a difference in terms of how challenging bias reduction can be.

5.2. The R Package NonProbEst for Estimation in Non-probability Surveys

The NonProbEst package for R can simplify the use of different methods to correct selection bias in non-probability surveys. It supports the user for estimation in a wide variety of contexts, depending on the available auxi-

liary information. This allows an easy application of many alternatives (calibration, PSA, PSA+calibration, Statistical Matching, model-calibration...) within a single package. Another important feature is that any kind of machine learning algorithm can be applied, without requiring expert knowledge by the user, in order to optimize the information provided by the auxiliary variables.

The package is very flexible, allowing different strategies depending on the situation and the available execution time. For example, a hyperparameter optimization process is carried out by default but it can be easily deactivated. Also, the Jackknife variance estimator can be computed with or without reweighting. However, some newer techniques like Kernel Weighting for PSA or Doubly Robust estimators have been proposed since its development so they are not included yet.

5.3. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

We propose a novel method for combining two of the most effective techniques in order to reduce the bias of non-probability surveys: Propensity Score Adjustment and Statistical Matching. The efficiency of the proposed method is tested in an extensive comparative study with simulations from three different datasets, thus considering its behavior under different conditions. Results show a certain advantage of the developed method over other proposals like the Doubly Robust estimator or shrinkage. Its estimations are always better than those given by PSA or SM alone.

The advantage of our training method is that it gives more importance in the prediction to those individuals who are more likely to appear in the population. By default, a model trained in a biased dataset might also produce biased predictions; however, if this bias is corrected by methods such as PSA, it is expected that the relationships established by the prediction model and its results are more similar to those present in the target population.

5.4. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

We present four different methods to estimate parameters based on the use of an important machine learning technique, the XGBoost algorithm, to predict the values of the target variable in the probability sample and also to determine the propensities of participating in the non-probability sample.

These are evaluated along classical estimators as well as other methods of estimation from web survey data that are more innovative.

To be as close as possible to other recent estimation works in non-probability surveys, we have replicated the experiment carried out by Chen et al. (2020). When comparing results from both simulations, we observe that estimators involving XGBoost provide better results overall in non-linear situations in comparison to the case where linear models are used. These results are relevant considering that, in practice, models will rarely be linear. In fact, they will likely be much more complex than the ones considered in this simulation. For this reason, we also compare the different estimators in two real datasets. The performance of XGBoost is shown to be better than classical machine learning models in terms of bias and Mean Square Error reduction.

However, these results can also be unreliable when the algorithms suffer from overfitting. Hyperparameter optimization has shown to be highly effective at controlling this issue. These kinds of procedures are therefore important when producing estimations.

The proposed method is also used to analyze a nonprobability survey sample on social effects of COVID-19. The results of this application show that selection bias correction techniques have the potential to provide substantial changes in the estimates of population means in nonprobability samples.

5.5. Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

We have established a two-step reweighting process in order to adjust the bias associated to surveys with an overlapping probability panel design. First, via calibration, the population nonresponse is addressed, understood as people who did not take part in the survey despite having been selected in the sample. Then, via PSA, the panel nonresponse is addressed, understood as people who participated in some of the measurements but did not follow up in further ones. This Propensity Score Adjustment is applied using state of the art machine learning techniques, such as XGBoost and hyperparameter optimization.

The proposed estimators are applied to the ESSOC in order to obtain reliable estimates, both cross-sectional and longitudinal, on the impact of COVID-19 on health and its determinants. The results show that the impact of the pandemic has hit differently across age groups and genders. More precisely, the self-perceived general health seems to have decreased more notably in older age groups and women, both according to the evolution of

cross-sectional estimates and longitudinal estimates. The gender gap, both in absolute and relative terms, has mostly grown as the pandemic advanced, meaning that the changes (mostly decreases in self-perceived general health) have been larger and worse in women in comparison to men.

5.6. Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain

We have addressed the problem of improving the estimates obtained from the union of a probability sample and a convenience sample. We introduce four methods for calculating weights that blend them together. These methods combine calibration and Propensity Score Adjustment using machine learning techniques for those situations where the variables of interest are observed in both samples.

We evaluate the behavior of the proposed estimators against other techniques for integrating probability and non-probability samples used in the literature. For this purpose, we have considered a simulation study with several sample sizes to cover different Missing At Random situations and we have compared the performance of standard logistic regression model with a modern machine learning algorithm (XGBoost) when estimating the propensity score. Our simulations show that the proposed estimator based on calibration and PSA techniques is very efficient at reducing self-selection bias and RMSRE with this kind of data. Also, the best performing techniques for the estimation of the propensity scores were those based on boosting, which guaranteed considerably lower bias and RMSRE in comparison to a similar estimator based on logistic regression and other techniques considered in the study. Hyperparameter tuning is also applied and the results prove its importance when using machine learning techniques in this context.

The application of $\hat{Y}_{CPSA-\alpha_0}$, the proposed method which has shown the most promising results at the simulation, to the ESPACOV II, a survey about the effects of the COVID-19 pandemic in Spain, provides the best correction of the impact of the deviations from population parameters in both samples.

Chapter 6

Future Research

The research presented in this dissertation lays the foundation for further issues which will be addressed in future works. One of the most interesting topics which should be explored is the evaluation in advance of the efficacy of a specific method when applied to a specific survey. So far, we have carried out comparative studies with simulations for which the real value for the parameter of interest is already known. This allowed a precise measurement of the resulting bias and Mean Square Error. However, the results have been very variable and dependent on each context. Therefore, being able to determine the expectable bias reduction even when said value is not available, which will be the case for a real case scenario, would be very useful. This would also allow us to compare techniques in order to determine the best alternative for each case. An in-depth analysis of the relation between known error estimation techniques, such as cross-validation, and the bias reduction after applying the methods would help for this matter.

The possibility of compensating the bias associated to a non-probability sample with a new auxiliary sampling would also be very interesting. Sometimes, the available data is too limited and therefore reducing its bias is too difficult. The model that estimates propensities could be used in order to define a new sampling design which specifically compensates its limitations. For example, these kinds of techniques may be especially useful when working with overlapping panel designs. The expectable non-response could be considered and compensated in advance for the following measurements.

Another frequently requested characteristic at the application of these methods is explicability. Obtaining accurate estimations with "black box"models is sometimes not enough. Being able to explain those estimations is also often useful. This would allow us to describe and interpret the origin of the bias. Thus it would be easier for a researcher to identify the problems and solve them in future works. In any case, being able to explain the self-selection or non-response bias would be valuable for many scenarios. Therefore, the use of interpretable models when applying the methods should be explored.

Finally, other statistical methodologies will be considered in order to further improve the robustness of the bias reduction techniques, similarly to the recent proposals by Rafei et al. (2022b) which uses Bayesian bootstrap methods in a fully model-based method. Implementations of all the novel state-of-the-art methods should be included in the NonProbEst package.

Bibliography

Anduiza, E. and Galais, C. Answering without reading: Imcs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, vol. 29(3), pp. 497–519, 2017.

Arcos, A., Contreras, J. M. and Rueda, M. M. A novel calibration estimator in social surveys. *Sociological Methods & Research*, vol. 43(3), pp. 465–489, 2014.

Beaumont, J.-F. Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, vol. 46(1), pp. 1–29, 2020.

Beaumont, J.-F. and Bissonnette, J. Variance estimation under composite imputation: The methodology behind sevani. *Survey Methodology*, vol. 37(2), pp. 171–179, 2011.

Bergstra, J., Bardenet, R., Bengio, Y. and KÃ©gl, B. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011.

Bergstra, J., Yamins, D. and Cox, D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 115–123. PMLR, 2013.

Berzofsky, M. E., McKay, T., Hsieh, Y. P. and Smith, A. Probability-based samples on twitter: Methodology and application. *Survey Practice*, vol. 11(2), p. 4936, 2018.

Bethlehem, J. Selection bias in web surveys. *International Statistical Review*, vol. 78(2), pp. 161–188, 2010.

Breidt, F. J., Opsomer, J. D. et al. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, vol. 32(2), pp. 190–205, 2017.

Breiman, L. Bagging predictors. *Machine learning*, vol. 24(2), pp. 123–140, 1996.

BREIMAN, L. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California, 1997.

BREIMAN, L. Random forests. *Machine learning*, vol. 45(1), pp. 5–32, 2001.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. *Classification and regression trees*. CRC press, 1984.

BRICKMAN BHUTTA, C. Not by the book: Facebook as a sampling frame. *Sociological methods & research*, vol. 41(1), pp. 57–88, 2012.

BURDEN, F. and WINKLER, D. Bayesian regularization of neural networks. *Artificial neural networks*, pp. 23–42, 2008.

CALLEGARO, M., MANFREDA, K. L. and VEHOVAR, V. *Web survey methodology*. SAGE, 2015.

CASTRO-MARTÍN, L., RUEDA, M. D. M. and FERRI-GARCÍA, R. Estimating general parameters from non-probability surveys using propensity score adjustment. *Mathematics*, vol. 8(11), p. 2096, 2020a.

CASTRO-MARTÍN, L., RUEDA, M. D. M. and FERRI-GARCÍA, R. Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics*, vol. 8(6), p. 879, 2020b.

CASTRO-MARTÍN, L., RUEDA, M. D. M. and FERRI-GARCÍA, R. Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *Journal of Computational and Applied Mathematics*, vol. 404, p. 113414, 2022.

CHEN, T. and GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, San Francisco California USA, 2016. ISBN 9781450342322.

CHEN, Y., LI, P. and WU, C. Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, vol. 115(532), pp. 2011–2021, 2020.

CHU, K. and BEAUMONT, J. F. The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In *Proceedings of the Survey Methods Section: SSC Annual Meeting*. 2019.

COPAS, J. The shrinkage of point scoring methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 42(2), pp. 315–331, 1993.

COPAS, J. B. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 45(3), pp. 311–335, 1983.

COUPER, M. P., GREMEL, G., AXINN, W., GUYER, H., WAGNER, J. and WEST, B. T. New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research*, vol. 73, pp. 221–235, 2018.

CSÁJI, B. C. ET AL. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24(48), p. 7, 2001.

DEVILLE, J.-C. and SÄRNDAL, C.-E. Calibration estimators in survey sampling. *Journal of the American statistical Association*, vol. 87(418), pp. 376–382, 1992.

DÍAZ DE RADA, V. Ventajas e inconvenientes de la encuesta por internet. *Papers*, vol. 97(1), pp. 193–223, 2012.

DÍAZ DE RADA, V., DOMÍNGUEZ, J. A. and PASADAS-DEL AMO, S. *Internet como modo de administración de encuestas*, vol. 59. CIS, 2019.

DISOGRA, C., COBB, C., CHAN, E. and DENNIS, J. M. Calibrating nonprobability internet samples with probability samples using early adopter characteristics. In *Joint Statistical Meetings (JSM), Survey Research Methods*, pp. 4501–4515. 2011.

DUA, D. and GRAFF, C. UCI machine learning repository. 2017.

ELLIOTT, M. N. and HAVILAND, A. Use of a web-based convenience sample to supplement a probability sample. *Survey methodology*, vol. 33(2), pp. 211–5, 2007.

FERRI-GARCÍA, R., BEAUMONT, J.-F., BOSA, K., CHARLEBOIS, J. and CHU, K. Weight smoothing for nonprobability surveys. *TEST*, pp. 1–25, 2021a.

FERRI-GARCÍA, R., CASTRO-MARTÍN, L. and RUEDA, M. D. M. Evaluating machine learning methods for estimation in online surveys with superpopulation modeling. *Mathematics and Computers in Simulation*, vol. 186, pp. 19–28, 2021b.

FERRI-GARCÍA, R. and RUEDA, M. D. M. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Statistics and Operations Research Transactions*, pp. 159–162, 2018. ISSN 2013-8830.

FERRI-GARCÍA, R. and RUEDA, M. D. M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLOS ONE*, vol. 15(4), p. e0231500, 2020. ISSN 1932-6203.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

GAO, Z., HOUSE, L. and BI, X. Impact of satisficing behavior in online surveys on consumer preference and welfare estimates. *Food Policy*, vol. 64, pp. 26–36, 2016.

GILLIGAN, C., KYPRI, K. and BOURKE, J. Social networking versus facebook advertising to recruit survey respondents: a quasi-experimental study. *JMIR research protocols*, vol. 3(3), p. e48, 2014.

GROVES, R. M. and HEERINGA, S. G. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 169(3), pp. 439–457, 2006.

GUTIÉRREZ, H. A. *Estrategias de muestreo diseño de encuestas y estimacion de parametros.*. Universidad Santo Tomas, Bogota (Colombia)., 2009.

HORVITZ, D. G. and THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, vol. 47(260), pp. 663–685, 1952.

IANNELLI, L., GIGLIETTO, F., ROSSI, L. and ZUROVAC, E. Facebook digital traces for survey research: Assessing the efficiency and effectiveness of a facebook ad–based procedure for recruiting online survey respondents in niche and difficult-to-reach populations. *Social Science Computer Review*, vol. 38(4), pp. 462–476, 2020.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. *An introduction to statistical learning*, vol. 112. Springer, 2013.

KALTON, G. and BRICK, J. M. Weighting schemes for household panel surveys. *Survey Methodology*, vol. 21(1), pp. 33–44, 1995.

KALTON, G. and CITRO, C. F. Panel surveys: Adding the fourth dimension. *Innovation: The European Journal of Social Science Research*, vol. 8(1), pp. 25–39, 1995. ISSN 1351-1610, 1469-8412.

KALTON, G., LEPKOWSKI, J. and LIN, T. K. Compensating for wave nonresponse in the 1979 ISDP research panel. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, vol. 372, p. 377. 1985.

KERN, C., LI, Y. and WANG, L. Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 2020.

KOHUT, A., KEETER, S., DOHERTY, C., DIMOCK, M. and CHRISTIAN, L. Assessing the representativeness of public opinion surveys. *Washington, DC: Pew Research Center*, 2012.

KUHN, M. Caret: classification and regression training. *Astrophysics Source Code Library*, pp. ascl–1505, 2015.

LEE, S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, vol. 22(2), p. 329, 2006.

LEE, S. and VALLIANT, R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, vol. 37(3), pp. 319–343, 2009.

LEPKOWSKI, J. M. Treatment of wave nonresponse in panel surveys. *Panel surveys*, 1989.

LIU, Y., GELMAN, A. and CHEN, Q. Inference from non-random samples using bayesian machine learning. *arXiv preprint arXiv:2104.05192*, 2021.

MANFREDA, K. L., BOSNJAK, M., BERZELAK, J., HAAS, I. and VEHOVAR, V. Web surveys versus other survey modes: A meta-analysis comparing response rates. *International journal of market research*, vol. 50(1), pp. 79–104, 2008.

MARKEN, S. Still listening: The state of telephone surveys. 2018. `https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx`. Accessed 2 March 2021.

MCDONALD, G. C. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1(1), pp. 93–100, 2009.

MORO, S., CORTEZ, P. and RITA, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

NEYMAN, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, vol. 97(4), pp. 558–625, 1934.

PÖTZSCHKE, S. and BRAUN, M. Migrant sampling using facebook advertisements: A case study of polish migrants in four european countries. *Social Science Computer Review*, vol. 35(5), pp. 633–653, 2017.

QUENOUILLE, M. H. Notes on bias in estimation. *Biometrika*, vol. 43(3/4), pp. 353–360, 1956.

RAFEI, A., ELLIOTT, M. R. and FLANNAGAN, C. A. Robust and efficient bayesian inference for non-probability samples. *arXiv preprint arXiv:2203.14355*, 2022a.

RAFEI, A., ELLIOTT, M. R. and FLANNAGAN, C. A. Robust model-based inference for non-probability samples. *arXiv preprint arXiv:2204.03215*, 2022b.

RINKEN, S., DOMÍNGUEZ-ÁLVAREZ, J.-A., TRUJILLO, M., LAFUENTE, R., SOTOMAYOR, R. and SERRANO-DEL ROSAL, R. Combined mobile-phone and social-media sampling for web survey on social effects of covid-19 in spain. In *Survey Research Methods*, vol. 14, pp. 165–170. 2020.

RIPLEY, B. D. *Pattern recognition and neural networks*. Cambridge university press, 2007.

RIVERS, D. Sampling for web surveys. In *Joint Statistical Meetings*, p. 4. 2007.

ROBBINS, M. W., GHOSH-DASTIDAR, B. and RAMCHAND, R. Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology*, vol. 9(5), pp. 1114–1145, 2021.

SERRANO DEL ROSAL, R., BIEDMA VELÁZQUEZ, L., DOMÍNGUEZ ÁLVAREZ, J. A., GARCÍA RODRÍGUEZ, M. I., LAFUENTE, R., SOTOMAYOR, R., TRUJILLO CARMONA, M. and RINKEN, S. Estudio social sobre la pandemia del covid-19 (espacov). 2020.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, vol. 323(6088), pp. 533–536, 1986.

SÁNCHEZ-CANTALEJO, C., RUEDA, M. D. M., SAEZ, M., ENRIQUE, I., FERRI, R., FUENTE, M. D. L., VILLEGAS, R., CASTRO, L., BARCELÓ, M. A., DAPONTE-CODINA, A., LORUSSO, N. and CABRERA-LEÓN, A. Impact of COVID-19 on the Health of the General and More Vulnerable Population and Its Determinants: Health Care and Social Survey-ESSOC, Study Protocol. *International Journal of Environmental Research and Public Health*, vol. 18(15), p. 8120, 2021.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

SCHONLAU, M. and COUPER, M. P. Options for conducting web surveys. *Statistical Science*, vol. 32(2), pp. 279–292, 2017.

SCHONLAU, M., VAN SOEST, A., KAPTEYN, A. and COUPER, M. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, vol. 37(3), pp. 291–318, 2009.

SNIJDERS, T. A. Estimation on the basis of snowball samples: how to weight? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, vol. 36(1), pp. 59–70, 1992.

SPANISH NATIONAL INSTITUTE OF STATISTICS. Life conditions survey. microdata. 2012.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58(1), pp. 267–288, 1996.

TOUZANI, S., GRANDERSON, J. and FERNANDES, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.

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

VALLIANT, R. and DEVER, J. A. Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, vol. 40(1), pp. 105–137, 2011.

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley, 2000.

VAN HOUWELINGEN, J. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, vol. 55(1), pp. 17–34, 2001.

WANG, L., GRAUBARD, B. I., KATKI, H. A. and LI, Y. Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 183(3), pp. 1293–1311, 2020.

WIŚNIOWSKI, A., SAKSHAUG, J. W., PEREZ RUIZ, D. A. and BLOM, A. G. Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, vol. 8(1), pp. 120–147, 2020.

WOLTER, K. *An Introduction to Variance Estimation*. 2007. ISBN 978-0-387-32917-8.

WU, C. and SITTER, R. R. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, vol. 96(453), pp. 185–193, 2001.

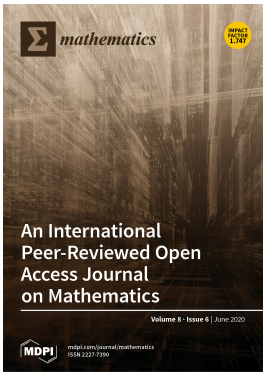
Part II

Appendices

Appendix A1

Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

Castro-Martín, Luis; Rueda, María del Mar; Ferri-García, Ramón (2020)
Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques
Mathematics, vol. 8, number 6, p. 879
DOI: 10.3390/math8060879



MATHEMATICS			
JCR Year	Impact factor	Rank	Quartile
2020	2.258	24/330	Q1

mathematics

MDPI

Article

Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques

Luis Castro-Martín, Maria del Mar Rueda * and Ramón Ferri-García

Department of Statistics and Operational Research, Faculty of Sciencies, University of Granada, 18071 Granada, Spain; luiscastro193@ugr.es (L.C.-M.); rferri@ugr.es (R.F.-G.)

* Correspondence: mrueda@ugr.es

Received: 7 May 2020; Accepted: 27 May 2020; Published: 1 June 2020

Abstract: Online surveys are increasingly common in social and health studies, as they provide fast and inexpensive results in comparison to traditional ones. However, these surveys often work with biased samples, as the data collection is often non-probabilistic because of the lack of internet coverage in certain population groups and the self-selection procedure that many online surveys rely on. Some procedures have been proposed to mitigate the bias, such as propensity score adjustment (PSA) and statistical matching. In PSA, propensity to participate in a nonprobability survey is estimated using a probability reference survey, and then used to obtain weighted estimates. In statistical matching, the nonprobability sample is used to train models to predict the values of the target variable, and the predictions of the models for the probability sample can be used to estimate population values. In this study, both methods are compared using three datasets to simulate pseudopopulations from which nonprobability and probability samples are drawn and used to estimate population parameters. In addition, the study compares the use of linear models and Machine Learning prediction algorithms in propensity estimation in PSA and predictive modeling in Statistical Matching. The results show that statistical matching outperforms PSA in terms of bias reduction and Root Mean Square Error (RMSE), and that simpler prediction models, such as linear and k-Nearest Neighbors, provide better outcomes than bagging algorithms.

Keywords: nonprobability surveys; machine learning; matching; propensity score adjustment; sampling

1. Introduction

Surveys are a fundamental tool for data collection in areas like social studies and health sciences. Probability sampling methods have been widely adopted by researchers in those areas, as well as by official statistics. The main reason is that it provides valid statistical inferences about large finite populations by using relatively small samples, based on a solid mathematical theory, with the right combination of a random sample design and an approximately design-unbiased estimator.

Over the last decade, new alternatives to survey sample data have became popular as data sources. Examples are big data and web surveys that have the potential of providing estimates in nearly real time, an easier data access, and lower data collection costs than traditional probability sampling [1]. Very often, the data-generating process of such sources is nonprobabilistic, given that the probability of being part of the sample is not known and/or is null for some groups of the target population, and, as a result, these methods produce nonprobability samples. There are serious issues on the use of nonprobability survey samples; the most relevant is that the data-generating process is unknown and may have serious coverage, nonresponse, and selection biases, which may not be ignorable and could

deeply affect estimates [2]. These biases tend to be more disruptive as the population size gets larger, regardless of the sample size [3].

In order to correct this selection bias produced by non-random selection mechanisms, some inference procedures are proposed in the literature. A first class of methods includes statistical models aiming to predict the non-sampled units of the population [4–6]. Specifying an appropriate super-population model capable of learning the variation of the target variables is important for this model-based method. For this approach, auxiliary features X must be available for each unit of the observed and the unobserved parts of the population. This situation is complicated in practice.

Some studies combine a nonprobability sample with a reference probability sample for constructing models for units in the latter or to adjust selection probabilities. The most important methods for this case are statistical matching and propensity score adjustment (PSA). There are many works studying the properties and performance of PSA [7–12], but there is not much of a bibliography that develops statistical matching in this context.

In this article, we apply machine learning prediction techniques to build statistical matching estimators and compare their performance with PSA estimators. Since there is not a sampling design that allows us to determine the main statistical properties (sampling distribution, expected value, variance, etc.) of random quantities calculated from the non-probability sample, we cannot include theoretical properties of the estimators obtained, but their behavior is studied through simulation studies that also include several propensity score techniques. Although PSA performance was compared with linear calibration in [10] and the combination of PSA with machine learning was already studied in [12], to the best of our knowledge, this is the first time that these methodologies are compared in practice and the first time that machine learning techniques are used for estimation with statistical matching from nonprobability samples.

The description of the conducted study is organized as follows: In Section 2, we introduce the notation and explain the estimation problem that can be solved with the aforementioned methods. In Sections 3 and 4, we describe the mathematical foundations of PSA and statistical matching, respectively, and their properties according to previous research. In Section 5, we briefly explain the ideas behind each of the algorithms tested in the study. In Section 6, we describe the data and the simulation study used to compare the performance of PSA and statistical matching, as well as the metrics used to measure it. Finally, in Sections 7 and 8, we show the results of the study and discuss some of their implications in the comparison between methods.

2. Background

Suppose that the finite population U consists of $i = 1, ..., N$ subjects. Let y be a survey variable and y_i be the y-value attached to the i-th unit,$i = 1, ..., N$.

Let s_v be a volunteer nonprobability sample of size n_v, obtained from $U_v \subset U$ observing the study variable y.

Without any auxiliary information, the population mean $\bar{Y}$ is usually estimated with the unweighted sample mean

$$\hat{\bar{Y}} = \sum_{k \in s_v} \frac{y_k}{n_v} \tag{1}$$

that produces biased estimates of the population mean. The size and direction of the bias depend on the proportion of the population with no chance of inclusion in the sample (coverage) and differences in the inclusion probabilities among the different members of the sample with a non-zero probability of taking part in the survey (selection) [2,13]. The selection bias cannot be estimated in practice for most survey variables of interest.

We consider the situation where there is a probability sample available and compare two inference methods to treat selection biases in a general framework. Let s_r be the reference probability sample selected under the sampling design (s_d, p_d) with π_i the first-order inclusion probability for the i-th individual. Let us assume that in s_r, we observe some other study variables that are common to

both samples, denoted by x. The available data are denoted by $\{(i, y_i, x_i), i \in s_v\}$, and $\{(i, x_i), i \in s_r\}$. We are interested in estimating a linear parameter $\theta_N = \sum_U a_i y_i$, where a_i are known constants. Examples include the population total $T_y = \sum_U y_i$, the population mean $\bar{Y}$, and the population proportion $p_A = \sum_U y_i / N$, where $y_i = 1$ if the unit i belongs to the interest group A, and 0 otherwise.

3. Propensity Score Adjustment

The most popular adjustment method in nonprobability settings is propensity score adjustment (PSA) or propensity weighting. This method, firstly developed by [14], was originally intended to correct the confounding bias in the experimental design context, and it is the most widely used method in practice [2,7–10,12,15–17]. In this approach, the propensity for an individual to participate in the volunteer survey is estimated by binning the data from both samples, s_r and s_v, together and training a machine learning model (usually logistic regression) on the variable δ, with $\delta_k = 1$ if $k \in s_v$ and $\delta_k = 0$ if $k \in s_r$. We assume that the selection mechanism of s_v is ignorable; this is:

$$P(\delta_k = i | y_k, \mathbf{x}_k) = P(\delta_k = i | \mathbf{x}_k), i = 0, 1; k \in s_v. \tag{2}$$

We also assume that the mechanism follows a parametric model:

$$P(\delta_k = 1 | y_k, \mathbf{x}_k) = p_k(\mathbf{x}) = \frac{1}{e^{-(\gamma^T \mathbf{x}_k)} + 1} \tag{3}$$

for some vector γ. We obtain the pseudo maximum likelihood of parameter γ and use the inverse of the estimated response propensity as weight for constructing the estimator [11]:

$$\hat{\theta}_{PSA1} = \sum_{k \in s_v} a_k y_k / \hat{p}_k(\mathbf{x}_k), \tag{4}$$

where $\hat{p}_k(\mathbf{x}_k)$ denotes the estimated response propensity for the individual $k \in s_v$. Alternative estimators can be constructed by slightly modifying the formula in (4) [18]:

$$\hat{\theta}_{PSA2} = \sum_{k \in s_v} (1 - \hat{p}_k(\mathbf{x}_k)) \, a_k y_k / \hat{p}_k(\mathbf{x}_k). \tag{5}$$

Other alternatives involve the stratification of propensities in a fixed number of groups, with the idea of grouping individuals with similar volunteering propensities. For instance, in [7,8], adjustment factors f_c are obtained for the cth strata of individuals:

$$f_c = \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_j^v / \sum_{j \in s_v} d_j^v}, \tag{6}$$

where s_r^c and s_v^c are individuals from the s_r and s_v sample respectively who belong to the cth group, while $d_k^r = 1/\pi_k$ and $d_j^v = 1/\hat{p}_j$ are the design weights for the kth individual of the reference sample and the jth individual of the volunteer sample, respectively. The final weights are:

$$w_j = f_c d_j^v = \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_j^v / \sum_{j \in s_v} d_j^v} d_j^v. \tag{7}$$

The weights are then used in the Horvitz–Thompson estimator:

$$\hat{\theta}_{PSA3} = \sum_{k \in s_v} w_k a_k y_k. \tag{8}$$

The approach used in [9] does not require the calculation of f_c. It only uses the average propensity within each group

$$\overline{\pi}_c = \sum_{k \in s_r^c \cup s_v^c} p_k(\mathbf{x}) / (n_r^c + n_v^c), \tag{9}$$

where n_r^c and n_v^c are the number of individuals from the reference and the volunteer sample, respectively, that belong to the cth group. The mean propensity for each member of the volunteer sample is used in the Horvitz–Thompson estimator:

$$\hat{\theta}_{PSA4} = \sum_c \sum_{k \in s_v^c} a_k y_k / \overline{\pi}_c. \tag{10}$$

PSA in nonprobability online surveys has been proven to be efficient if the selection mechanism is ignorable and the right covariates are used for modeling [7]. If some of these conditions do not apply, the use of PSA can induce biased estimates that would need further adjustments [9]. The combination of PSA and calibration has shown successful results in terms of bias removal [8,10].

Several machine learning models have been suggested as alternatives to logistic regression for the estimation of propensity scores in the experimental design context, with promising results. Ref. [19,20] examined the performance of various classification and regression trees (CART) for PSA in sample balancing. Other applications of machine learning algorithms in PSA involve their use in nonresponse adjustments; more precisely, they have been studied using Random Forests as propensity predictors [21]. Regarding the nonprobability sampling context covered in this study, [12] presented a simulation study using decision trees, k-Nearest Neighbors, Naive Bayes, Random Forests, and a Gradient Boosting Machine that support the view given in [6] about machine learning methods being used for removing selection bias in nonprobability samples. All of those algorithms, along with Discriminant Analysis and Model Averaged Neural Networks, will be used for propensity estimation in this study. Further details can be consulted in Section 5.

4. Statistical Matching

Statistical matching (also known as data fusion, data merging, or synthetic matching) is a model-based approach introduced by [22] and further developed by [23] for nonresponse in probability samples. The idea in this context is to model the relationship between y_k and x_k using the volunteer sample s_V in order to predict y_k for the reference sample.

Suppose that the finite population $\{(i, y_i, x_i), i \in U\}$ can be viewed as a random sample from the superpopulation model:

$$y_i = m(x_i) + e_i, i = 1, 2, ..., N, \tag{11}$$

where $m(x_i) = E_m(y_i|x_i)$ and the random vector $e = (e_1, ..., e_N)'$ is assumed to have zero mean.

Under the design-based approach, the usual estimator of a population's linear parameter is the Horvitz–Thompson estimator given by:

$$\hat{\theta}_{HT} = \sum_{k \in s_r} a_k y_k d_k \tag{12}$$

where $d_k = 1/\pi_k$ is the sampling weight of the unit k that is design-unbiased, consistent for θ, and asymptotically normally distributed under mild conditions [24]. This estimator cannot be calculated because y_k is not observed for the units $k \in s_r$; thus, we substitute y_i by the predicted values from the above model. Thus, the matching estimator is given by:

$$\hat{\theta}_{SM} = \sum_{s_r} a_k \hat{y}_k d_k, \tag{13}$$

where $\hat{y}_k$ is the predicted value of y_k.

The key is how to predict the values of y_k. Usually, the linear regression model is considered for estimation; $E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\beta$ is easy to implement in most of the existent statistical packages, but several drawbacks have to be considered. Parametric models require assumptions regarding variable selection, the functional form and distributions of variables, and specification of interactions. If any of these assumptions are incorrect, the bias reduction could be incomplete or nonexistent. Contrary to statistical modeling approaches that assume a data model with parameters estimated from the data, more advanced machine learning algorithms aim to extract the relationship between an outcome and predictor without an a priori data model. These methods have not been widely applied in the statistical matching literature. Now, we propose the use of machine learning methods as an alternative to linear regression modeling. The ML prediction methods considered in this article are described in the following section.

5. Prediction Modeling

5.1. Generalized Linear Models (GLM)

The most basic regression model consists of calculating coefficients, β, of linear regression based on input data. The coefficients that satisfy the optimality criteria based on minimizing the Ordinary Least Squares are estimated with the formula $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. This method is only stable as long as $\mathbf{X}'\mathbf{X}$ is relatively close to a unit matrix [25]. Quite often, covariates suffer from multicollinearity. For those cases, ridge regression proposes an identity term to control instability: $\beta = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, where $k \geq 0$ can be chosen arbitrarily or via parameter tuning. β can also be considered as the posterior mean of a prior normal distribution with zero mean and a variance of $\mathbf{I}\sigma^2/k$ [26]. From that point of view, Bayesian estimates can be obtained via Gibbs sampling.

Instead, the Least Absolute Shrinkage and Selection Operator (LASSO) regression [27] proposes using a penalty parameter, α, according to the following optimization problem:

$$\begin{array}{l} argmin \sum_{i=1}^{N}(y_i - \alpha - \sum_j \beta_j x_{ij})^2 \\ subject\ to \sum_j |\beta_j| \leq t. \end{array} \tag{14}$$

t is a hyperparameter that forces the shrinkage of the coefficients. In this case, coefficients are allowed to be equal to zero. Therefore, the main difference is that LASSO allows the optimization procedure to select variables, while ridge regression may produce very small coefficients for some cases without reaching zero. Alternatively, LASSO coefficients can be estimated considering the posterior mode of prior Laplace distributions. Bayesian estimates can then be calculated as described in [28]. Ridge and LASSO are both considered standard penalized regression models [29].

For PSA, these methods can be used for estimating the propensities. First, the target variable for the model is defined as $y_i = 1$ if $k \in s_v$ and $y_i = 0$ if $k \in s_r$. The pseudo maximum likelihood can then be optimized via logistic regression or any of its variants described above.

For statistical matching, the target variable for the model is the survey variable itself. Therefore, the model is trained with the volunteer sample and then used to obtain the estimated responses for the reference sample.

5.2. Discriminant Analysis

When the predicted variable is discrete, Discriminant Analysis can be used for classification of individuals. Let y be the dependent variable with K classes, π_k the probability of an individual of belonging to the kth class, $\mathbf{X}$ the matrix of covariates $n \times p$, and $f_k(\mathbf{x})$ the joint distribution of $\mathbf{x}$ conditioned to y taking the kth class. As described in [30], Linear Discriminant Analysis (LDA) assigns an individual the class that maximizes the probability:

$$P(y_i = k|\mathbf{x} = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x}_i)}, k = 1, ..., K. \tag{15}$$

Assuming that $\mathbf{X}|y = k$ follows a multivariate Gaussian distribution $N_p(\mu_k, \Sigma)$, LDA works by assigning an instance the class for which the coefficient $\delta_k(\mathbf{x}_i)$ defined as

$$\delta_k(\mathbf{x}_i) = \mathbf{x}_i^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k) \tag{16}$$

is largest. Note that the decision depends on a linear combination of the multivariate Gaussian distribution parameters; hence, the classifier gets the name of Linear Discriminant Analysis. When used for PSA, $K = 2$ and, as a result, the outcome of LDA is the posterior probability obtained in (15) for the class $\delta = 1$.

LDA can provide good results; however, its simplicity can be a handicap in some cases where relationships between covariates and target are complex, and if the covariates are correlated, its performance gets worse [31]. For these reasons, alternatives considering smoothing, such as Penalized Discriminant Analysis (FDA) or Shrinkage Discriminant Analysis (SDA), can be used. The former expands the covariate matrix and applies penalization coefficients in the calculation of thresholds [32], while the latter performs a shrinkage of covariates similar to that performed in the ridge or LASSO models.

LDA is only suitable for classification and, therefore, it cannot be used for statistical matching when the survey variable is continuous. However, its probabilistic nature makes it appropriate for estimating propensities in PSA, as described above.

5.3. Decision Trees, Bagged Trees, and Random Forests

Decision trees sequentially split the input data via conditional clauses until they reach a terminal node, which assigns a specific class or value. This process results in the following estimation for the expectance $E_m(y_i|\mathbf{x}_i)$:

$$E_m(y_i|\mathbf{x}_i) = \begin{cases} \overline{y(s^{J_1})} & \{i \in s / \mathbf{x}_i \in J_1\} \\ \dots & \dots \\ \overline{y(s^{J_k})} & \{i \in s / \mathbf{x}_i \in J_k\}, \end{cases} \tag{17}$$

where $\overline{y(s^{J_i})}$ denotes the mean of y among the members of the sampled population, s, meeting the criteria of the i-th terminal node.

Bagged trees combine this approach with bagging [33]. Bagging averages the predictions of multiple weak classifiers (in this case, m unpruned trees). In order for them to complement each other, they are trained with a bootstrapped subsample of the complete dataset. Therefore:

$$E_m(y_i|\mathbf{x}_i) = \frac{\sum_{j=1}^{m} \phi_j(\mathbf{x}_i)}{m}, \; \phi_j(\mathbf{x}_i) = \begin{cases} \overline{y(s^{J_1^j})} & \{i \in s / \mathbf{x}_i \in J_1^j\} \\ \dots & \dots \\ \overline{y(s^{J_k^j})} & \{i \in s / \mathbf{x}_i \in J_k^j\}, \end{cases} \tag{18}$$

where $\overline{y(s^{J_i^j})}$ denotes the mean of y among the members of the sampled population, s, meeting the criteria of the i-th terminal node of the j-th tree. This technique is known to improve the accuracy of the predictions [34]. Alternatively, Random Forests can also be used for both regression and classification using weak classifiers [35]. In this algorithm, the input variables for each weak classifier are a random subset of all of the covariates, instead of taking the whole $\mathbf{x}_i$ vector as in bagged trees.

This approach is easy to apply for statistical matching. As usual, a model is trained using the volunteer sample in order to predict a response based on the covariates. Said model is then applied to the reference sample covariates. However, tree-based models are not good for estimating probabilities [36]. They can still be used for PSA, taking the proportion of weak classifiers that agree as the estimated propensity.

5.4. Gradient Boosting Machine

Gradient Boosting Machine (GBM) also works as an ensemble of weak classifiers. Boosting is an iterative process that trains subsequent models, giving more importance to the data for which previous models failed. This idea can be interpreted as an optimization problem [37], and, therefore, it is suitable for the gradient descent algorithm [38]. Then, the estimates for y are:

$$E_m(y_i|\mathbf{x}_i) = v^T J(\mathbf{x}_i), \tag{19}$$

where $J(\mathbf{x}_i)$ stands for a matrix of terminal nodes of m decision trees and v is a vector representing the weight of each tree. GBM has improved previous state-of-the-art models for some cases [39].

GBM can be used for PSA and statistical matching in the same way as the previous ensemble models considered.

5.5. k-Nearest Neighbors

k-Nearest Neighbors is "one of the most fundamental and simple classification methods" [40]. It does not need training. The algorithm simply averages the value of the target variable for the k individuals closest to the estimated individual (its k nearest neighbors), given a certain distance dependent on the covariates. This is:

$$E_m(y_i|\mathbf{x}_i) = \frac{\sum_{j \in s / d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k)})} y_j}{k} \tag{20}$$

where $x_{(1)}$,...,$x_{(n-1)}$ are, respectively, the individuals closest to and furthest from x_i. Choosing the right k is important for the proper performance of the algorithm.

For classification, k-Nearest Neighbors would usually simply predict the most repeated label among its k-nearest neighbors. However, it can instead take into account the proportion in order to estimate probabilities. This idea can be applied for PSA, taking $y_i = 1$ if $k \in s_v$ and $y_i = 0$ if $k \in s_r$, as always. For statistical matching, k-Nearest Neighbors can also be used normally to predict the responses.

5.6. Naive Bayes

The Naive Bayes algorithm is a classifier (that is, it can only be used to predict discrete variables) based on the Bayes theorem. In this study, Naive Bayes has been used only for propensity estimation in PSA. In this case, the Bayes theorem can be used with the probabilities that the participants have of being part of the volunteer sample and the occurrence of a given vector for $\mathbf{X}$, that is, the values of the covariates for a given individual i.

$$\hat{p}_i(\mathbf{x}_i) = \frac{P(\delta_i = 1)P(\mathbf{X} = x_i|\delta_i = 1)}{P(\mathbf{X} = \mathbf{x}_i)}. \tag{21}$$

The Naive Bayes classifier is simple in its reasoning, but can provide precise results in PSA under certain conditions [12]. On the other hand, predictions from Naive Bayes can turn unstable when covariates with high cardinality (e.g., numerical variables) are present, as discrete domains are required for computation of probabilities [41].

As was the case with discriminant analysis, Naive Bayes works naturally with probabilities; therefore, it is suitable for estimating propensities in PSA, but not for statistical matching if the survey variable is continuous.

5.7. Neural Networks with Bayesian Regularization

Neural networks calculate the expectance of y_i as:

$$E_m(y_i|\mathbf{x}_i) = g\left(\sum_{k=1}^{L} v_k f_k(\cdot) + b\right), \tag{22}$$

where g and f_k stand for the activation functions, v_k are the weights of the k-th neuron, and b is the activation threshold [42]. The inputs follow an iterative process through one or more hidden layers until reaching the last layer, which produces the final output. The weights are initialized randomly and then optimized via gradient descent with the backpropagation algorithm [43].

Overfitting is an important problem for neural networks so prior distributions can be imposed to v_k weights as a regularization method. They are then optimized to maximize the posterior density or the likelihood, as described in [42]. Another option is bagging of neural networks, as explained in [44]. The same neural network model is fitted using different seeds, and the results are averaged to obtain the predictions. This approach is known as Model Averaged Neural Networks.

Neural networks have already been considered for superpopulation modeling [45]. They are the state of the art for many domains [46]; Bayesian neural networks in particular are "fairly robust with respect to the problems of overfitting and hyperparameter choice" [47].

Since they work as universal approximators [48], neural networks can be used for PSA and statistical matching in the same way as generalized linear models.

6. The Simulation Study

6.1. Data

All of the experiments were performed using three different populations. In addition, two different sampling strategies were selected for each one in order to recreate the behavior of the estimates under the lack of representativeness of the potentially sampled subpopulation and under selection mechanisms tied to individual features (e.g., voluntariness).

The first population, which will be referred to as P1, corresponds to the microdata of the Spanish Life Conditions Survey (2012 edition) [49]. It collects data about economic and life conditions variables for 28,610 adults living in Spain. We took the mean health, as reported by the individuals themselves on a scale from 1 to 5, as the objective variable to estimate. The algorithms were trained using the 56 most related variables, excluding "health issues in the last six months", "chronic conditions", "household income difficulties", and "civil status" (as they are too correlated with the target variable). The first sampling strategy for this population, which will be referred to as P1S1, was a simple random sampling excluding the individuals without internet access. In the second sampling strategy, P2S2, we also included a propensity to participate in the sample using the formula $Pr(yr) = \frac{yr^2 - 1900^2}{1996^2 - 1900^2}$, where yr is the year in which the individual was born. This way, linear models should have more problems learning the relations.

BigLucy [50], P2, was chosen as the second population. It consists of various financial variables of 85,396 industrial companies of a city for a particular fiscal year. The target variable chosen was the annual income in the previous fiscal year. The algorithms were trained using the size of the company (small, medium, or big), the number of employees, the company's income tax, and whether it is ISO certified. The first sampling strategy for this population, P2S1, was simple random sampling among the companies with SPAM options that are not small companies. This approach tested whether the models were able to correctly estimate the annual income for companies that were not in the training data. The second sampling strategy, P2S2, was simple random sampling among the companies with SPAM options, including a propensity to participate calculated as $Pr(taxes) = min(taxes^2/30, 1)$, where $taxes$ is the company's income tax. This scenario is similar but it implies a quadratic dependence.

The Bank Marketing Data Set [51], P3, is the third population. It includes information about 41,188 phone calls related to direct marketing campaigns of a Portuguese banking institution. Our goal is to predict the mean contact duration. A total of 18 variables were used for training. We excluded two of the dataset variables, the month, and whether the client has subscribed for a term deposit in order to make the inference more difficult. For the first sampling strategy, P3S1, we applied simple random sampling among the clients contacted more than three times. For the second sampling strategy, P3S2, we applied simple random sampling among the clients contacted more than twice. Surprisingly, filtering less led to worse estimations for some cases.

6.2. Simulation

Each population and sampling strategy was simulated using various sample sizes: 1000, 2000, and 5000. The same size is taken for the convenience sample and for the reference sample. For each sample size, 500 simulations were executed. In each simulation, PSA (using weights defined in Formula (4) in Section 3), PSA with stratification (using weights defined in Formula (10) in Section 3), and statistical matching estimates were obtained using several predictive algorithms.

For PSA (with and without stratification), the following classification algorithms were used: Logistic regression (*glm*), generalized linear model via penalized maximum likelihood (*glmnet*), Naive Bayes (*naivebayes*), k-Nearest Neighbors (*knn*), C4.5 decision tree (*J48*), Bagged Trees (*treebag*), Random Forests (*rf*), Gradient Boosting Machine (*gbm*), Model Averaged Neural Network (*avNNet*), Linear Discriminant Analysis (*lda*), Penalized Discriminant Analysis (*pda*), and Shrinkage Discriminant Analysis (*sda*).

For statistical matching, the following regression algorithms were used: linear regression (*glm*), Ridge regression with and without Bayesian priors (*bridge* and *ridge* respectively), LASSO regression via penalized maximum likelihood (*glmnet*), LARS-EN algorithm (*lasso*) and using Bayesian priors on the estimates (*blasso*), k-Nearest Neighbors (*knn*), Bagged Trees (*treebag*), Gradient Boosting Machine (*gbm*) and Bayesian-regularized Neural Networks (*brnn*).

These represent standard variants from different model types: Linear regression, penalized regression, Bayesian models, prototype models, trees, gradient boosting, neural networks, and discriminant analysis. All of the methods were trained using default hyperparameters, except for k-Nearest Neighbors, Naive Bayes, and C4.5, because their performance improved greatly after hypeparameter tuning. Said tuning was performed with bootstrap. The framework used for training, optimization, and prediction was *caret* [52], an R [53] package.

Different metrics are considered for evaluating each scenario: Relative mean bias, relative standard deviation, and relative Root Mean Square Error (RMSE).

$$RBias\ (\%) = \left(\frac{\sum_{i=1}^{500} \hat{p}_{yi}}{500} - p_y \right) \times \frac{100}{p_y} \tag{23}$$

$$RStandard\ deviation\ (\%) = \sqrt{\frac{\sum_{i=1}^{500} (\hat{p}_{yi} - \hat{\hat{p}}_y)^2}{499}} \times \frac{100}{p_y} \tag{24}$$

$$RMSE\ (\%) = \sqrt{RBias^2 + RSD^2}, \tag{25}$$

where p_y is the value of the target variable, $\hat{\hat{p}}_y$ is the mean of the 500 estimations of p_y, and $\hat{p}_{yi}$ is the estimation of p_y in the i-th simulation.

In order to rank each estimator, the mean efficiency, the median efficiency, and the number of times it has been among the best are measured. An estimator is considered to be among the best when its *RMSE* differs from the minimum *RMSE* by less than 1%. The efficiency is defined as follows:

$$Efficiency\ (\%) = \frac{Baseline - RMSE}{Baseline} \times 100, \tag{26}$$

where the baseline is the *RMSE* of using the sample average as the estimation.

To complete the comparison analysis, the results of relative bias, standard deviation, and RMSE were analyzed using linear mixed-effects regression. This approach provides estimates of the effect size of each adjustment method and algorithm. Datasets were considered random effects, while adjustment (matching, PSA, or PSA with propensity stratification) and algorithm (*glm*, *gbm*, *glmnet*, *knn*, and *treebag*, as they were the only algorithms used in all adjustments) were the fixed effects variables. All three response variables take non-negative values, and the interpretation is the same: The lower their value is, the better (less biased and/or less variable) the estimations are. Following this rule, negative Beta coefficients indicate that a given factor is contributing to better estimations, and vice versa for positive coefficients.

7. Results

Tables A1–A3 in the Appendix A show, respectively, the resulting biases, deviations, and RMSEs. In general, it can be observed that statistical matching outperforms PSA, which outperforms PSA with stratification. Nevertheless, all three methods consistently reduce the sample bias.

In terms of machine learning algorithms, basic linear models seem the most robust. Others, like Naive Bayes, can achieve outstanding results, but only for some cases. It is also interesting noting that C4.5 trees can even lead to worse estimations than simply using the sample average.

The final ranking confirming this impression can be seen in Table 1. Statistical matching with linear or ridge regression has the best mean efficiency and has been among the best more than the rest of approaches. This is not a surprise, since their simplicity avoids overfitting, presumably one of the main problems of matching. Ridge regression should be preferred if the data suffer from multicollinearity. Otherwise, linear regression alone is very effective (and faster).

Table 1. Mean and median efficiency (%) of each estimator and the number of times it has been among the best.

	Mean	Median	Best		Mean	Median	Best
matching ridge	61.5	63.8	10	psa gbm	30.7	28.8	3
matching glm	61.5	64.2	10	psa strat naive	30.5	32.1	3
matching glmnet	61	62.8	7	psa strat knn	25.6	24.7	0
matching brnn	57.3	61.7	7	matching lasso	24.6	14.4	3
matching blasso	57.1	59.9	6	psa strat glm	24.5	28.6	1
matching bridge	55.8	61.2	7	psa strat lda	23.4	27.5	0
matching knn	55.8	51.7	3	psa strat sda	23.2	27.2	0
psa glm	46.4	53.8	5	psa strat pda	23.2	27.2	0
psa sda	46.2	51.7	4	psa strat avNNet	21.8	23.4	0
psa glmnet	46.1	53.4	3	psa strat glmnet	21.7	28.1	1
psa lda	46	51.2	3	psa strat gbm	16.8	16.2	0
psa pda	45.7	51.7	4	psa treebag	10.1	11.6	0
psa naive	41.2	56.9	6	psa strat treebag	7.6	3.5	0
psa knn	38.5	42.4	3	psa strat rf	3.6	4.4	0
psa avNNet	34.2	33.2	0	psa rf	−4.5	7.8	0
matching gbm	32.2	34.9	0	psa strat J48	−23.9	3.8	0
matching treebag	31.4	49.1	1	psa J48	−36.7	7.9	0

The results of linear mixed-effects modeling can be consulted in Tables 2–4. It is noticeable how linear models, LASSO with LARS-EN algorithm, and k-Nearest Neighbors outperform Bagged Trees in all metrics (modulus of relative bias, relative standard deviation, and RMSE), while there is no evidence that GBM is different from any of them. Regarding adjustment methods, PSA (both with and without propensity stratification) showed significantly more bias and RMSE than statistical matching. In the case of standard deviation, there is also evidence that PSA without propensity stratification provides

higher values (deviation) than matching. Altogether, these results would indicate that matching has a larger effect on bias reduction than PSA.

Table 2. Linear mixed-effects model on the modulus of relative bias ($|RBias|$) considering adjustment methods and algorithms as fixed effects and datasets as random effects. Reference levels: Bagged Trees (*treebag*) algorithm and matching adjustment.

Coefficient	**Estimate**	**Std. Error**	**D. f.**	***t* Value**	**IC 95%**	***p*-Value**
(Intercept)	12.755	8.47	5.1276	1.506	[−8.857; 34.367]	0.19104
glm	−3.359	1.224	258.00	−2.745	[−5.769; −0.950]	0.00647
glmnet	−2.753	1.224	258.00	−2.250	[−5.163; −0.343]	0.02530
knn	−2.960	1.224	258.00	−2.419	[−5.370; −0.551]	0.01624
gbm	−0.624	1.224	258.00	−0.510	[−3.034; 1.785]	0.61042
psa	6.844	0.948	258.00	7.221	[4.978; 8.710]	5.79×10^{-12}
psa strat	9.601	0.948	258.00	10.129	[7.734; 11.467]	$<2 \times 10^{-16}$
Group	**Variance**	**Std. Dev.**				
Dataset	424.21	20.596				
Residual	40.42	6.358				
Dataset	**Sampling**	**Intercept**				
P1	P1S1	1.590				
P1	P1S2	4.522				
P2	P2S1	53.693				
P2	P2S2	13.029				
P3	P3S1	4.110				
P3	P3S2	−0.414				

Table 3. Linear mixed-effects model on relative standard deviation considering adjustment methods and algorithms as fixed effects and datasets as random effects. Reference levels: *treebag* algorithm and Matching adjustment.

Coefficient	**Estimate**	**Std. Error**	**D. f.**	***t* Value**	**IC 95%**	***p*-value**
(Intercept)	3.686	0.518	14.41	7.112	[2.578; 4.795]	4.5×10^{-6}
glm	−2.095	0.430	258.00	−4.866	[−2.942; −1.247]	2.0×10^{-6}
glmnet	−2.224	0.430	258.00	−5.167	[−3.071; −1.376]	4.8×10^{-7}
knn	−1.986	0.430	258.00	−4.614	[−2.833; −1.138]	6.2×10^{-6}
gbm	−2.020	0.430	258.00	−4.694	[−2.868; −1.173]	4.4×10^{-6}
psa	0.623	0.333	258.00	1.869	[−0.033; 1.280]	0.0628
psa strat	0.404	0.333	258.00	1.213	[−0.252; 1.061]	0.2262
Group	**Variance**	**Std. Dev.**				
Dataset	0.834	0.913				
Residual	5.002	2.236				
Dataset	**Sampling**	**Intercept**				
P1	P1S1	2.479				
P1	P1S2	2.703				
P2	P2S1	4.391				
P2	P2S2	4.064				
P3	P3S1	4.226				
P3	P3S2	4.254				

Table 4. Linear mixed-effects model on RMSE considering adjustment methods and algorithms as fixed effects and datasets as random effects. Reference levels: *treebag* algorithm and matching adjustment.

Coefficient	Estimate	Std. Error	D. f.	*t* value	IC 95%	*p*-value
(Intercept)	14.092	8.374	5.13	1.683	[−7.271; 35.455]	0.15174
glm	−4.094	1.222	258.00	−3.351	[−6.500; −1.688]	0.00093
glmnet	−3.515	1.222	258.00	−2.877	[−5.920; −1.109]	0.00435
knn	−3.639	1.222	258.00	−2.978	[−6.045; −1.233]	0.00317
gbm	−1.288	1.222	258.00	−1.054	[−3.694; 1.118]	0.29272
psa	6.744	0.946	258.00	7.126	[4.880; 8.607]	1.0×10^{-11}
psa strat	9.246	0.946	258.00	9.771	[7.383; 11.110]	$<2 \times 10^{-16}$
Group	**Variance**	**Std. Dev.**				
Dataset	414.5	20.359				
Residual	40.3	6.348				
Dataset	**Sampling**	**Intercept**				
P1	P1S1	2.436				
P1	P1S2	5.367				
P2	P2S1	54.549				
P2	P2S2	14.515				
P3	P3S1	5.949				
P3	P3S2	1.737				

8. Discussion

Nonprobability samples are increasingly common due to the growing internet penetration and the subsequent rise of online questionnaires. These questionnaries are a faster, less expensive, and more comfortable method of information collection in comparison to traditional ones. However, samples obtained with this technique deal with several sources of bias: Despite the increasing internet penetration, large population groups (less educated or elderly people) are still not properly represented. In addition, questionnaires are often administered with non-probabilistic sampling methods (e.g., snowballing), which imply that the selection is controlled by the interviewees themselves, causing a selection bias.

In this study, we focus on two of the proposed methods to reduce biases produced by nonprobability sampling: PSA and matching. We also compare the outcomes when the predictive modeling, required in both methods, is done through linear regression and through machine learning algorithms. PSA and matching require a probability sample on which the target variable has not been measured. The unit sampling performed in the simulations captures different self-selection scenarios in nonprobability sampling, while probability samples are drawn by simple random sampling with no sources of bias. This canonical representation is not usual, as reference samples are mostly drawn with complex sampling methods and the amount of bias is non-null. Further research could take into account these imperfect situations.

Results show that statistical matching provides better results than PSA on bias reduction and RMSE, regardless of the dataset and selection mechanism. In addition, linear models and k-nearest neighbors provided, on average, better results in terms of bias reduction than more complex models, such as GBM and Bagged Trees. These results are relevant since, even though there are comparative studies between adjustment techniques in nonprobability surveys [11,54], to the best of our knowledge, no comparison has been done before between these two methods.

Before closing, several limitations of our analysis should be mentioned. Given that the datasets used for simulation are real-life examples, we cannot ensure whether a selection bias mechanism is Missing At Random (MAR) or Missing Not At Random (MNAR), as the causality relationships are not known. It is known that the selection mechanism makes a difference in terms of how challenging bias reduction can be, but, in this study, it was not possible to assess.

In the near future, it is planned to explore how to combine PSA and statistical matching techniques. Shrinkage is a natural way to improve the available estimates in terms of the mean squared error that has been used by many authors in other contexts (e.g., [55–57]). The idea is to shrink the estimator $\hat{\theta}_{SM}$ towards the estimator $\hat{\theta}_{PSA}$ and obtain $\hat{\theta}_{srk} = K\hat{\theta}_{SM} + (1-K)\hat{\theta}_{PSA}$, where K is a constant satisfying $0 < K < 1$.

Another way can be considered by taking into account that most machine learning models allow weighting of the data used for training. Therefore, the weights obtained via PSA for the volunteer sample can be used when training the model used for statistical matching, since it is trained with said sample.

Author Contributions: Conceptualization, M.d.M.R.; formal analysis, M.d.M.R.; funding acquisition, M.d.M.R.; methodology, R.F.-G.; software, L.C.-M.; supervision, R.F.-G.; validation, L.C.-M.; writing, L.C.-M. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the Ministerio de Economia, Industria y Competitividad, Spain, under Grant MTM2015-63609-R and, in terms of the third author, an FPU grant from the Ministerio de Ciencia, Innovación y Universidades, Spain (FPU17/02177).

Acknowledgments: We would like to thank the anonymous referees for their remarks and comments that have improved the presentation and the contents of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Relative mean bias (%) of each estimator for each population, sampling method, and sample size. The best values among the methods are shown in bold.

	P1S1			P1S2			P2S1			P2S2			P3S1			P3S2		
Estimator	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**
baseline	8.4	8.5	8.5	12.9	12.8	12.8	70.6	70.4	70.4	32.7	32.6	32.9	13.4	13.2	13.3	5.9	5.8	5.9
matching blasso	3.3	3	2.7	6.6	5.8	5.2	24.6	**24.6**	**24.6**	12.6	12.6	12.6	8.9	5.1	2.7	3	**0.4**	2
matching bridge	3.4	3.1	2.8	6.9	6.1	5.3	24.5	**24.6**	**24.5**	12.3	12.6	12.7	9.1	4.8	2.3	4	**0.5**	1.5
matching brnn	**2.4**	2.3	2.2	**4.8**	4.5	4.3	25.3	**24.7**	**24.7**	13.3	13.3	13.3	4.3	2.5	**0**	**0.4**	2.1	4
matching gbm	5.3	5.3	5.4	8.8	8.9	8.9	46.2	46.7	47.1	17	17.4	17.5	**0.4**	2.3	5	5.8	5.3	3.2
matching glm	**2.6**	2.4	2.5	**4.7**	4.7	4.6	24.4	**24.6**	**24.7**	12.6	12.7	12.7	**0.6**	**0.5**	0.8	3.4	3.3	3
matching glmnet	**2.6**	2.6	2.5	**4.8**	4.7	4.8	25.5	25.6	25.6	12.7	12.9	12.8	**0.7**	0.8	0.9	3.2	3.2	3.2
matching knn	5	4.4	3.6	7.5	6.9	6.1	34.2	34.1	34	7	5.8	4.2	6.1	5.8	5.5	0.9	**0.5**	**0.3**
matching lasso	7.1	7.2	7.2	10.8	11	11.1	65.5	65.5	65.5	29.9	29.8	29.7	6.8	6.4	6.6	1.8	1.3	1.2
matching ridge	**2.4**	2.5	2.5	**4.7**	4.7	4.7	24.5	**24.6**	**24.7**	12.6	12.8	12.7	0.5	0.9	0.8	3.2	3.1	3.2
matching treebag	3.5	3.7	4.3	6.5	6.1	6.4	45.5	45.7	46	16.4	16.6	16.6	6.6	**0.1**	8.8	10.8	6	1.7
psa avNNet	3.9	4.2	3.9	6.3	6.7	6.6	66.8	64.2	62.6	8.6	6.9	10	10.5	8.9	10	5.1	4.5	4.6
psa gbm	5.9	5.7	5.4	9.4	9.2	9	66.1	66.6	67.2	15.1	15.3	15.6	10.9	11.1	11.5	1.3	1.3	1.4
psa glm	3.4	3.5	3.5	5.6	5.8	5.8	67.7	68	68.1	15.1	15.1	15.1	4.8	5.2	5.1	1.3	1.3	1.3
psa glmnet	3.7	3.6	3.6	5.9	6	5.9	66.6	66.7	66.9	15.2	15.1	15.1	5.4	5.6	5.4	1.2	1.1	1.1
psa J48	4.6	5	5.2	9.5	10.5	10.9	70.6	70.4	69.6	23.5	22.1	15.1	19.3	22.3	23.4	4.2	2.2	2
psa knn	4.3	4.2	4.1	7.3	7.1	7	68.4	68.5	66.1	18.8	18.2	13.2	7	7.5	7.9	**0.6**	**0.6**	**0.6**
psa lda	3.7	3.6	3.6	6.1	6.2	6.3	67.3	67.2	67.1	14.8	14.9	14.8	6.5	6.2	6.4	**0.2**	**0.3**	**0.2**
psa naive	**2.2**	**1**	**0.1**	**4.7**	**3.6**	**2.5**	**22.4**	**24.8**	29.8	**4.5**	6.4	7.1	3.4	4.1	4.3	4.3	4.4	5.3
psa pda	3.6	3.6	3.6	6.2	6.3	6.2	67.7	67.2	67	14.8	14.9	14.8	6.8	6.4	6.3	**0.5**	**0.2**	**0.3**
psa rf	7	7.1	7.4	11	11.1	11.4	117.2	125.3	134.9	26.4	30.7	30.9	11.6	11.9	12.4	4.1	4.8	5.1
psa sda	3.5	3.7	3.6	6.2	6.2	6.2	67.1	67	67	14.8	14.7	14.8	6.3	6.2	6.3	**0.5**	**0.3**	**0.3**
psa treebag	6.8	7	7.3	10.9	11	11.4	66.2	66	59.6	13.3	23.4	25.6	12.1	12.4	12.7	4.4	4.7	5.6
psa strat avNNet	6.4	6.5	6.5	9.4	9.5	9.4	68.1	68.5	67.6	17.9	16.4	18.3	12.8	13	12.8	3.5	3.2	3.5
psa strat gbm	7.2	7.1	7	11	10.7	10.6	66.3	65.2	64.1	22.8	22.9	23	12.2	12.2	12	3.9	4.2	4.2
psa strat glm	6.2	6.1	6.1	9	8.8	8.8	71.3	69.6	66.9	23	22.9	23	10.8	10.7	10.8	2.9	3	3
psa strat glmnet	6.3	6.2	6.1	9.1	8.9	8.9	77.3	78.7	81	23	22.9	23	10.9	10.9	10.9	3.1	3	3
psa strat J48	6.2	6.9	6.9	10.5	11.5	11.8	70.6	70.4	70	27.4	26.9	22.5	14.8	17.7	18.8	**0.5**	**0.2**	1.4
psa strat knn	6.7	6.6	6.6	9.1	9.1	9.1	69.5	69.7	67.6	16.9	15.5	11.2	11.4	11.1	11.1	3.4	3.6	3.5
psa strat lda	6.2	6.2	6.2	9.1	9.1	9	71.4	69.9	67.2	22.7	22.7	22.8	10.5	10.5	10.5	3.4	3.5	3.6
psa strat naive	5.6	5.5	5.1	6.4	5.8	5	79.8	79.2	77.4	**3.4**	**3.4**	**2.9**	10.4	10.4	10.4	2.5	1.9	**0.3**
psa strat pda	6.2	6.1	6.1	9.2	9.1	9.1	72	70	67.5	22.8	22.7	22.8	12.6	12.8	12.9	3.6	3.5	3.6
psa strat rf	7.4	7.5	8.1	12.5	12.4	12.1	82	83.3	85	21.4	26.8	30.2	10.5	10.6	10.5	5.5	5.5	5.6
psa strat sda	6.3	6.2	6.1	9.3	9.1	9.1	70.8	70	66.7	22.8	22.7	22.8	12.9	13.2	13.1	3.4	3.6	3.6
psa strat treebag	7.9	7.9	8.1	12.4	12.4	12.4	68.4	68.7	67.1	6.9	20.5	24.3	9.8	8.9	11.2	5.9	5.6	5.7

Table A2. Relative deviation (%) of each estimator for each population, sampling method, and sample size. The best values among the methods are in bold.

	P1S1			**P1S2**			**P2S1**			**P2S2**			**P3S1**			**P3S2**		
Estimator	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**
baseline	1.1	0.8	0.5	1.3	0.9	0.5	1.7	1.2	0.7	2.4	1.7	1	3.1	1.9	0.8	3	2.1	1.2
matching blasso	1.5	1.1	0.6	1.8	1.3	0.8	1.5	1.2	0.7	1.7	1.2	0.7	3.9	3.1	1.5	3.8	2.7	1.6
matching bridge	1.6	1.1	0.7	2	1.4	1.1	1.7	2.3	0.7	1.8	1.2	0.7	5.1	3.5	1.6	4.1	3	1.6
matching brnn	1.6	1	0.6	1.9	1.4	0.8	2.1	1.1	0.7	1.9	1.2	0.8	6.2	5.1	2.3	5.3	4	1.6
matching gbm	1.4	1	0.6	1.7	1.2	0.7	1.3	0.9	0.6	1.7	1.2	0.7	7.6	4.3	1.6	6.6	5.1	3.6
matching glm	1.4	1	0.6	2	1.4	0.9	1.6	1.1	0.6	1.7	1.2	0.7	4.1	2.7	1.2	4	2.6	1.4
matching glmnet	1.5	1	0.6	2	1.3	0.8	1.6	1.1	0.6	1.6	1.2	0.7	3.9	2.8	1.1	4	2.5	1.4
matching knn	1.6	1	0.7	2	1.4	0.9	1.3	0.8	0.6	1.7	1.3	0.9	4.3	2.8	1.3	3.9	2.8	1.5
matching lasso	1.2	0.8	0.5	1.4	1	0.6	1.7	1.2	0.7	2	1.4	0.9	3.9	2.5	1.1	3.5	2.4	1.4
matching ridge	1.6	1	0.6	1.9	1.3	0.8	1.5	1.1	0.7	1.8	1.2	0.7	4	2.7	1.2	3.9	2.5	1.4
matching treebag	1.4	1.1	0.7	1.9	1.4	0.9	1.4	1	0.7	1.6	1.2	0.8	11	6.4	1.5	9	5.6	2
psa avNNet	1.6	1.1	0.9	2	1.4	1.1	8.8	14.6	16	6.8	3.2	5.3	5	6	3.9	3.4	3	2
psa gbm	1.3	0.9	0.5	1.6	1	0.7	2.7	1.8	1.1	1.8	1.3	0.7	3.9	2.6	1.1	4.1	2.7	1.4
psa glm	1.5	1	0.6	2	1.4	0.8	3	2	1.3	1.7	1.2	0.8	4.1	2.7	1.2	4	2.6	1.3
psa glmnet	1.5	1	0.6	2	1.3	0.8	2.7	1.9	1.3	1.7	1.2	0.8	4	2.6	1.1	3.7	2.6	1.4
psa J48	2.6	1.8	1.2	2.9	2.1	1.2	1.7	1.2	1.6	8.7	8.5	6.5	21.9	19.9	18.6	25.2	17.9	11.5
psa knn	1.9	1.2	0.7	2.1	1.6	0.9	3.4	2.6	1.4	3.4	2.5	1.6	4.9	3.3	1.5	4.5	3.2	1.9
psa lda	1.5	1	0.6	1.9	1.2	0.8	3.6	2.6	1.6	1.9	1.2	0.7	3.5	2.4	1	3.6	2.4	1.3
psa naive	4.3	2.3	1.2	3.4	2.4	1.4	8.4	8.4	5.6	17.2	4.8	0.8	13.1	8	3.8	10.7	6.4	5.9
psa pda	1.5	1	0.5	1.8	1.2	0.7	3.9	2.8	1.6	1.7	1.2	0.7	3.6	2.3	1	3.6	2.6	1.3
psa rf	1.3	1	0.6	1.5	1.1	0.7	11.7	9.7	8.7	5.2	3.6	3.2	5.2	3.6	2.4	4.5	3.3	2.1
psa sda	1.4	1	0.6	1.8	1.2	0.7	3.9	2.7	1.6	1.7	1.2	0.7	3.5	2.4	1	3.6	2.5	1.3
psa treebag	1.6	1.2	1.3	1.8	1.2	0.8	9.7	13.9	8.8	21.1	8.7	4.2	5.7	4.3	2.8	5.3	4	2.4
psa strat avNNet	1.2	0.8	0.5	1.6	1.1	0.7	5.4	5.5	7.2	3.6	1.9	3.7	3.2	2.3	1.1	4	2.7	1.6
psa strat gbm	1.2	0.8	0.5	1.4	1	0.6	4.9	4.4	3.6	1.5	1.1	0.8	3.5	2.2	0.9	3.5	2.2	1.2
psa strat glm	1.2	0.8	0.5	1.5	1	0.6	7.8	6.7	4.5	1.6	1.1	0.8	3.2	2.1	0.9	3.5	2.1	1.2
psa strat glmnet	1.2	0.8	0.5	1.5	1.1	0.6	6.4	4.9	2.5	1.7	1.1	0.7	3.2	1.9	0.9	3.3	2.2	1.2
psa strat J48	2.2	1.1	0.7	2.4	1.5	0.8	2.1	1.4	4.2	5.4	4.9	4	15.4	12.3	10.5	20.1	14.6	10
psa strat knn	1.2	0.9	0.5	1.7	1.2	0.7	2.7	1.9	1.4	5	3.7	2	3.2	2.3	1.2	3.4	2.5	1.3
psa strat lda	1.2	0.8	0.5	1.5	1	0.6	7.1	6.6	4.8	1.7	1.2	0.8	3.2	2	0.9	3.3	2.3	1.3
psa strat naive	1.6	1.1	0.6	2.7	1.7	0.9	4.1	3.6	2.7	2.8	2	1.3	3.1	2	0.9	6	4.8	4
psa strat pda	1.2	0.8	0.5	1.4	1	0.6	7.6	6.8	4.9	1.7	1.3	0.8	2.9	1.9	0.9	3.2	2.2	1.2
psa strat rf	1.2	0.8	0.4	1.4	0.9	0.6	3.2	2.5	1.7	4.2	3.3	2.9	3.1	1.9	0.9	3.3	2.1	1.2
psa strat sda	1.2	0.8	0.5	1.5	1.1	0.6	7.5	7.1	4.4	1.8	1.3	0.8	3	2	1	3.3	2.2	1.2
psa strat treebag	1.2	0.8	0.5	1.4	0.9	0.6	5.6	7.7	4.5	20.8	8.8	4	10.5	10.4	3.7	3.1	2.2	1.3

Table A3. Relative Root Mean Square Error (RMSE) (%) of each estimator for each population, sampling method, and sample size. The best values among the methods are shown in bold.

	P1S1			P1S2			P2S1			P2S2			P3S1			P3S2		
Estimator	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**	**1000**	**2000**	**5000**
baseline	8.5	8.5	8.5	12.9	12.9	12.8	70.6	70.4	70.5	32.8	32.7	32.9	13.8	13.4	13.3	6.6	6.2	6
matching blasso	3.6	3.2	2.8	6.9	6	5.3	24.7	**24.6**	**24.6**	12.7	12.7	12.6	9.7	6	3	4.8	**2.7**	2.6
matching bridge	3.8	3.3	2.8	7.1	6.3	5.4	24.6	**24.7**	**24.5**	12.5	12.6	12.7	10.4	5.9	2.8	5.7	3	2.2
matching brnn	**2.9**	**2.5**	2.5	**5.2**	**4.7**	4.4	25.4	**24.7**	**24.7**	13.4	13.4	13.3	7.6	5.6	2.3	5.3	4.5	4.3
matching gbm	5.5	5.4	5.4	9	9	8.9	46.2	46.7	47.1	17.1	17.4	17.5	7.6	4.9	5.2	8.8	7.3	4.8
matching glm	**3**	**2.6**	2.5	**5.1**	**4.9**	4.7	**24.4**	**24.6**	**24.7**	12.7	12.7	12.7	**4.2**	**2.8**	**1.4**	5.2	4.2	3.3
matching glmnet	**3**	**2.8**	2.6	**5.2**	**4.9**	4.9	25.5	25.6	25.6	12.8	12.9	12.8	**4**	**2.9**	**1.4**	5.2	4	3.4
matching knn	5.3	4.6	3.6	7.8	7	6.2	34.2	34.1	34	7.2	6	4.3	7.5	6.5	5.7	**4**	**2.9**	**1.5**
matching lasso	7.2	7.3	7.3	10.9	11.1	11.1	65.6	65.5	65.5	30	29.8	29.7	7.8	6.8	6.7	**3.9**	**2.7**	**1.8**
matching ridge	**2.9**	**2.7**	2.6	**5.1**	**4.9**	4.8	**24.5**	**24.6**	**24.7**	12.7	12.9	12.7	**4**	**2.9**	**1.5**	5.1	4	3.5
matching treebag	3.8	3.8	4.3	6.7	6.2	6.4	45.5	45.7	46	16.5	16.6	16.7	12.8	6.4	9	14	8.2	2.7
psa avNNet	4.2	4.3	4	6.6	6.8	6.7	67.4	65.8	64.6	11	7.6	11.3	11.7	10.8	10.7	6.1	5.4	5
psa gbm	6.1	5.8	5.5	9.5	9.3	9	66.1	66.6	67.2	15.2	15.3	15.6	11.6	11.4	11.5	4.3	3	2
psa glm	3.7	3.6	3.5	6	6	5.9	67.8	68.1	68.1	15.2	15.1	15.2	6.4	5.8	5.2	4.2	3	1.9
psa glmnet	4	3.8	3.7	6.2	6.1	6	66.7	66.8	66.9	15.3	15.1	15.1	6.7	6.2	5.6	**3.9**	**2.8**	**1.8**
psa J48	5.3	5.3	5.4	9.9	10.7	10.9	70.6	70.4	69.6	25	23.7	16.4	29.2	29.9	29.9	25.5	18.1	11.7
psa knn	4.7	4.4	4.2	7.6	7.3	7.1	68.5	68.5	66.1	19.1	18.4	13.3	8.5	8.2	8	4.5	3.3	2
psa lda	4	3.7	3.6	6.4	6.3	6.3	67.4	67.2	67.2	14.9	14.9	14.8	7.4	6.7	6.4	**3.6**	**2.4**	**1.3**
psa naive	4.8	**2.6**	**1.2**	5.8	4.3	**2.9**	**23.9**	26.2	30.3	17.8	8	7.1	13.5	9	5.7	11.6	7.8	8
psa pda	3.9	3.7	3.6	6.5	6.5	6.3	67.8	67.2	67	14.9	15	14.9	7.7	6.8	6.3	**3.7**	**2.6**	**1.3**
psa rf	7.1	7.1	7.4	11.1	11.2	11.4	117.8	125.7	135.2	26.9	30.9	31.1	12.7	12.4	12.7	6.1	5.8	5.5
psa sda	3.8	3.8	3.7	6.5	6.3	6.3	67.2	67.1	67	14.9	14.8	14.8	7.3	6.6	6.4	**3.6**	**2.5**	**1.3**
psa treebag	7	7.1	7.4	11.1	11.1	11.4	66.9	67.4	60.2	25	24.9	26	13.3	13.1	13	6.9	6.1	6.1
psa strat avNNet	6.5	6.6	6.5	9.5	9.6	9.4	68.3	68.7	68	18.3	16.5	18.6	13.2	13.2	12.8	5.3	4.2	3.8
psa strat gbm	7.3	7.2	7	11.1	10.8	10.7	66.5	65.3	64.2	22.9	22.9	23	12.7	12.4	12	5.2	4.8	4.4
psa strat glm	6.3	6.2	6.2	9.1	8.8	8.8	71.7	69.9	67.1	23.1	22.9	23	11.2	10.9	10.8	4.5	3.7	3.2
psa strat glmnet	6.4	6.2	6.2	9.2	9	9	77.5	78.8	81	23.1	22.9	23	11.3	11.1	10.9	4.5	3.8	3.3
psa strat J48	6.6	7	7	10.7	11.6	11.9	70.6	70.5	70.2	27.9	27.3	22.9	21.4	21.5	21.6	20.1	14.6	10.1
psa strat knn	6.8	6.7	6.6	9.3	9.2	9.1	69.5	69.7	67.6	17.6	15.9	11.3	11.8	11.3	11.2	4.8	4.3	3.7
psa strat lda	6.3	6.2	6.2	9.2	9.1	9	71.7	70.2	67.4	22.7	22.7	22.9	11	10.7	10.5	4.8	4.2	3.8
psa strat naive	5.9	5.6	5.2	6.9	6	5.1	79.9	79.3	77.4	**4.4**	**3.9**	**3.2**	10.8	10.6	10.5	6.5	5.2	4
psa strat pda	6.3	6.2	6.2	9.3	9.2	9.1	72.4	70.4	67.7	22.8	22.7	22.9	12.9	12.9	13	4.9	4.2	3.8
psa strat rf	7.5	7.6	8.1	12.5	12.5	12.1	82	83.3	85	21.8	27	30.3	10.9	10.8	10.5	6.4	5.9	5.8
psa strat sda	6.4	6.2	6.2	9.4	9.1	9.1	71.2	70.3	66.8	22.8	22.8	22.8	13.2	13.3	13.1	4.7	4.3	3.8
psa strat treebag	8	8	8.1	12.5	12.4	12.4	68.6	69.1	67.2	21.9	22.3	24.6	14.4	13.7	11.8	6.6	6	5.9

References

1. Rada, D. Ventajas e inconvenientes de la encuesta por Internet. *Pap. Rev. Sociol.* **2012**, *97*, 193–223.
2. Elliott, M.R.; Valliant, R. Inference for nonprobability samples. *Stat. Sci.* **2017**, *32*, 249–264. [CrossRef]
3. Meng, X.L. Statistical paradises and paradoxes in big data (I), Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **2018**, *12*, 685–726. [CrossRef]
4. Royall, R.M.; Herson, J. Robust estimation in finite populations I. *J. Am. Stat. Assoc.* **1973**, *68*, 880–889. [CrossRef]
5. Valliant, R.; Dorfman, A.H.; Royall, R.M. *Finite Population Sampling and Inference: A Prediction Approach*; John Wiley: New York, NY, USA, 2000; No. 04, QA276. 6, V3.
6. Buelens, B.; Burger, J.; van den Brakel, J.A. Comparing inference methods for non-probability samples. *Int. Stat. Rev.* **2018**, *86*, 322–343. [CrossRef]
7. Lee, S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *J. Off. Stat.* **2006**, *22*, 329–349.
8. Lee, S.; Valliant, R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **2009**, *37*, 319–343. [CrossRef]
9. Valliant, R.; Dever, J.A. Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **2011**, *40*, 105–137. [CrossRef]
10. Ferri-García, R.; Rueda, M.D.M. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *Stat. Oper. Res. Trans.* **2018** , *1*, 159–182.
11. Valliant, R. Comparing Alternatives for Estimation from Nonprobability Samples. *J. Surv. Stat. Methodol.* **2020**, *8*, 231–263. [CrossRef]
12. Ferri-García, R.; Rueda, M.D.M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE* **2020**, *15*, e0231500. [CrossRef] [PubMed]
13. Couper, M.P. The future of modes of data collection. *Public Opin. Q.* **2011**, *75*, 889–908. [CrossRef]
14. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983** , *70*, 41–55. [CrossRef]
15. Taylor, H. Does internet research work? *Int. J. Mark.* **2000**, *42*, 1–11. [CrossRef]
16. Taylor, H.; Bremer, J.; Overmeyer, C.; Siegel, J.W.; Terhanian, G. The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 US elections. *Int. J. Mark. Res.* **2001**, *43*, 127–135.
17. Schonlau, M.; Van Soest, A.; Kapteyn, A. Are 'Webographic' or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? *Surv. Res. Methods* **2007**, *1*, 155–163. [CrossRef]
18. Schonlau, M.; Couper, M.P. Options for conducting web surveys. *Stat. Sci.* **2017**, *32*, 279–292. [CrossRef]
19. Lee, B.K.; Lessler, J.; Stuart, E.A. Improving propensity score weighting using machine learning. *Stat. Med.* **2010**, *29*, 337–346. [CrossRef]
20. Phipps, P.; Toth, D. Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Ann. Appl. Stat.* **2012**, *6*, 772–794. [CrossRef]
21. Buskirk, T.D.; Kolenikov, S. Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Surv. Methods Insights Field* **2015**, 1–17. [CrossRef]
22. Rivers, D. Sampling for Web Surveys. In Proceeding of the Joint Statistical Meetings, Salt Lake City, UT, USA, 1 August 2007.
23. Beaumont, J.F.; Bissonnette, J. Variance Estimation under Composite Imputation. The methodology behind SEVANI. *Surv. Methodol.* **2011**, *37*, 171–179.
24. Fuller, W.A. Regression estimation for survey samples. *Surv. Methodol.* **2002**, *28*, 5–24.
25. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, 12, 55–67. [CrossRef]
26. Hsiang, T. A Bayesian view on ridge regression. *J. R. Soc. Ser. D* **1975**, *24*, 267–268. [CrossRef]
27. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]

28. Park, T.; Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
29. Van Houwelingen, J. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat. Neerl.* **2001**, *55*, 17–34. [CrossRef]
30. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin, Germany, 2013.
31. Hastie, T.; Buja, A.; Tibshirani, R. Penalized discriminant analysis. *Ann. Stat.* **1995**, *23*, 73–102. [CrossRef]
32. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin, Germany, 2009.
33. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
34. Sutton, C.D. Classification and regression trees, bagging, and boosting. *Handb. Stat.* **2005**, *24*, 303–329.
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, Bergamo, Italy, 5–7 September 2015; pp. 625–632.
37. Breiman, L. *Arcing the Edge. Tech. Rep., Technical Report 486*; Statistics Department, University of California: Berkeley, CA, USA, 1997.
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 1189–1232. [CrossRef]
39. Touzani, S.; Granderson, J.; Fernandes, S. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy Build.* **2018**, *158*, 1533–1543. [CrossRef]
40. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
41. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer International Publishing: Cham, Switzerland, 2015.
42. Okut, H. Bayesian regularized neural networks for small n big p data. In *Artificial Neural Networks-Models and Applications*; IN-TECH: Munich, Germany, 2016.
43. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
44. Ripley, B.D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Cambridge, UK, 1996.
45. Breidt, F.J.; Opsomer, J.D. Model-assisted survey estimation with modern prediction techniques. *Stat. Sci.* **2017**, *32*, 190–205. [CrossRef]
46. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
47. Baesens, B.; Viaene, S.; Poel, V.d.D.; Vanthienen, J.; Dedene, G. Bayesian neural network learning for repeat purchase modelling in direct marketing. *Eur. J. Oper. Res.* **2002**, *138*, 191–211. [CrossRef]
48. Csáji, B.C. Approximation with artificial neural networks. *Fac. Sci. Etvs Lornd Univ. Hung.* **2001**, *24*, 7.
49. National Institute of Statistics (INE). Life Conditions Survey. Microdata. Available online: https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176807menu=resultados&&idp=1254735976608#!tabs-1254736195153 (accessed on 30 May 2020).
50. Gutiérrez, H.A. *Estrategias de Muestreo Diseño de Encuestas y Estimacion de Parametros*; Universidad Santo Tomas: Bogota, Colombia, 2009.
51. Moro, S.; Cortez, P.; Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.* **2014**, *62*, 22–31. [CrossRef]
52. Kuhn, M. *Caret: Classification and Regression Training*; R Package Version 6.0-81; R Foundation for Statistical Computing: Vienna, Austria, 2018.
53. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
54. Chen, J.K.T.; Valliant, R.L.; Elliott, M.R. Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **2019**, *68*, 657–681. [CrossRef]
55. James, W.; Stein, C. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 311–319.

56. Copas, J.B. The shrinkage of point scoring methods. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1993**, *42*, 315–331. [CrossRef]
57. Arcos, A.; Contreras, J.M.; Rueda, M. A Novel Calibration Estimator in Social Surveys. *Sociol. Methods Res.* **2014** *43*, 465–489. [CrossRef]

Appendix A2

The R package NonProbEst for estimation in non-probability surveys

Rueda, María del Mar; Ferri-García, Ramón; Castro-Martín, Luis (2020)
The R package NonProbEst for estimation in non-probability surveys
The R Journal, vol. 12, number 1, pp. 406–418
DOI: 10.32614/RJ-2020-015

The R Journal

Volume 12/1, June 2020

A peer-reviewed, open-access publication of the R Foundation for Statistical Computing

Contents

STATISTICS & PROBABILITY			
JCR Year	Impact factor	Rank	Quartile
2020	3.984	12/125	Q1

The R Package NonProbEst for Estimation in Non-probability Surveys

by María del Mar Rueda, Ramón Ferri-García and Luis Castro-Martín

Abstract Different inference procedures are proposed in the literature to correct selection bias that might be introduced with non-random sampling mechanisms. The R package **NonProbEst** enables the estimation of parameters using some of these techniques to correct selection bias in non-probability surveys. The mean and the total of the target variable are estimated using Propensity Score Adjustment, calibration, statistical matching, model-based, model-assisted and model-calibratated techniques. Confidence intervals can also obtained for each method. Machine learning algorithms can be used for estimating the propensities or for predicting the unknown values of the target variable for the non-sampled units. Variance of a given estimator is performed by two different Leave-One-Out jackknife procedures. The functionality of the package is illustrated with example data sets.

Introduction

Since sampling theory was formalized in the beginning of the 20th century, surveys have been the main tool to obtain information from society and nature. Traditional surveys used telephone or face-to-face interviews for questionnaire administration, as well as mailing lists. However, the increase of costs, linked to the decrease in response rates, and the development of information and communication technologies have favored the use of new survey modes such as online or smartphone questionnaires. These modes make the sampling process cheaper and faster, but tend to amplify bias from several sources. More precisely, online surveys are often performed through a non-probability sampling, using self-selection procedures without a defined sampling frame where the inclusion probabilities are known or with deficient sampling frames with coverage issues, leading to higher levels of selection bias (Elliott, Michael R. and Valliant, Richard, 2017).

Some techniques can be used to correct selection bias in online non-probability surveys. A good overview of the various methods is given in Elliott, Michael R. and Valliant, Richard (2017). There are three important approaches: the pseudo-design based inference (or pseudo-randomisation (Buelens, Bart et al., 2018)), statistical matching and predictive inference.

In the pseudo-design based inference, the idea is to construct weights to correct for selection bias. The first method is estimating response probabilities and using them in Horvitz-Thompson or Hajek type estimators to account for unequal selection probabilities. The most used method to estimate response probabilities is Propensity Score Adjustment (see e.g. Lee, Sunghee and Valliant, Richard (2009)). This method uses a probability reference sample in addition to a non-probability convenience sample to construct a response propensity model. Sample matching is another approach also applied to tackle selection bias. A predictive model, with the target variable as the dependent variable, is built using data from the non-probability sample. This model is subsequently applied to a probability sample (where the target variable is not measured) to predict values of its individuals for an estimation of the population values. Similarly, predictive methods are based on superpopulation models. In this approach, a predictive model is fitted for the analysis variable from the sample and used to project the sample to the full population. This approach (that can be used with probability and non-probability samples) allows researchers to use the auxiliary information about covariates in different methods for predicting the unknown values. Most of these methods require special software for their implementation. The package **NonProbEst** implements some of these techniques.

The paper is structured as follows. First, we introduce the notation used throughout the paper and we discuss the different ways to do inference for non-probability surveys. In section 2.3 we briefly comment on the usefulness of Machine Learning (ML) Techniques in this context. Then, we describe the R package NonProbEst. In section 2.5 we briefly describe the use of the functions, including suitable examples, for each method.

Statistical methodology

Let U denote a finite population with N units, $U = \{1, \ldots, k, \ldots, N\}$. Let s_V be a volunteer non-probability sample of size n_V, self-selected from an online population U_V which is a subset of the total target population U. Let y be the variable of interest in the survey estimation. Without any auxiliary information, the population total of y, Y, is usually estimated with the following Horvitz-Thompson

type estimator:

$$\hat{Y}_{HT} = \sum_{k \in s_V} w_{vk} y_k \tag{1}$$

being w_{vk} a weight of the unit k set by the researcher to adjust the lack of response, lack of coverage, voluntariness, ... (e.g. by means of post-stratification). A simple choice is $w_{vk} = N/n_V$, that is, consider the sample of volunteers as if it was obtained with a simple random sampling design of the population U.

This estimator has a bias induced by various mechanisms regarding their application. The most important are the selection bias (due to the difference between sampled and nonsampled individuals on the probability to participate in a survey) and the coverage bias (the online population U_v is not the same of the target population U).

The key to successful weighting to remove the bias in non-probability surveys lies in the use of powerful auxiliary information. Auxiliary information can be available in different forms. We distinguish three different cases, called InfoTP, InfoES and InfoEP, depending on the information at hand.

- InfoTP: Only the population totals of the auxiliary variables are known (often called control totals). Possible sources of information are a census of the target population, an administrative register, ... One of the simplest and most frequently used control totals occurs when the information consists of known counts for a set of population groups.
- InfoES: The auxiliary variable values are available for every element in a probability sample. This reference survey is conducted on the same target population than the non-probability survey, with the main difference that the former has a better coverage and higher response rates than the latter, thus it is adequate to represent the behavior that the target population should have when a probability survey is performed on it.
- InfoEP: The auxiliary variable values are available for every element in the whole population. An example of this is when statistical agencies use auxiliary variables specified in different existing registers, for all the elements in the population.

We will now explain the main methods used to treat these biases depending on the type of information that is available.

InfoTP

Calibration

Let $\mathbf{x}_k$ be the value taken on unit k by a vector of auxiliary variables which population total is assumed to be known $\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$. The calibration estimation of Y consists in the computation of a new vector of weights w_k for $k \in s$ which modifies as little as possible the original sample weights, w_{vk}, which have the desirable property of producing unbiased estimations, respecting at the same time the calibration equations

$$\sum_{k \in s_V} w_k \mathbf{x}_k = \mathbf{X}. \tag{2}$$

Given a pseudo-distance $G(w_k, w_{vk})$, the calibration process consists in finding the solution to the minimization problem

$$\min_{w_k} \{ \sum_{k \in s_V} G(w_k, w_{vk}) \} \tag{3}$$

while respecting the calibration equation (2). Several distances were defined in Deville, Jean-Claude and Särndal, Carl Erik (1992), being the linear distance one of the most commonly used. The resulting estimator of Y under the chi-square distance is the general regression estimator

$$Y_{reg} = \sum_{s_V} w_k y_k = \sum_{s_V} d_k y_k + (\mathbf{X} - \sum_{s_V} w_{vk} \mathbf{x}_k)' \hat{B}_{s_V} \tag{4}$$

where $\hat{B}_s$ is

$$\hat{B}_{s_V} = T_s^{-1} \sum_{s_V} w_{vk} \mathbf{x}_k y_k \tag{5}$$

being $T_s = \sum_{s_V} w_{vk} \mathbf{x}_k \mathbf{x}_k'$.

It is proved in Bethlehem, Jelke (2010) that bias can be reduced through calibration only when the non-response due to volunteering has a Missing At Random scheme, while it cannot be equally done in Not Missing at Random situations (which are the most frequent).

InfoSP

Propensity Score Adjustment

The Propensity Score Adjustment method was originally developed by Rosenbaum, Paul R and Rubin, Donald B (1983) which sought to reduce the confounding bias between treatment and control groups in experimental designs. This approach would be considered in sampling research as well in combination with a reference sample (Rubin, Donald B, 1986), but it was not proposed for online surveys until the early 2000's (Taylor, H. et al., 2001).

It is expected that a sample collected by online recruitment would not follow the principles of a probability sampling, especially in those cases that the survey is filled by volunteer respondents. In such a situation, every individual is associated to a probability of participating in the survey which depends on her or his characteristics.

The propensity for an individual to take part on the non-probability survey is obtained by training a predictive model (often a logistic regression) on the dichotomous variable, I_{s_V}, which measures whether a respondent from the combination of both samples took part in the volunteer survey or in the reference survey. Covariates used in the model, $\mathbf{x}$, are measured in both samples (in contrast to the target variable which is only measured in the non-probability sample), thus the formula to compute the propensity of taking part in the volunteer survey with a logistic model, π, can be displayed as

$$\pi(\mathbf{x}) = \frac{1}{e^{-(\gamma^T \mathbf{x})} + 1} \tag{6}$$

for some vector γ, as a function of the model covariates.

We denote by s_R the reference sample and w_{Rk} the original design weight of the k individual in the reference sample

Several options for using the propensity scores in estimation are listed below:

- We can use the inverse of the estimated response propensity as a weight for constructing the estimator (Valliant, Richard, 2020):

$$\hat{Y}_{PSA1} = \sum_{k \in s_V} w_{Vk} y_k / \hat{\pi}(\mathbf{x}_k) = \sum_{k \in s_V} y_k w_k^{PSA1} \tag{7}$$

  where $\hat{\pi}(\mathbf{x}_k)$ is the estimated response propensity for the individual k of the volunteer sample as predicted using covariates $\mathbf{x}$.

- Alternatively, the approach proposed in Schonlau, Matthias and Couper, Mick P. (2017) can be used to obtain weights for a Horvitz-Thompson type estimator using propensity scores. Weights are defined as

$$w_k^{PSA2} = \frac{1 - \hat{\pi}(\mathbf{x}_k)}{\hat{\pi}(\mathbf{x}_k)} \tag{8}$$

  and resulting estimator for the population total is given by

$$\hat{Y}_{PSA2} = \sum_{k \in s_V} y_k w_k^{PSA2} \tag{9}$$

- Valliant, Richard and Dever, Jill A. (2011) use the propensity scores to post-stratify the sample. The process is: sort the combined sample by $\hat{\pi}(\mathbf{x}_k)$; split the combined sample into g classes (g = 5 as the conventional choice following Cochran, William G (1968)), each of which has about the same number of cases in the combined sample; and compute an average propensity, $\bar{\pi}_g$ within subclass g. Use $\bar{\pi}_g$ as the weight adjustment for every person in the subclass. Resulting estimator is:

$$\hat{Y}_{PSA3} = \sum_{g} \sum_{k \in s_{V_g}} w_{Vk} y_k / \bar{\pi}_g = \sum_{g} \sum_{k \in s_{V_g}} y_k w_k^{PSA3} \tag{10}$$

- Following the approach described in Lee, Sunghee and Valliant, Richard (2009) propensity scores are divided in g classes, where all units may have the same propensity score or at least be

in a very narrow range and an adjustment factor is calculated as:

$$f_g = \frac{\sum_{k \in s_{R_g}} w_{Rk} / \sum_{k \in s_R} w_{Rk}}{\sum_{k \in s_{V_g}} w_{Vk} / \sum_{k \in s_V} w_{Vk}} \tag{11}$$

where s_{R_g} is the set of individuals in the reference sample that are in the gth class of propensity scores and s_{V_g} is the set of individuals in the volunteer sample that are in the gth class of propensity scores. Finally, the adjusted weights w^{PSA4} are the product of the original weights and the adjustment factor; following the same notation, the adjusted weight for individual k in s_{V_g} (i. e. the individual k of the gth propensity class in the volunteer sample) is computed as

$$w_k^{PSA4} = w_{Vk} f_g \tag{12}$$

and the estimator is given by

$$\hat{Y}_{PSA4} = \sum_g \sum_{k \in s_{V_g}} y_k w_k^{PSA4} \tag{13}$$

Research findings have shown that PSA successfully removes bias in some situations, but at the cost of increasing the variance (Lee, Sunghee and Valliant, Richard, 2009). Valliant, Richard and Dever, Jill A. (2011) showed that the estimation of a variable using PSA must be complemented with further weighting adjustment in order to make estimates less biased. The use of PSA with further calibration is studied in Lee, Sunghee and Valliant, Richard (2009) and Ferri-García, Ramón and Rueda, Maria del Mar (2018), concluding that calibration adjustments are helpful if they are applied using the right covariates.

Variance estimation in PSA is not a simple issue. Valliant, Richard (2020) proposes an estimator of the variance for an estimator of a mean, $\hat{\bar{y}}$, based on linearization, but this estimator does not take into account the randomness of weight estimation, therefore it will tend to underestimate the variance.

Jackknife's variance estimator (Quenouille, Maurice H (1956)) can be seen as an acceptable alternative in nonprobability samples after applying PSA. Let $\hat{\bar{y}} = \frac{1}{N} \sum_{k \in s_V} w_k^{PSA} y_k$ be the estimator of the mean of y, his Leave-One-Out Jackknife estimator of the variance is given by:

$$\hat{V}(\hat{\bar{y}}) = \frac{n-1}{n} \sum_{j=1}^{n} (\bar{y}_{(j)} - \bar{y})^2 \tag{14}$$

where $\bar{y}_{(j)}$ is the value of the estimator $\hat{\bar{y}}$ after dropping unit j from s_V and where $\bar{y}$ is the mean of values $\bar{y}_{(j)}$.

Given that PSA weights are estimated from the available data, the exclusion of one unit can have an impact on the values of w_i and affect the variability of the estimator. This variability can be taken into account if propensities are recalculated for each of the n Leave-One-Out partitions. Thus a Jackknife estimator with recalculating weights is defined as:

$$\hat{V}_{rw}(\hat{\bar{y}}) = \frac{n-1}{n} \sum_{j=1}^{n} (\bar{y}_{rw(j)} - \bar{y}_{rw})^2 \tag{15}$$

where $\bar{y}_{rw(j)} = \frac{1}{N} \sum_{k \in s_V - \{j\}} w_k^{PSA}(j) y_k$, with $w_k^{PSA}(j)$ the PSA weight obtained from the sample $s_V - \{j\}$ and $\bar{y}_{rw}$ is the mean of values $\bar{y}_{rw(j)}$.

Statistical matching

The statistical matching method was introduced by Rivers, D. (2007). The idea is to model the relationship between y_k and $\mathbf{x}_k$ using the volunteer sample s_V in order to predict y_k for the reference sample. That is, the matching estimator is given by:

$$\hat{Y}_{SM} = \sum_{s_R} \hat{y}_k w_{Rk}$$

being $\hat{y}_k$ the predict value of y_k.

The key is how to predict the values y_k. Usually $\hat{y}_k = \mathbf{x}_k' \hat{\beta}$ being $\hat{\beta} = \sum_{k \in s_V} y_k \mathbf{x}_k / \sum_{k \in s_V} \mathbf{x}_k' \mathbf{x}_k$ but other methods can be considered as donor imputation (Rivers, D., 2007) or fractional donor imputation (Kim, J.K and Fuller, W., 2004).

A major drawback of matching is that the precision of the non-probability sample reduces to

the standard error of the reference sample (Buelens, Bart et al., 2018). These authors also justify that matching is based on strong ignorability assumptions and can lead biased estimators if the assumptions are not met.

InfoUP

The prediction approach is based on superpopulation models, which assume that the population under study $\mathbf{y} = (y_1, ..., y_N)'$ is a realization of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)'$ having a superpopulation model ξ. To incorporate auxiliary information $\mathbf{x}_k$ available for all $k \in U$ on assume a superpopulation for y built on some mean function of $\mathbf{x}$:

$$Y_k = m(\mathbf{x}_k) + e_k, \quad k = 1, ..., N. \tag{16}$$

The random vector $e = (e_1, ..., e_N)'$ is assumed to have zero mean and a positive definite covariance matrix which is diagonal (Y_k are mutually independent).

Using a set of covariates, $\mathbf{x}$, measured in s_V and $\bar{s}_V = U - s_V$ it is possible to estimate the values of y in $\bar{s}_V$ with regression modeling such that the estimated value of y for an individual k can be calculated through the following expression:

$$\hat{y}_k = E_m(y_k|\mathbf{x}_k) \tag{17}$$

m alludes to the specific model which provides the expectation of y_k, and $\mathbf{x}_k$ are the values of the k-th individual in the covariates $\mathbf{x}$.

We can use the auxiliary information in several ways to define several estimators:

- the model-based estimator:

$$\hat{Y}_m = \sum_{k \in s_V} y_k + \sum_{k \in \bar{s}_V} \hat{y}_k \tag{18}$$

- the model-assisted estimator:

$$\hat{Y}_{ma} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s_V} (y_k - \hat{y}_k) w_{Vk} \tag{19}$$

- the model-calibrated estimator:

$$\hat{Y}_{mcal} = \sum_{k \in s_V} y_k w_k^{CAL} \tag{20}$$

  where w_k^{CAL} are such that they minimize $\sum_{k \in s} G\left(w_k^{CAL}, w_{Vk}\right)$, where $G(\cdot, \cdot)$ is a particular distance function, subject to

$$\sum_{k \in s_V} w_k^{CAL} \hat{y}_k = \sum_{k \in U} \hat{y}_k.$$

Usually the linear regression model is used, $E_m(y_k|\mathbf{x}_k) = \mathbf{x}_k'\beta$ and the above estimators can be rewritten as a type of regression estimators.

Prediction estimators need complete information about the auxiliary variables (InfoEP) and can fail if the model is not true, but might potentially be fruitful to correct for selection bias in informative sampling (Buelens, Bart et al., 2018).

Use of machine learning algorithms in non-probability samples

The emerging data sources like Big Data can be used in combination to traditional survey samples for construct more valid inferences. Machine Learning (ML) methods can be used for the matter, given their known advantages in high dimensional environments. There are several types of learning algorithms but for this package we focus on classification and regression. Classification aims to identify the category to which a new observation belongs while regression is used for prediction in real-valuated variables. Both are trained with known observations to make predictions based on some covariates.

There is a vast spectrum of classification and regression algorithms to take into account, starting from the basic linear and logistic regressions and its extensions, like Ridge regression (Hoerl, Arthur E and Kennard, Robert W, 1970). Other examples are decision trees which uses tree-like graphs , like the C4.5 (Quinlan, J Ross, 1993). More modern approaches even build ensembles of decision trees with outstanding results, like XGBoost (Chen, Tianqi and Guestrin, Carlos, 2016). During the last few

years, deep learning models have been dramatically improving the state-of-the-art (LeCun, Yann et al., 2015). However, many other techniques are still being widely used and developed, like some bayesian methods (Park, Trevor and Casella, George, 2008). Having so many different options, choosing the right learning algorithm for each problem is key for obtaining optimal results.

Regarding survey research, the use of ML algorithms has been studied in the last few years for deriving model-assisted estimators (Montanari, Giorgio E and Ranalli, M Giovanna (2007); Baffetta, Federica et al. (2009); Breidt, F Jay et al. (2017)). In the prediction approach ML algorithms uses the sample to train a model capturing the behaviour of a target variable which is to be estimated, and applies it to the nonsampled individuals to obtain population-level estimates. Applications of machine learning algorithms in PSA for nonresponse propensity have been studied for classification and regression trees (Phipps, Polly et al., 2012) and Random Forests (Buskirk, Trent D and Kolenikov, Stanislav, 2015); their efficacy on reducing nonresponse bias in comparison to logistic regression depends on the available covariates and the complexity of the relationships. (Chen, Jack Kuang Tsung et al., 2019) use LASSO for calibrating non-probability surveys. (Buelens, Bart et al., 2018) review existing inference methods to correct for selection bias and recommend adding ML methods to deal with non-probability samples.

NonProbEst allows the use of a wide variety of classification and regression algorithms for model-based, model-assisted and model-calibrated estimators, matching and PSA (which only works with classification). It offers so many alternatives by relying on **caret** (Max Kuhn, 2018), a well known machine learning package.

The R package NonProbEst

The package **NonProbEst** implements in R a set of techniques for estimation in non-probability surveys, using various approaches which correspond to several frameworks. Functions in the package allow to obtain calibration weights via `calib_weights`, propensity scores via `propensities` and matching predictions for a reference sample via `matching`. Propensity scores can be transformed into weights by all of the approaches mentioned in previous sections via functions `lee_weights`, `sc_weights`, `valliant_weights`, `vd_weights`. These weights can be used for estimation of total, mean and proportion of a given target variable measured in a sample using functions `total_estimation`, `mean_estimation`, `prop_estimation`. Alternatively, total and mean can also be calculated using a model-based, a model-assisted or a model-calibrated approach with the functions `model_based`, `model_assisted` and `model_calibrated` respectively. The variance of the estimators can be calculated using the Leave-One-Out Jackknife method, this is, recalculating the set of weights after substracting one unit or not, by means of the functions `generic_jackknife_variance` and `jackknife_variance`, and without recalculating the weights via `fast_jackknife_variance`. Frequentist confidence intervals of the estimates can be directly computed with the `confidence_interval` function.

Calibration weights are obtained using the `calib` function of the **sampling** package (Yves Tillé and Alina Matei, 2016) for g-weights computation. `calib_weights` offers a wrapper for calculation of final weights straight from the dataset. Functions that require prediction techniques, such as `propensities`, `matching`, `model_based`, `model_assisted`, `model_calibrated` and `jackknife_variance`, use the `train` function from the **caret** package (Max Kuhn, 2018). This function allows the user to use any of the algorithms in the large list of functions which are covered by `train`, with the possibility of optimizing hyperparameters for a better performance of the predictors. For propensity estimation, only classification algorithms should be used as the target variable is binary (participation in the probability sample vs participation in the non-probability sample). Case weights are used to balance both classes (for models that accept them). For matching, model-based and model-assisted estimations, algorithms should account for the type of variable of the target feature.

Note that weighting formulas for PSA from Lee, Sunghee (2006) and Valliant, Richard and Dever, Jill A. (2011) require applying a stratification procedure. In both `lee_weights` and `vd_weights` the same procedure is applied: the vector of propensities is sorted increasingly, and the individuals are equally divided in g strata of the same length according to their position in the sorted vector. g is defined by the user, and the procedure results in a vector with the strata number (from 1 to g) to which a given individual corresponds. This stratification avoids errors that could arise from the lack of unique values.

Three datasets are available in the package: `sampleP`, `sampleNP` and `population`. These fictitious datasets were created as described in Ferri-García, Ramón and Rueda, Maria del Mar (2018); `sampleP` represents a probability sample of size $n_r = 500$ extracted by simple random sampling from a frame covering the entire population, while `sampleNP` represents a non-probability sample of size $n_v = 1000$ extracted by simple random sampling from a frame covering only the subpopulation of individuals who have access to Internet. The dataset of the complete population of size $N = 50000$ is

available in `population`. Variables available in each dataset differ, with `sampleNP` having the largest amount of variables. In the aforementioned dataset, three variables (`vote_gen`, `vote_pens`, `vote_pir`) measuring whether an individual would vote to a given party ("gen", "pens" or "pir") in an election or not. Probabilities of voting to party "gen", "pens" or "pir" are higher if the individual is a woman, and elder person and has access to the Internet, respectively. These variables are only measured in `sampleNP`, meaning that adjustment methods have to be applied in order to produce reliable estimates of voting intentions. For the matter, the rest of the available variables in the dataset, which are also included in `sampleP` (except for the language) and `population`, can be used. `education_primaria`, `education_secundaria`, `education_terciaria` are three disjunct variables measuring the education level of the individual (Primary, Secondary or Tertiary Education), while `age` and `sex` measures the numeric age and the gender (0 female, 1 male). Finally, `language` measures whether the individual's native language is the official language or not. The absence of certain variables in the datasets accounts for real situations where not all the information is available at individual level.

It must be mentioned that the use of `jackknife_variance` for calculating the variance of the estimators via Leave-One-Out Jackknife will be computationally slower than the `fast_jackknife_variance` alternative. Recalculating the weights in each iteration means that the weighting procedure has to be repeated as many times as individuals are in the non-probability sample. If Propensity Score Adjustment is used for weighting, the models have to be rebuilt in each iteration, resulting in larger computation times which will depend on the computational costs of the algorithms used for propensity estimation. Note that `generic_jackknife_variance` will behave similarly if the estimator passed as argument involves predictive modelling algorithms or other costly procedures. To show the difference of procedures, we calculated the Leave-One-Out Jackknife estimated variance of the estimator of the mean for the variable `vote_pir` in a non-probability sample of size $n_v = 100$ extracted by simple random sampling on the `sampleNP` dataset, using a probability sample of size $n_r = 100$ extracted by simple random sampling on the `sampleP` dataset as the reference sample data. Considering a population of $N = 50000$, variance estimates of the estimator weighted by PSA using different algorithms were computed, measuring the computation elapsed time. All the calculations were performed in a Intel(R) Core(TM) i7-3770 CPU up to 3.40GHz. Results can be consulted in Table 1

Weight recalculation	PSA algorithm	R function	Elapsed time (seconds)
No	Logistic regression	glm	0.004999876
Yes	Logistic regression	glm	75.56034
Yes	CART	rpart	102.3409
Yes	Random Forest	rf	203.7737
Yes	GBM	gbm	453.731
Yes	Neural Network	nnet	719.733

Table 1: Total elapsed time of Leave-One-Out Jackknife variance estimation under recalculation of weights in each iteration for a set of predictive models, with sample sizes of 100 for both the probability and the non-probability sample

In this example, the variance estimation with recalculations takes more than 15000 times the seconds that it takes without recalculations if logistic regression is the method used for propensity estimation, and almost 144000 times if feed-forward neural networks are used. Time differences might be different depending on the data, the estimator and the algorithm, but they will be largely appreciable in all cases.

In order to ilustrate how the resources in the package can be used for estimation in non-probability surveys, some examples of each adjustment covered by the package are developed in the following section.

Inference in non-probability samples with NonProbEst

InfoTP: Calibration

Suppose that a non-probability sample of 1000 individuals recruited via online surveying is available for estimating the vote intention in a given election. For the matter, `sampleNP` will be used as the non-probability sample data.

```
> library(NonProbEst)
> head(sampleNP)
  vote_gen vote_pens vote_pir education_primaria education_secundaria education_terciaria age sex language
1        0         1        0                  1                    0                   0  66   1        1
```

```
2        0         0        1                  0                    0                  1  30   1         1
3        1         0        0                  0                    1                  0  62   0         1
4        0         0        1                  1                    0                  0  33   0         1
5        0         0        1                  0                    1                  0  30   0         1
6        0         0        0                  1                    0                  0  69   1         1
```

Some auxiliary information is available in the sample; more precisely, individual data on education, age, gender and language (as described in the previous Section) can be used for mitigating the effects of coverage error. Population totals are available for all of these auxiliar variables, as they have been measured for the entire population. They can be retrieved from the `population` dataset:

```
> head(population)
  education_primaria education_secundaria education_terciaria age sex language
1                  0                    1                   0  39   1        1
2                  0                    0                   1  55   0        1
3                  1                    0                   0  35   0        1
4                  1                    0                   0  58   1        1
5                  1                    0                   0  36   1        1
6                  0                    1                   0  61   1        1
> totals <- colSums(population)
> totals
  education_primaria education_secundaria  education_terciaria                 age                 sex            language
               25287                10546                14167             2539340               24430               45429
```

If the variables of which population totals are available are not disjunct, Raking calibration can be applied in order to estimate cell counts and account for the lack of information. This can be done with the `calib_weights` function; in this case, the `Xs` argument were the dataset `sampleNP` selecting the auxiliar variables only. Other arguments involve the totals previously obtained and the initial weights, which allows the user to specify whether sampling design weights were used or not. In the latter case, unitary weights should be provided as a vector of ones of length equal to the number of individuals in the non-probability sample. Population size and method to be used by the `calib` function from **sampling** have to be specified.

```
> covariates <- colnames(sampleNP)[4:9]
> initial_weights <- rep(1, nrow(sampleNP))
> w <- calib_weights(sampleNP[, covariates], totals, initial_weights,
        N = 50000, method = "raking")
```

Once we obtain the weights, estimates for the mean (proportion if the variable is binary) or the total of any variable present in the non-probability sample can be obtained using `mean_estimation` or `total_estimation` respectively. For example, the estimated proportion of votes for each party can be obtained with the following code:

```
> mean_estimation(sampleNP, w, "vote_gen", N = 50000)
  vote_gen
0.09824163
> mean_estimation(sampleNP, w, "vote_pens", N = 50000)
vote_pens
0.3726149
> mean_estimation(sampleNP, w, "vote_pir", N = 50000)
 vote_pir
0.3905399
```

If these estimates are compared to those which would be obtained if no adjustment was used, the effect of calibration is notorious. As the presence of "gen" voters in the sample is MCAR, estimates do not differ, but in the case of "pens" voters whose presence is MAR, the calibration approach gives a larger estimate which can be explained by the fact that the overrepresentation of younger people in the sample has been corrected up to a point. To a much lesser extent, this correction is also noticeable in the estimation of vote to "pir" (presence of their voters in the sample is NMAR).

```
> sum(sampleNP$vote_gen)/nrow(sampleNP)
[1] 0.096
> sum(sampleNP$vote_pens)/nrow(sampleNP)
[1] 0.346
> sum(sampleNP$vote_pir)/nrow(sampleNP)
[1] 0.404
> sum(sampleNP$vote_gen)/nrow(sampleNP) -
+        mean_estimation(sampleNP, w, "vote_gen", N = 50000)
   vote_gen
-0.00224163
> sum(sampleNP$vote_pens)/nrow(sampleNP) -
```

```
+        mean_estimation(sampleNP, w, "vote_pens", N = 50000)
  vote_pens
-0.02661494
> sum(sampleNP$vote_pir)/nrow(sampleNP) -
+        mean_estimation(sampleNP, w, "vote_pir", N = 50000)
  vote_pir
0.01346014
```

The variance of the estimates can be assessed through Leave-One-Out Jackknife, both with or without reweighting in each iteration. In the former case, a function must be created by the user for such a task. In the following lines, a function example is developed for estimating the variance on the estimation of the proportion of votes for the "pir" party:

```
### Leave-One-Out Jackknife variance estimation with reweighting
> estimator <- function(s){
     initial_weights <- rep(1, nrow(s))
     w <- calib_weights(s[,covariates], totals, initial_weights, N = 50000,
               method = "raking")
     return(mean_estimation(s, w, "vote_pir", N = 50000))
     }
> v_r <- generic_jackknife_variance(sampleNP, estimator, N = 50000)
> v_r
[1] 0.0003352199
### Leave-One-Out Jackknife variance estimation without reweighting
> v_nr <- fast_jackknife_variance(sampleNP, w,
     estimated_vars = "vote_pir", N = 50000)
> v_nr
    vote_pir
0.0003189449
```

These estimates of the variance can be used for the construction of confidence intervals for the estimation of the proportion via `confidence_interval` function. This function requires the point estimator and the standard deviation as arguments, with the option to fix the confidence level. If not specified by the user, the confidence interval is calculated at 95% confidence level.

```
> ic_r <- confidence_interval(
        mean_estimation(sampleNP, w, "vote_pir", N = 50000),
        sqrt(v_r)
        )
> ic_r
lower.vote_pir upper.vote_pir
     0.3546549      0.4264249
> ic_nr <- confidence_interval(
        mean_estimation(sampleNP, w, "vote_pir", N = 50000),
        sqrt(v_nr)
        )
> ic_nr
lower.vote_pir upper.vote_pir
     0.3555368      0.4255429
```

InfoSP: Propensity Score Adjustment

Suppose that, in addition to the non-probability sample, a probability sample of the same target population is available as auxiliary information. The target variable is not measured, but some other variables which are also available in the non-probability sample have been measured on it. For the matter, `sampleP` will be used as data from the probability sample.

```
> head(sampleP)
  education_primaria education_secundaria education_terciaria age sex
1                  1                    0                   0  35   1
2                  0                    0                   1  64   0
3                  1                    0                   0  55   1
4                  0                    1                   0  61   1
5                  0                    0                   1  35   0
6                  1                    0                   0  51   1
```

In order to reduce the selection bias, Propensity Score Adjustment can be used in this case for reweighting. This procedure is implemented in the `propensities` function; it requires both samples, the list of covariates to be used to build the models for propensity estimation, and three arguments regarding technical aspects of the adjustment: the prediction algorithm (must match any of the list of `caret` supported algorithms), a boolean indicating whether smoothing of propensities is applied or not, and a vector of strings specifying the preprocessing procedures to be passed to `train` (by default, preprocessing is not applied). Further arguments to be passed to `train` can be specified.

In this example, the propensity of participating will be estimated using k-Nearest Neighbors with further smoothing and a parameter grid of all the odd numbers between 3 and 11 for optimization of k. The covariates will be all the variables measured in `sampleP`. The result will be a list with two vectors: the estimated propensities for individuals in the non-probability (convenience) and the probability (reference) sample respectively.

```
> covariates <- colnames(sampleP)
> pi <- propensities(sampleNP, sampleP, covariates,
        algorithm = "knn", smooth = T,
        tuneGrid = data.frame(k = seq(3, 11, by = 2)))
> summary(pi$convenience)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3079  0.6249  0.6873  0.6834  0.7584  0.9995
> summary(pi$reference)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3079  0.5384  0.6388  0.6236  0.6998  0.9469
```

The propensities must be subsequently transformed into weights for their application in survey estimation. Transformations available in **NonProbEst** include approaches developed by Lee, Sunghee (2006) and Lee, Sunghee and Valliant, Richard (2009) in the `lee_weights` function, Valliant, Richard and Dever, Jill A. (2011) in the `vd_weights` function, Schonlau, Matthias and Couper, Mick P. (2017) in the `sc_weights` function and Valliant, Richard (2020) in the `valliant_weights` function. `lee_weights` and `vd_weights` require propensities of both samples and a number of strata (5 by default), while `sc_weights` and `valliant_weights` only require propensities of the non-probability sample.

For example, if we want to apply propensities via weights developed in Valliant, Richard and Dever, Jill A. (2011) for the estimation of voting intention to party "pir", we can do it with the following code:

```
> wi <- vd_weights(convenience_propensities = pi$convenience,
                   reference_propensities = pi$reference)
> summary(wi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.233   1.376   1.493   1.505   1.632   2.011
> mean_estimation(sample = sampleNP, weights = wi,
                  estimated_vars = "vote_pir")
 vote_pir
0.4006072
#Estimation of the 95% confidence interval
> estim <- mean_estimation(sample = sampleNP, weights = wi,
                  estimated_vars = "vote_pir")
> std_dev <- fast_jackknife_variance(sample = sampleNP, weights = wi,
                          estimated_vars = "vote_pir", N = 50000)
> confidence_interval(estimation = estim, std_dev = std_dev, confidence = 0.95)
lower.vote_pir upper.vote_pir
     0.4001341      0.4010803
```

Note that for those weights that are calculated by means of propensity stratification, propensities of the individuals in the convenience and reference sample are needed. If they are calculated by inverting propensities, only those for the individuals in the convenience sample are needed. For example, if we calculate weights via the formula developed in Schonlau, Matthias and Couper, Mick P. (2017), the code is:

```
> wi <- sc_weights(propensities = pi$convenience)
> summary(wi)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0004998 0.3185741 0.4549419 0.5044062 0.6003197 2.2479720
```

Apart from direct estimation, resulting weights can be used as inputs in the `initial_weights`

argument of the `calib_weights` function for the estimation with PSA and calibration, or with the package **survey** (Thomas Lumley, 2018) for more complex analysis.

InfoUP: superpopulation estimators

In this case, in addition to the non-probability sample, the population itself is avaliable for some covariates. However, the target variable is only measured in the non-probability sample. For the matter, `sampleNP` will be used as the non-probability sample data and `population` will be used as the population data.

The model-based estimator can be used to estimate the population total (or mean) for the target variable. In this example, the expected number of votes for "pens" will be estimated with regularized logistic regression as learning algorithm. This procedure is implemented in the `model_based` function. It requires the sample, the population, the covariates names and the target variable as arguments. In our example, the specific algorithm and a normalization preprocessing are passed to change default behaviour. Since no optimization strategy is specified in this case, a default bootstrap will be applied.

```
> covariates <- c("education_primaria", "education_secundaria",
        "education_terciaria", "age", "sex", "language")
> mySample = sampleNP
> mySample$vote_pens = factor(mySample$vote_pens, c(0, 1), c('F', 'T'))
> model_based(mySample, population, covariates, "vote_pens",
        positive_label = 'T', algorithm = "glmnet",
        proc = c("center", "scale"))
[1] 18282.51
```

If the proportion of votes has to be estimated, rather than the total, it would be as simple as adding the `estimate_mean` argument as follows:

```
> model_based(mySample, population, covariates,
        "vote_pens", positive_label = 'T', algorithm = "glmnet",
        proc = c("center", "scale"), estimate_mean = TRUE)
[1] 0.366757
```

Alternatively, model-calibrated estimator can be used to achieve higher efficiency in some situations. In that case, design weights have to be specified in the argument "weights", in addition to the rest of arguments previously described. If no sampling design was followed in data collection, which is the case that we suppose in our example, we can specify unitary weights by turning the parameter to 1, as it is done in the following code:

```
> model_calibrated(sample_data = mySample, weights = 1, full_data = population,
+                  covariates = covariates, estimated_var = "vote_pens",
+                  positive_label = 'T', algorithm = "glmnet",
+                  proc = c("center","scale"), estimate_mean = TRUE)
[1] 0.365945
```

Conclusion and future developments

In this paper we show how the **NonProbEst** package can simplify the application of different weighting methods to correct selection bias in non-probability surveys. This package is, to the best of our knowledge, the first package that supports the user beyond estimation in PSA, PSA+calibration, statistical matching or model-calibration. Another important feature is that a wide range of ML techniques can be used to optimize the information provided by the auxiliary variables.

Additional features will be integrated in future versions of the package. Some simplified wrappers will be developed for some methods so non-expert users can also easily apply them, more parameters will be avaliable for estimation and further support for weighted models will be added. Also, other techniques for variance estimation can be considered. Many of these features can already be applied combining **NonProbEst** with the **survey** package, as noted before.

Regarding Machine Learning, methods for variable selection will be studied as well as the use of more advanced deep learning libraries outside of **caret**'s scope. Variable selection would help explaining the bias and choosing the best covariates for its correction. Better deep learning libraries would allow the use of state-of-the-art algorithms.

Acknowledgments

This work is partially supported by Ministerio de Economía y Competitividad of Spain (grant MTM2015-63609-R) and by Ministerio de Ciencia, Innovación y Universidades (grant FPU17/02177).

Bibliography

Baffetta, Federica, Fattorini, Lorenzo, Franceschi, Sara, and Corona, Piermaria. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, 113(3):463–475, 2009. URL https://doi.org/10.1016/j.rse.2008.06.014. [p6]

Bethlehem, Jelke. Selection bias in web surveys. *International Statistical Review*, 78(2):161–188, 2010. URL https://doi.org/10.1111/j.1751-5823.2010.00112.x. [p3]

Breidt, F Jay, Opsomer, Jean D, and others. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205, 2017. URL https://doi.org/10.1214/16-sts589. [p6]

Buelens, Bart, Burger, Joep, and van den Brakel, Jan A. Comparing inference methods for non-probability samples. *International Statistical Review*, 86(2):322–343, 2018. URL https://doi.org/10.1111/insr.12253. [p1, 5, 6]

Buskirk, Trent D and Kolenikov, Stanislav. Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, page 17, 2015. [p6]

Chen, Jack Kuang Tsung, Valliant, Richard L., and Elliott, Michael R. Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):657–681, 2019. URL https://doi.org/10.1111/rssc.12327. [p6]

Chen, Tianqi and Guestrin, Carlos. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016. [p5]

Cochran, William G. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968. URL https://doi.org/10.2307/2528036. [p3]

Deville, Jean-Claude and Särndal, Carl Erik. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992. URL https://doi.org/10.1080/01621459.1992.10475217. [p2]

Elliott, Michael R. and Valliant, Richard. Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017. URL https://doi.org/10.1214/16-sts598. [p1]

Ferri-García, Ramón and Rueda, Maria del Mar. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Statistics and Operations Research Transactions*, 1(2):159–182, 2018. [p4, 6]

Hoerl, Arthur E and Kennard, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. URL https://doi.org/10.1080/00401706.1970.10488634. [p5]

Kim, J.K and Fuller, W. Fractional hot deck imputation. *Biometrika*, 91(3):559–578, 2004. URL https://doi.org/10.1093/biomet/91.3.559. [p4]

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey. Deep learning. *nature*, 521(7553):436–444, 2015. URL https://doi.org/10.1038/nature14539. [p6]

Lee, Sunghee. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2):329–349, 2006. [p6, 10]

Lee, Sunghee and Valliant, Richard. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343, 2009. URL https://doi.org/10.1177/0049124108329643. [p1, 3, 4, 10]

Max Kuhn. *caret: Classification and Regression Training*, 2018. URL https://CRAN.R-project.org/package=caret. R package version 6.0-81. [p6]

Montanari, Giorgio E and Ranalli, M Giovanna. Multiple and ridge model calibration for sample surveys. In *Proceedings of the Workshop in Calibration and estimation in surveys, Ottawa*, 2007. [p6]

Park, Trevor and Casella, George. The bayesian lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008. URL https://doi.org/10.1198/016214508000000337. [p6]

Phipps, Polly, Toth, Daniell, and others. Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6(2):772–794, 2012. URL https://doi.org/10.1214/11-AOAS521. [p6]

Quenouille, Maurice H. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956. URL https://doi.org/10.2307/2332914. [p4]

Quinlan, J Ross. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993. [p5]

Rivers, D. Sampling for web surveys. *Presented in Joint Statistical Meetings*, 2007. Salt Lake City, UT. [p4]

Rosenbaum, Paul R and Rubin, Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. URL https://doi.org/10.1093/biomet/70.1.41. [p3]

Rubin, Donald B. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94, 1986. URL https://doi.org/10.2307/1391390. [p3]

Schonlau, Matthias and Couper, Mick P. Options for conducting web surveys. *Statistical Science*, 32(2): 279–292, 2017. URL https://doi.org/10.1214/16-sts597. [p3, 10]

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W., and Terhanian, G. The record of internet-based opinion polls in predicting the results of 72 races in the november 2000 us elections. *International Journal of Market Research*, 43(2):127–135, 2001. URL https://doi.org/10.1177/147078530104300203. [p3]

Thomas Lumley. *survey: Analysis of Complex Survey Samples*, 2018. URL https://CRAN.R-project.org/package=survey. [p11]

Valliant, Richard. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), 2020. URL https://doi.org/10.1093/jssam/smz003. [p3, 4, 10]

Valliant, Richard and Dever, Jill A. Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137, 2011. URL https://doi.org/10.1177/0049124110392533. [p3, 4, 6, 10]

Yves Tillé and Alina Matei. *sampling: Survey Sampling*, 2016. URL https://CRAN.R-project.org/package=sampling. R package version 2.8. [p6]

María del Mar Rueda
Department of Statistics and Operations Research
University of Granada
Spain
ORCiD: 0000-0002-2903-8745
mrueda@ugr.es

Ramón Ferri-García
Department of Statistics and Operations Research
University of Granada
Spain
ORCiD: 0000-0002-9655-933X
rferri@ugr.es

Luis Castro
Department of Statistics and Operations Research
University of Granada
Spain
luiscastro193@correo.ugr.es

Appendix A3

Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

Castro-Martín, Luis; Rueda, María del Mar; Ferri-García, Ramón (2022)
Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys
Journal of Computational and Applied Mathematics, vol. 404, p. 113414
DOI: 10.1016/j.cam.2021.113414

MATHEMATICS, APPLIED			
JCR Year	Impact factor	Rank	Quartile
2020	2.621	36/265	Q1

Journal of Computational and Applied Mathematics 404 (2022) 113414

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys

Luis Castro-Martín, María del Mar Rueda *, Ramón Ferri-García

Department of Statistics and Operations Research, University of Granada, Spain

ARTICLE INFO

Article history:
Received 24 September 2020

Keywords:
Nonprobability surveys
Machine learning techniques
Propensity score adjustment
Survey sampling

ABSTRACT

The convenience of online surveys has quickly increased their popularity for data collection. However, this method is often non-probabilistic as they usually rely on selfselection procedures and internet coverage. These problems produce biased samples. In order to mitigate this bias, some methods like Statistical Matching and Propensity Score Adjustment (PSA) have been proposed. Both of them use a probabilistic reference sample with some covariates in common with the convenience sample. Statistical Matching trains a machine learning model with the convenience sample which is then used to predict the target variable for the reference sample. These predicted values can be used to estimate population values. In PSA, both samples are used to train a model which estimates the propensity to participate in the convenience sample. Weights for the convenience sample are then calculated with those propensities. In this study, we propose methods to combine both techniques. The performance of each proposed method is tested by drawing nonprobability and probability samples from real datasets and using them to estimate population parameters.

1. Introduction

Survey samplers have long been using probability samples from one or more sources to make valid and efficient inferences on finite population parameters. Methods for combining two or more probability samples were also developed to increase the efficiency of estimators for a given cost. Dual frame and multiple frame methods for survey estimation, developed in [1] and [2] respectively, are an example of such techniques.

Due to technological innovations, large amounts of inexpensive data (commonly known as Big Data) and data from non-probability samples are now accessible. Big data include administrative data, social media data, internet of things and scraped data from websites, and satellite images. Big Data and data from web panels have the potential of providing estimates in near real time, unlike traditional data derived from probability samples. Statistical agencies are now taking modernization initiatives into account to find new ways to integrate data from a variety of sources and to produce "real-time" official statistics. On the other hand, a review by [3] concludes that the potential of probability sampling cannot be reached by nonprobability samples, even if correction methods are applied.

Inferences from Big Data and nonprobability surveys have important sources of error. Given the characteristics of these data collection procedures, selection bias is particularly relevant. Following notation from [4], in a situation where U is the target population to which survey results are supposed to be generalized, a nonprobability selection ensures that sample individuals will be drawn from a population of potentially covered individuals, $U_{pc} \subset U$. This is the case

* Corresponding author.
E-mail addresses: luiscastro193@ugr.es (L. Castro-Martín), mrueda@ugr.es (M.d.M. Rueda), rferri@ugr.es (R. Ferri-García).

https://doi.org/10.1016/j.cam.2021.113414
0377-0427/

of internet and smartphone surveys, where the population with the necessary devices for taking part in the survey are a subset of the total population. The bias produced by this issue is commonly known as coverage error. In addition, if the participation in the survey is conditioned to a selection mechanism, the sample will be eventually drawn from an actually covered population, $U_{ac} \subset U_{pc}$. Following the previous example, internet surveys with an opt-in scheme (such as snowball samples in social media websites) would recruit volunteer respondents willing to participate, hence not all of the potentially covered population would have a non-zero probability of being drawn. This is commonly known as self-selection bias.

Some techniques to mitigate selection bias can be applied if a probability sample, drawn from U with a sampling design (d_s, p_s) and negligible sources of bias, is available. From all of them, Propensity Score Adjustment (PSA) and Statistical Matching have gained interest from the research community. PSA, originally developed for reducing selection bias in non-randomized clinical trials [5], was adapted to nonprobability surveys in the works of [6] and [7]. This method aims to estimate the propensity to participate in the survey of each individual by taking into account how would have the sample been if it was drawn with a probability sampling design. Its efficacy at reducing selection bias has been repeatedly proven [6–10], although requires a proper specification of the model and the variables to be included on it, and further adjustments such as calibration. Statistical Matching [11,12] is a rather predictive approach; the nonprobability sample is used to develop a prediction model on the target variable, which is subsequently used for prediction in the probability sample. It remains unclear which of the methods is more efficient, although a recent experiment by [13] showed better results for Statistical Matching in terms of efficiency.

In this study, we treat the problem of integrating the information provided by probability and nonprobability surveys (or Big Data). We develop a set of procedures which combine the results provided by PSA and Statistical Matching to obtain survey estimates, and compare their efficiency to that of the mentioned methods on their own. The combination of results from multiple sources has been studied in survey research, and the promising results provide some evidence that the application of these methods could be fruitful in the nonprobability survey context. Furthermore, predictive modeling allows to incorporate auxiliary information as training weights or parameter configuration, hence a two-step approach can be applied. Our initial hypothesis is that the combination of multiple sources for estimation in nonprobability survey sampling has the potential to overcome current methods in terms of bias reduction and efficiency of the estimators.

The remainder of the article is organized in four sections. After introducing the problem of estimation in Section 2, in Section 3, new estimators are proposed based on different approach to integrate data. Some simulation experiments are carried out to check the finite size sample properties of the proposed estimators in Section Section 4. Finally, Section Section 5 presents the concluding remarks.

2. The problem of estimation with non-probability samples

Let U denote a finite population with N units, $U = \{1, \ldots, k, \ldots, N\}$. Let s_V be a volunteer non-probability sample of size n_V, self-selected from an online population U_V which is a subset of the total target population U and s_R a reference probabilistic sample of size n_{rs} selected from U under a sampling design (s_d, p_d) with $\pi_i = \sum_{s_r \ni i} p_d(s_r)$ the first order inclusion probability for the ith individual. Let y be the variable of interest in the survey estimation. Let $\mathbf{x}_k$ be the value taken on unit k by a vector of auxiliary variables. Covariates $\mathbf{x}$ have been measured on both samples, while the variable of interest y has been measured only in the volunteer sample. We denote by $w_{Rk} = 1/\pi_k$ the original design weight of the k individual in the reference sample.

A matching estimator is defined by:

$$\hat{Y}_{SM} = \sum_{s_R} \hat{y}_k w_{Rk}$$

being $\hat{y}_k$ the predicted value of y_k.

The key is how to predict the values y_k. Formal working linear regression models, relating the study variable y to the vector of auxiliary variables are usually considered to develop efficient estimators of the total Y. Suppose a working population model, $E_m(y_i) = m(x_i, \beta) = m_i$ for $i \in U$ is assumed to hold for the sample s_V where E_m denotes model expectation and the mean function m_i is specified. Using the data from the sample s_V we obtain an estimator $\hat{\beta}$ which is consistent for β if the model is correctly specified and thus the estimator $\hat{Y}_{SM}$ is consistent if the model for the study variable is correctly specified but the estimator will be biased if the model for the study variable is incorrectly specified. Parametric models require assumptions regarding variable selection, the functional form and distributions of variables, and specification of interactions. Contrary to statistical modeling approaches that assume a data model with parameters estimated from the data, more advanced machine learning algorithms aim to extract the relationship between an outcome and predictor without an a priori data model. These methods have been recently applied in the statistical matching context in [13].

In recent years, propensity score adjustment (PSA) has increasingly been used as a means of correcting selection bias in online surveys. The efficacy of PSA at removing selection bias from online surveys has been discussed in numerous studies (see e.g. [6]; [7]; [8]; [9]).

It is expected that a sample collected by online recruitment would not follow the principles of a probability sampling, especially in those cases that the survey is filled by volunteer respondents. We can define an indicator variable I as follows:

$$I_i = \begin{matrix} 1 & i \in s_V \\ 0 & i \notin s_V \end{matrix}, i = 1, 2, \ldots, N \tag{1}$$

Propensity scores, π_i, can be defined as the propensity of the ith individual of participating in the survey, this is, the probability that $I_i = 1$. The propensity score of the individual can be formulated, following notation in [14], as the expected value of I conditional on her/his target variable and covariates' value:

$$\pi_i = E[I_i|\mathbf{x}_i, y_i] = P(I_i = 1|\mathbf{x}_i, y_i) \tag{2}$$

The probability reflects the selection mechanism of the non-probability sample. Depending on the mechanism, the conditional probability might vary. If the selection is Missing Completely At Random (MCAR), then $P(I_i = 1|\mathbf{x}_i, y_i) = P(I_i = 1)$ and estimates obtained from s_V would be unbiased. If the selection is Missing At Random (MAR), then $P(I_i = 1|\mathbf{x}_i, y_i) = P(I_i = 1|\mathbf{x}_i)$. When the selection mechanism is Missing Not At Random (MNAR) or MAR, Propensity Score Adjustment (PSA) can be applied to remove the bias induced by such mechanisms. Although the real propensity cannot be obtained, it can be estimated if a reference survey is available. The reference survey must have been conducted on the same target population than the online survey but collected in a more adequate manner regarding coverage and response issues.

The propensity for an individual to take part on the non-probability survey is obtained by training a predictive model (often a logistic regression) on the dichotomous variable, I_{s_V}, which measures whether a respondent from the combination of both samples took part in the volunteer survey or in the reference survey. Covariates used in the model, $\mathbf{x}$, are measured in both samples (in contrast to the target variable which is only measured in the non-probability sample), thus the formula to compute the propensity of taking part in the volunteer survey with a logistic model, π, can be displayed as

$$\pi(\mathbf{x}) = \frac{1}{e^{-(\gamma^T\mathbf{x})} + 1} \tag{3}$$

for some vector γ, as a function of the model covariates.

We can use the inverse of the estimated response propensity as a weight for constructing the estimator [15]:

$$\hat{Y}_{PSA} = \sum_{k \in s_V} w_{V\,k} y_k / \hat{\pi}(\mathbf{x}_k) = \sum_{k \in s_V} y_k w_k^{PSA} \tag{4}$$

where $\hat{\pi}(\mathbf{x}_k)$ is the estimated response propensity for the individual k of the volunteer sample as predicted using covariates $\mathbf{x}$.

3. Proposed estimators by combining probability and non-probability samples

In this section, we will explore new ways of doing the integration of data of probability and non-probability samples.

3.1. Shrinkage

Shrinkage is a natural way to improve the available estimates, in terms of the mean squared error. For example, composite estimators are used in small area estimation (see [16,17]). [18] applies shrinkage in regression analysis and [19] uses this technique to predict a binary response on the basis of binary explanatory variables. Similarly, [20] propose a shrinkage calibration estimator in cluster sampling.

We propose an estimator based on composite information, as follows:

$\hat{Y}_{srk} = K\hat{Y}_{SM} + (1 - K)\hat{Y}_{PSA}$, where K is a constant satisfying $0 < K < 1$.

Theorem 1. *The optimum value for k in the sense of minimum variance into the class of estimators $\hat{Y}_{srk}$ is*

$$k_{opt} = \frac{AV(\hat{Y}_{PSA}) - cov(\hat{Y}_{SM}, \hat{Y}_{PSA})}{AV(\hat{Y}_{SM}) + AV(\hat{Y}_{PSA}) - 2cov(\hat{Y}_{SM}, \hat{Y}_{PSA})}. \tag{5}$$

Proof. The variance of $\hat{Y}_{srk}$ is given by

$$\begin{aligned} V(\hat{Y}_{srk}) &= V(K\hat{Y}_{SM} + (1 - K)\hat{Y}_{PSA}) \\ &= K^2 V(\hat{Y}_{SM}) + (1 - K)^2 V(\hat{Y}_{PSA}) + 2K(1 - K)cov(\hat{Y}_{SM}, \widehat{p}_{rq}). \end{aligned}$$

By denoting $V_1 = V(\hat{Y}_{SM})$, $V_2 = V(\hat{Y}_{PSA})$ and $C = cov(\hat{Y}_{SM}, \hat{Y}_{PSA})$, the variance of $\hat{Y}_{srk}$ can be expressed as

$$V(\hat{Y}_{srk}) = K^2 V_1 + (1 - K)^2 V_2 + 2K(1 - K)C.$$

The first derivative of $V(\hat{Y}_{srk})$ with respect to K is

$$\frac{\partial V(\hat{Y}_{srk})}{\partial k} = 2KV_1 - 2(1-K)V_2 + 2(1-2K)C = 0;$$
$$K_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C}.$$

The second derivative is

$$\frac{\partial V(\hat{Y}_{srk})}{\partial^2 K} = 2V(\hat{Y}_{SM} - \hat{Y}_{PSA}) > 0,$$

and we conclude that K_{opt} really minimizes $AV(\hat{Y}_{srk})$. □

Note. Usually samples s_V and s_p are independents, thus $K_{opt} = \frac{V_2}{V_1+V_2}$.

The optimal coefficient K_{opt} depends on population variances, which are usually unknown in practice, and so $\hat{Y}_{srk_{opt}}$ cannot be calculated.

The following estimator can be defined

$$\hat{Y}_{op} = \widehat{K}_{opt}\hat{Y}_{SM} + (1 - \widehat{K}_{opt})\hat{Y}_{PSA}$$

where $\widehat{K}_{opt}$ denotes that estimates are substituted for the variances and covariances in (5).

3.2. Double robust estimator

We assume a working population model, $E_m(y_i) = \mu(\mathbf{x}_i) = m_i, i = 1, \ldots, N$. A new estimator which combines probability and non-probability samples can be defined by using the idea of the difference estimator ([21], pag. 222).

The total Y can be written as:

$$Y = \sum_U \hat{y}_k + \sum_U (y_k - \hat{y}_k)$$

being $\hat{y}_k = \hat{m}_k$ the predicted value of the y_k under the population model. We estimate each term by using the weighted estimator obtained from the reference probabilistic sample and the volunteer sample respectively:

$$\hat{Y}_{DR} = \sum_{s_R} \hat{y}_k w_{Rk} + \sum_{s_V} w_k^{PSA}(y_k - \hat{y}_k).$$

The estimator $\hat{Y}_{DR}$ is double robust: it is consistent if either the model for the propensities or the model for the study variable is correctly specified.

If the working outcome regression model for y is linear, $E_m(y_i) = \beta\mathbf{x}$, this estimator coincides with the estimator proposed by [14].

3.3. Training data with PSA weights

Most machine learning models allow considering weights for the training data. We also propose an estimator which uses w_k^{PSA} for $k \in s_V$ when training the model which predicts $\hat{y}_k$ for $k \in s_R$. The estimation would then be: $\sum_{s_R} \hat{y}_k w_{Rk}$

For example, if the chosen model is linear regression, a predictor for Statistical Matching would be obtained as

$$E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta$$

where β coefficients are optimized in order to minimize the following Mean Square Error:

$$MSE(s_V) = \frac{\sum_{s_V}(\hat{y}_k - y_k)^2}{n_V}$$

The proposed estimator would simply minimize the following weighted Mean Square Error instead:

$$MSE(s_V) = \frac{\sum_{s_V} w_k^{PSA}(\hat{y}_k - y_k)^2}{\sum_{s_V} w_k^{PSA}}.$$

Thus the proposed estimator will be obtained with algorithm 1:

- Calculate w_k^{PSA} for $k \in s_V$ by using some machine learning classification algorithm described in Ferri and Rueda (2020).
- Train a model $E_m(y_i|\mathbf{x}_i)$ using x_k for $k \in s_V$ weighted with w_k^{PSA} for $k \in s_V$. Often, this means minimizing the weighted Mean Square Error defined above. However, each machine learning model may have its own weighting mechanism.
- Obtain $\hat{y}_k$ for $k \in s_R$ using the model trained in the previous step.
- Estimate the total as $\hat{Y}_{tr} = \sum_{s_R} \hat{y}_k w_{Rk}$

4. Simulation study

4.1. Data

We have chosen 3 datasets for the simulation study. Also, for each one of them, 2 different non-probabilistic sampling strategies are used for the volunteer sample. The probabilistic sampling strategy for the reference sample is always a simple random sampling among the whole population. The volunteer samples include the target variable while the reference samples do not contain that information.

The first population is the Hotel Booking Demand Dataset [22], denoted as P1. It contains booking information for a city hotel and a resort hotel. In total, it consists of 119,390 bookings due to arrival between the 1st of July of 2015 and the 31st of August 2017. The target is estimating the mean number of week nights (Monday to Friday) the guests book to stay at the hotel. The first non-probabilistic sampling strategy, denoted as S1, is a random sampling where the bookings from the resort hotel have 10 times more probability of being chosen than the bookings from the city hotel. The second sampling strategy, denoted as S2, is a random sampling where the bookings from the city hotel have 5 times more probability of being chosen than the bookings from the resort hotel. In both cases, 28 covariates were used. The only variables excluded as covariates were the target, the hotel type, the reservation status and the reservation status date.

The second population is BigLucy [23], denoted as P2. It contains financial information about 85,396 industrial companies. In this case, the target is estimating the mean annual income in the previous year. The first non-probabilistic sampling strategy, denoted as S1, is a simple random sampling among the companies with SPAM options, excluding those labeled as "small companies". The second sampling strategy, denoted as S2, considers a propensity to participate in the volunteer sample calculated as $Pr(taxes) = min(taxes^2/30, 1)$, where *taxes* is the company's income tax in the previous year, among the companies with SPAM options. The covariates used are: the number of employees, the company's income tax, the size (small, medium or big) and whether it is ISO certified.

The third population, denoted as P3, consists of a study conducted in 2012 by the Spanish National Institute of Statistics about the economic and life conditions of 28,610 adult individuals [24]. The target is estimating the mean self-reported health on a scale from 1 to 5. For the first sampling strategy, denoted as S1, a simple random sample is taken among the individuals with internet access. For the second one, denoted as S2, a propensity to participate defined as $Pr(yr) = \frac{yr^2 - 1900^2}{1996^2 - 1900^2}$, where *yr* is the year the individual was born, is added to the internet restriction. 56 health-related covariates are used, avoiding those too correlated with the target variable like health issues in the last 6 months or chronic conditions.

4.2. Simulation

We have performed simulations for the 4 proposed estimators, including both variants of shrinkage. For each one, every dataset with their corresponding sampling strategies has been simulated 500 times for each sample size. 1000, 2000 and 5000 have been used as sample size, taking the same size for both samples (the volunteer and the reference ones). The machine learning model chosen for every method is logistic regression, given its proven reliability [13].

In order to evaluate the results for the simulations, 3 metrics are calculated: the relative mean bias, the relative standard deviation and the relative Root Mean Square Error. These metrics are defined as follows:

$$RBias\ (\%) = \left| \frac{\sum_{i=1}^{500} \hat{Y}^{(i)}}{500} - Y \right| \cdot \frac{100}{Y} \tag{6}$$

$$RStandard\ deviation\ (\%) = \sqrt{\frac{\sum_{i=1}^{500} (\hat{Y}^{(i)} - \hat{\bar{Y}})^2}{499}} \cdot \frac{100}{Y} \tag{7}$$

$$RMSE\ (\%) = \sqrt{RBias^2 + RSD^2} \tag{8}$$

with $\hat{Y}^{(i)}$ the estimation of Y in the ith simulation and $\hat{\bar{Y}}$ the mean of the 500 estimations.

Finally, in order to compare each method, the mean and median efficiency is obtained as well as the number of times it has been among the best. The efficiency of a method is defined as follows:

$$Efficiency\ (\%) = \frac{Baseline - RMSE}{Baseline} \cdot 100 \tag{9}$$

where the baseline is the RMSE of using the unweighted sample mean for the estimation. Also, a method is considered to be among the best when its RMSE differs from the best RMSE by less than 1%.

Table 1
Relative mean bias (%) of each population and sample size for each method.

	Baseline	Matching	PSA	Training	Chen	K_1	K_2
P1S1 1000	18.9	4.5	5.5	4.5	4.6	5.2	5
P1S1 2000	18.9	4.9	5.5	4.8	4.9	5.1	5.2
P1S1 5000	18.6	4.8	4.6	4.7	4.8	4.9	4.7
P1S2 1000	9.2	5	4.1	4.1	4.1	4.5	5
P1S2 2000	9.2	4.9	4.2	3.9	4.1	4.4	4.4
P1S2 5000	9.1	4.7	3.9	3.6	3.8	4.3	4.3
P2S1 1000	70.6	24.4	67.7	23.6	24.4	46	35.2
P2S1 2000	70.4	24.6	68	23.7	24.5	46.2	35.4
P2S1 5000	70.4	24.7	68.1	23.7	24.5	46.3	35.3
P2S2 1000	32.7	12.6	15.1	10.9	11.9	13.7	13.7
P2S2 2000	32.6	12.7	15.1	10.9	11.9	13.6	13.7
P2S2 5000	32.9	12.7	15.1	11	12	13.7	13.8
P3S1 1000	8.4	2.6	3.4	2.3	2.3	2.9	2.9
P3S1 2000	8.5	2.4	3.5	2.2	2.3	3	3
P3S1 5000	8.5	2.5	3.5	2.1	2.3	3	3
P3S2 1000	12.9	4.7	5.6	4.1	4.5	5.2	5.2
P3S2 2000	12.8	4.7	5.8	4.1	4.3	5.2	5.2
P3S2 5000	12.8	4.6	5.8	4	4.2	5.1	5.1

Table 2
Relative RMSE (%) of each population and sample size for each method.

	Baseline	Matching	PSA	Training	Chen	K_1	K_2
P1S1 1000	19.1	5.6	6.3	5.4	5.5	6	5.8
P1S1 2000	18.9	5.4	5.9	5.3	5.3	5.5	5.6
P1S1 5000	18.7	5	8.6	4.9	5.6	5.9	6.3
P1S2 1000	9.5	5.9	5.7	5	5.3	5.5	5.9
P1S2 2000	9.3	5.3	4.8	4.4	4.7	4.9	4.9
P1S2 5000	9.2	4.8	4.2	3.8	4	4.5	4.4
P2S1 1000	70.6	24.4	67.8	23.7	24.4	46	35.3
P2S1 2000	70.4	24.6	68.1	23.7	24.5	46.2	35.4
P2S1 5000	70.5	24.7	68.1	23.7	24.5	46.3	35.4
P2S2 1000	32.8	12.7	15.2	11.1	12	13.8	13.8
P2S2 2000	32.7	12.7	15.1	11	12	13.7	13.8
P2S2 5000	32.9	12.7	15.2	11	12	13.7	13.8
P3S1 1000	8.5	3	3.7	2.8	2.7	3.3	3.3
P3S1 2000	8.5	2.6	3.6	2.4	2.5	3.1	3.1
P3S1 5000	8.5	2.5	3.5	2.2	2.3	3.1	3.1
P3S2 1000	12.9	5.1	6	4.6	4.9	5.6	5.6
P3S2 2000	12.9	4.9	6	4.4	4.6	5.3	5.3
P3S2 5000	12.8	4.7	5.9	4.1	4.3	5.2	5.2

4.3. Results

The results obtained for the bias and RMSE can be consulted in Tables 1 and 2 respectively. Table 3 contains the summary comparing each method. Both shrinkage estimators are referred to as K_1, for $K_1 = s_r/(s_r + s_v)$, and K_2, for $K_2 = V(\hat{\theta}_{PSA})/(V(\hat{\theta}_{PSA}) + V(\hat{\theta}_{SM}))$. The double robust estimator is referred to as *Chen*. The estimator which uses PSA weights when training the Statistical Matching model is referred to as *Training*.

As it can be observed, Training always obtains the best estimations. Even though its difference from Matching is small, the most interesting point is that even in the case where PSA outperforms Matching, Training is still better. Chen offers very similar results, although slightly worse.

Shrinkage simply produces values between Matching and PSA. Also, there is not much difference between both variants because the variance of Matching and PSA is usually similar.

5. Conclusions

Selection bias, a growing issue in survey sampling and empirical sciences due to new questionnaire administration methods, appears when a sample is drawn from a potentially covered population which is different on its composition to the target population. If a sample drawn from the target population is available, some methods can be applied to adjust for selection bias in the nonprobability sample. Propensity Score Adjustment (PSA) and Statistical Matching are the

Table 3
Mean and median efficiency (%) of each method and times it has been among the best.

	Mean	Median	Best
Training	65.8	66.4	18
Chen	64	65.2	18
Matching	61.8	64.2	14
K_2	57.3	58	10
K_1	55	58.2	10
PSA	46.6	53.9	6

most important methods up to date, both of them showing an increase in efficiency when applied to the estimation of a population parameter. In this context, it is feasible that a combination of both methods could result in an advantage in terms of bias and error reduction, especially given that they can be complemented as they have different outcomes (weights in PSA and predictions in Matching). Previous work by [14] proved that a doubly-robust estimator could provide acceptable results, with good properties.

In this study, shrinkage methods to combine two estimates, doubly-robust estimation and the use of PSA weights in the training of models to be used for Statistical Matching are compared in terms of bias and RMSE. The results are obtained from simulations with three different datasets to enable the study of the behavior of such methods under different conditions. Results show a certain advantage of the training method developed in this paper over the model-assisted estimator, and an advantage of both of them over Statistical Matching. Shrinkage and PSA stand far below, although they offer competitive results under certain circumstances.

The advantage of the training method is that it gives more importance in the prediction to those individuals who are more likely to appear in the population. By default, a model trained in a biased dataset might also produce biased predictions; however, if this bias is corrected by methods such as PSA, it is expected that the relationships established by the prediction model and its results are more similar to those present in the target population. This also applies to the model-assisted estimator, where the prediction errors in the nonprobability sample with the largest importance are those with a higher probability of being present in a random sample from the target population.

Our study has some limitations to be noted: first, although a variety of datasets have been used, the suitability of each method might be influenced by the data itself. The results presented here need further replicability in a wider range of datasets and scenarios in order to have the full picture. Secondly, only one prediction algorithm (linear regression models) was used in the study. Previous research showed that modern Machine Learning prediction techniques can be advantageous in removing selection bias with PSA [25], although it remains unclear for Statistical Matching [13]. Further research could introduce these algorithms in the adjustment methods presented here and compare them to the linear regression case. Finally, the theoretical properties of some of the methods proposed here (shrinkage and training) have to be developed, although these properties should not be very different from those of the dual frame estimation (in the case of shrinkage) or those from the Statistical Matching estimator (in the case of training).

Acknowledgment

This work is partially supported by Ministerio de Economía y Competitividad of Spain grant MTM2015-63609-R and PID2019-106861RB-I00 /AEI 10.13039/501100011033.

References

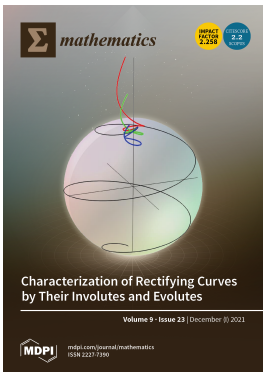
[1] H.O. Hartley, Multiple frame surveys, in: Proceedings of the Social Statistics Section, American Statistical Association, Vol. 19, Washington, DC, 1962, pp. 203–206.
[2] F. Mecatti, A single frame multiplicity estimator for multiple frame surveys, Surv. Methodol. 33 (2) (2007) 151–157.
[3] C. Cornesse, A.G. Blom, D. Dutwin, J.A. Krosnick, E.D. De Leeuw, S. Legleye, J. Pasek, D. Pennay, B. Phillips, J.W. Sakshaug, et al., A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research, J. Surv. Stat. Methodol. 8 (1) (2020) 4–36, http://dx.doi.org/10.1093/jssam/smz041.
[4] M.R. Elliott, R. Valliant, et al., Inference for nonprobability samples, Statist. Sci. 32 (2) (2017) 249–264, http://dx.doi.org/10.1214/16-STS598.
[5] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55, http://dx.doi.org/10.1093/biomet/70.1.41.
[6] S. Lee, Propensity score adjustment as a weighting scheme for volunteer panel web surveys, J. Off. Stat. 22 (2) (2006) 329.
[7] S. Lee, R. Valliant, Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment, Sociol. Methods Res. 37 (3) (2009) 319–343, http://dx.doi.org/10.1177/0049124108329643.
[8] R. Valliant, J.A. Dever, Estimating propensity adjustments for volunteer web surveys, Sociol. Methods Res. 40 (1) (2011) 105–137, http://dx.doi.org/10.1177/0049124110392533.
[9] R. Ferri-García, M. Rueda, Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys, SORT 1 (2018) 159–162.
[10] L. Castro-Martín, M. Rueda, R. Ferri-García, Estimating general parameters from non-probability surveys using propensity score adjustment, Mathematics 8,11 (2020) 2096.
[11] D. Rivers, Sampling for web surveys, in: Joint Statistical Meetings, 2007.
[12] J.-F. Beaumont, J. Bissonnette, Variance estimation under composite imputation: The methodology behind sevani, Surv. Methodol. 37 (2) (2011) 171–179.

[13] L. Castro-Martín, M. Rueda, R. Ferri-García, Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques, Mathematics 8 (6) (2020) 879, http://dx.doi.org/10.3390/math8060879.
[14] Y. Chen, P. Li, C. Wu, Doubly robust inference with nonprobability survey samples, J. Amer. Statist. Assoc. (2019) 1–11, http://dx.doi.org/10.1080/01621459.2019.1677241.
[15] R. Valliant, Comparing alternatives for estimation from nonprobability samples, J. Surv. Stat. Methodol. 8 (2) (2020) 231–263, http://dx.doi.org/10.1093/jssam/smz003.
[16] W. James, C. Stein, Estimation with quadratic loss, in: Breakthroughs in Statistics, Springer, 1992, pp. 443–460, http://dx.doi.org/10.1007/978-1-4612-0919-5_30.
[17] J.N. Rao, Small-area estimation, in: Wiley StatsRef: Statistics Reference Online, 2014, pp. 1–8, http://dx.doi.org/10.1002/9781118445112.stat03310.pub2.
[18] J.B. Copas, Regression, prediction and shrinkage, J. R. Stat. Soc. Ser. B Stat. Methodol. 45 (3) (1983) 311–335, http://dx.doi.org/10.1111/j.2517-6161.1983.tb01258.x.
[19] J. Copas, The shrinkage of point scoring methods, J. R. Stat. Soc. Ser. C. Appl. Stat. 42 (2) (1993) 315–331, http://dx.doi.org/10.2307/2986235.
[20] A. Arcos, J.M. Contreras, M.M. Rueda, A novel calibration estimator in social surveys, Sociol. Methods Res. 43 (3) (2014) 465–489, http://dx.doi.org/10.1177/0049124113507906.
[21] C.-E. Särndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer Science & Business Media, 2003.
[22] N. Antonio, A. de Almeida, L. Nunes, Hotel booking demand datasets, Data Brief 22 (2019) 41–49, http://dx.doi.org/10.1016/j.dib.2018.11.126.
[23] H.A. Gutiérrez, Estrategias de Muestreo Diseño de Encuestas y Estimacion de Parametros, Universidad Santo Tomas, Bogota (Colombia), 2009.
[24] National institute of statistics (INE), Life conditions survey, microdata, 2012.
[25] R. Ferri-García, M. Rueda, Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys, PLoS One 15 (4) (2020) e0231500, http://dx.doi.org/10.1371/journal.pone.0231500.

Appendix A4

On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

Castro-Martín, Luis; Rueda, María del Mar; Ferri-García, Ramón; Hernando-Tamayo, César (2021)
On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures
Mathematics, vol. 9, number 23, p. 2991
DOI: 10.3390/math9232991



MATHEMATICS			
JCR Year	Impact factor	Rank	Quartile
2020	2.258	24/330	Q1

mathematics

MDPI

Article

On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures

Luis Castro-Martín, María del Mar Rueda *, Ramón Ferri-García and César Hernando-Tamayo

Department of Statistics and Operational Research, University of Granada, 18011 Granada, Spain; luiscastro193@ugr.es (L.C.-M.); rferri@ugr.es (R.F.-G.); cesarhernando@ugr.es (C.H.-T.)
* Correspondence: mrueda@ugr.es

Abstract: In the last years, web surveys have established themselves as one of the main methods in empirical research. However, the effect of coverage and selection bias in such surveys has undercut their utility for statistical inference in finite populations. To compensate for these biases, researchers have employed a variety of statistical techniques to adjust nonprobability samples so that they more closely match the population. In this study, we test the potential of the XGBoost algorithm in the most important methods for estimation that integrate data from a probability survey and a nonprobability survey. At the same time, a comparison is made of the effectiveness of these methods for the elimination of biases. The results show that the four proposed estimators based on gradient boosting frameworks can improve survey representativity with respect to other classic prediction methods. The proposed methodology is also used to analyze a real nonprobability survey sample on the social effects of COVID-19.

Keywords: nonprobability surveys; machine learning techniques; propensity score adjustment; survey sampling

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

Academic Editors: Leonid V. Bogachev and Amir Mosavi

Received: 6 October 2021
Accepted: 19 November 2021
Published: 23 November 2021

1. Introduction

Survey sampling theory, since its foundation in the 20th century with the works of Jerzy Neyman [1,2], has been the gold standard for applied research in the empirical sciences. Its methods have been primarily developed for contexts where a probability sampling is feasible; under this assumption, survey sampling methods allow us to obtain reliable estimates from a sample of a population, with an associated measure of the variability that arises from the randomness of the sample.

Traditional questionnaire administration modes, such as face-to-face or telephone surveys, have met (to a large extent) the conditions that guarantee probability sampling for a long time. However, in the last few years the winds of change have brought other data sources into the picture in response to the growing issues of those traditional modes (such as drops in response rates or increase of costs). The increasing prevalence of nonprobability surveys, such as web panels, interception surveys or large volume datasets collected automatically that are often used in big data (e.g., lists of tweets or transactions), has brought positive aspects like reducing survey time and cost per respondent, as well as enabling more possibilities for questionnaire design. On the other hand, collecting a strict probability sample using such methods is largely difficult because of the frame undercoverage that arises from drawing the sample from a subset of the target population (such as internet users) and the fact that the respondents are self-selected for many of those methods. These issues make methods for nonprobability samples even more important.

When using the aforementioned data sources for finite population inference, adjusting for selection bias should be considered. Among the various techniques to remove bias in web surveys, we could underline propensity score adjustment (PSA). This method, originally developed for reducing selection bias in non-randomized clinical trials [3], is commonly used for dealing with missing data [4], and was adapted to nonprobability

surveys in the work of [5,6]. Among the alternatives, we could mention the statistical matching method, which is also known as *mass imputation* in the literature, which was developed in [7] as a technique to address selection bias in web surveys by means of predictive modelling.

These methods are often used using logistic models (to estimate the propensity to participate in the survey of each individual) and linear regressions (to predict the values of the interest variable), which may entail several disadvantages for large populations in comparison to modern prediction methods such as ML algorithms.

In recent decades, numerous machine learning (ML) methods have emerged that have proven to be more suitable for regression and classification than linear regression methods. Although there has been an exponential increase in the use of these techniques in many areas [8–10], their application in the context of sampling in finite populations has been limited. A model-assisted estimator based on a neural network with skip-layer connections was developed in [11] . A design-based model-assisted estimator using KNN (K-nearest neighbor method) was developed in [12,13]. Spline regression and random forests in post-stratification were used in [14]. The effects of bagging on non-differentiable survey estimators including sample distribution functions and quantile were invesigated in [15].

Recently, ML algorithms have been considered in the literature for the treatment of nonprobability samples. A simulation study using certain ML predictive algorithms (decision trees, k-nearest neighbors, Naive Bayes, Random Forest and Gradient Boosting Machine) is performed in [16]. Their findings showed that ML methods have the potential to remove selection bias in nonprobability samples to a greater extent than logistic regression in some scenarios. This view had been previously supported by [17]. The use of linear models and some ML algorithms in PSA to estimate propensities and in imputation for statistical matching was compared in [18]. Other recent papers that use Regression Trees and boosting algorithms to remove bias in web surveys are [19,20].

A common machine learning algorithm under the Gradient Boosting framework is XGBoost [21]. The use of this algorithm is motivated by the promising results obtained with boosting algorithms in general and Gradient Boosting Machines (GBM) in particular; for instance, the simulation study from [16] showed that Gradient Boosting Machines can lead to selection bias reductions in situations of high dimensionality, or where the selection mechanism is Missing At Random (MAR). Boosting algorithms have been applied in propensity score weighting for non-randomized experiments, including Gradient Boosting Machines [22–27], showing on average better results than conventional parametric regression models. Given its theoretical advantage over GBM, which could lead to even better results in a broader range of situations, XGBoost will be used for this research to test its adequacy for mitigating selection bias in volunteer samples and lay a baseline performance result. We will apply this algorithm for several estimators based on different approaches.

The paper is organized as follows. In Section 2, the existing methods for correcting selection bias in volunteer samples using a reference probability sample are described. In Section 3, the XGBoost method is presented and its use for estimating population mean in our context is proposed. The results from several simulation studies are presented in Section 4. An application to a real survey is presented in Section 5. Finally, the findings and their implications are discussed in Section 6.

2. Context

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. Let s_V be a convenience (or volunteer) nonprobability sample of size s_V. Let y be the variable of interest in the survey estimation.

The population mean, $\overline{Y}$, can be estimated with the naive estimator based on the sample mean of y in s_V:

$$\hat{\overline{Y}} = \sum_{i \in s_V} \frac{y_i}{n_V} \tag{1}$$

If the convenience sample s_V suffers from selection bias, this estimator will provide biased results. This can happen if there is an important fraction of the population with zero chance of being included in the sample (coverage bias) and if there are significant differences in the inclusion probabilities among the different members of the population (selection bias) [28,29].

Let s_R be a reference sample of size n_R selected from U under a probability sampling design (s_R, p_R) with $\pi_i = \sum_{s_R \ni i} p_R(s_R)$ (where s_R denotes the samples which contain the unit i) the first order inclusion probability for individual i, we denote by $d_i = 1/\pi_i$ the design weights for the units in the reference sample. Let $\mathbf{x}_i$ be the values presented by individual i for a vector of covariates $\mathbf{x}$. Those covariates are common to both samples, while we only have measurements of the variable of interest y for the individuals in the convenience sample.

In this context, propensity score adjustment (PSA) can be used to reduce the selection bias that would affect the unweighted estimates. This approach aims to estimate the propensity of an individual to be included in the nonprobability sample by combining the data from both samples, s_R and s_V, and training a predictive model on the variable δ, with $\delta_i = 1$ if $i \in s_V$ and $\delta_i = 0$ if $i \in s_R$. PSA assumes that the selection mechanism of s_V is ignorable and follows a parametric model:

$$P(\delta_i = 1|\mathbf{x}_i) = p_i(\mathbf{x}) = \frac{1}{e^{-(\gamma' \mathbf{x}_i)} + 1} \tag{2}$$

for some vector γ. The procedure is to estimate the parameter γ by using logistic regression and transform the estimated propensities to weights by inverting them $w_i^{\log} = 1/\hat{p}_i$ where $\hat{p}_i = \hat{p}_i(\mathbf{x}_i) = (e^{-(\hat{\gamma}' \mathbf{x}_i)} + 1)^{-1}$ is the estimated propensity for the individual $i \in s_V$ based on logistic regression. Thus the inverse propensity score weighting estimator (IPSW) [30] is:

$$\hat{\bar{Y}}_{IPSW} = \frac{1}{\sum_{i \in s_V} w_i^{\log}} \sum_{i \in s_V} y_i w_i^{\log} \tag{3}$$

Propensities can be transformed into weights using other procedures, such as stratifying the vector of propensities to form groups of individuals with similar propensities and assign all individuals in a group the same weight [6,31].

If the design weights are used in the computation of γ, the estimator $\hat{\bar{Y}}_{IPSW}$ is valid provided the participation rate is small, given that the optimization procedure leads to the pseudologlikelihood function developed in [32] which provides an unbiased and consistent estimator of the propensities except for an extra term that depends on the size of s_V relative to U, and therefore can be considered as negligible if $U \gg s_V$. A modification of PSA is the TrIPW estimator developed in [19], that uses a modified version of the Classification And Regression Trees (CART) algorithm [33], and does not require the participation rate to be small. Although IPSW and TrIPW can be considered PSA approaches, the methodology of the latter is slighty different, as it takes into account design weights in the tree building by definition, while in the IPSW approach it is not required to use design weights. The propensity for each individual $i \in s_V$ is estimated as:

$$\tilde{p}_i^{CART} = \frac{\#(l(i) \cap s_V)}{\#(l(i))} \tag{4}$$

where $l(i)$ represents the terminal node of the CART algorithm trained on U in which i-th individual of s_V lies. The formula above represents the proportion of population individuals that would be classified in the terminal node 1 and also belong to s_V. Given that $U - s_V$ is not available, the propensity described above has to be estimated from the information contained in the available samples using a modified CART algorithm and

estimating proportions by taking design weights into account to be used for estimating population and subpopulation sizes as follows:

$$\hat{p}_i{}^{CART} = \frac{\#(l(i) \cap s_V)}{\hat{\#}(l(i))} = \frac{\#(l(i) \cap s_V)}{\sum_{j \in l(i) \cap s_R} \frac{1}{\pi_j}} \tag{5}$$

where π_j is the first order inclusion probability for individual j in s_R. The equation above substitutes the unknown number of individuals from the population that would fit in $l(i)$ by its estimated value through the sum of the sampling weights of individuals from s_R that belong to $l(i)$. These values $\hat{p}_i{}^{CART}$ are now used to construct a Hajek type estimator of $\overline{Y}$ as:

$$\hat{\overline{Y}}_{TrIPW} = \frac{1}{\sum_{i \in s_V} w_i^{CART}} \sum_{i \in s_V} y_i w_i^{CART} \tag{6}$$

where $w_i^{CART} = 1/\hat{p}_i{}^{CART}$. This non-parametric approach shows acceptable results under non-linearity conditions [19].

In a similar way to PSA, propensity scores are used to measure the similarity between the covariates of the probabilistic and nonprobability samples. The new approach is called Kernel Weighting [34]. These propensity scores were made through the use of logistic regression, as explained previously.

For $j \in s_R$ we compute the distance of its estimated propensity score from each i in the nonprobability sample (whose result varies from -1 to 1) as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \hat{p}_i(\mathbf{x}_i) - \hat{p}_j(\mathbf{x}_j) \tag{7}$$

Then, a zero-centered kernel function is applied to smooth distances. Thus, the pseudoweights can be calculated:

$$k_{ij} = \frac{K\{d(\mathbf{x}_i, \mathbf{x}_j)/h\}}{\sum_{j \in s_V} K\{d(\mathbf{x}_i, \mathbf{x}_j)/h\}} \tag{8}$$

where $K(\cdot)$ is the applied kernel function (i.e., Gaussian):

$$K\big(d(\mathbf{x}_i, \mathbf{x}_j); h\big) \propto \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2h^2}\right) \tag{9}$$

and h is the bandwidth. To calculate the optimal bandwidth, Silverman's method is used [35]:

$$h = 0.9 \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-\frac{1}{5}} \tag{10}$$

where $\hat{\sigma}$ is the square root of the variance, IQR is the interquartile range and n is the length of the distances vector. Finally the KW weight is given by:

$$w_i = \sum_{j \in s_R} k_{ij} d_j \tag{11}$$

and the KW estimator of the population mean is:

$$\hat{\overline{Y}}_{KW} = \frac{1}{\sum_{i \in s_V} w_i^{KW}} \sum_{i \in s_V} y_i w_i^{KW}. \tag{12}$$

Another variation of KW is Boosted Kernel Weighting. Its only difference is the usage of machine learning instead of logistic regression to get the propensities [20]. These authors use four ML methods: model-based recursive partitioning, conditional random forests, gradient boosting machines and model-based boosting to estimate propensities and deduce in their simulation study that boosting methods result in KW with lower bias in several settings without increasing variance.

PSA is often used for reducing selection bias in nonprobability surveys, but empirical evidence of its effectiveness is mixed. A study with four web panel surveys was developed in [36], showing that the reduction in bias is likely to be partial and unpredictable . Alternative methods for selection bias adjustment are based in superpopulation models. Statistical matching (SM) is an approach developed by [7] and applied to nonresponse treatment in [37]. This method aims to predict y in the probability sample (where y has not been measured) using covariates $\mathbf{x}$ and the volunteer sample s_V to fit the models that will be used to predict values of y in the reference sample. SM assumes that y is a realization of a superpopulation random variable Y, which follows a functional relationship with the set of covariates $\mathbf{x}$ such that:

$$y_i = m(\mathbf{x}_i) + e_i,\ i = 1, 2, \ldots, N, \tag{13}$$

It is often assumed that the relationship between y and $\mathbf{x}$ is linear, meaning that $m(x_i) = \beta \mathbf{x}_i$, the random vector $e = (e_1, \ldots, e_N)'$ is assumed to have zero mean and the coefficients β can be estimated by the usual methods in linear regression such as Ordinary Least Squares or maximum likelihood. The matching estimator is then given by:

$$\hat{\bar{Y}}_{SM} = \frac{1}{\sum_{i \in s_R} d_i} \sum_{s_R} \hat{y}_i d_i \tag{14}$$

where $\hat{y}_i$ the imputed value of y_i and d_i the design weight of the individual i in s_R.

It remains unclear which of the two methods (PSA or SM) is more efficient, although a recent experiment by [18] showed a higher efficiency of statistical matching.

Recently, [32] proposed a new doubly robust estimator based on the previous linear model (13), and showed that this estimator can be conveniently used for inferences from nonprobability samples. The estimator is defined as:

$$\hat{\bar{Y}}_{DR} = \frac{1}{\sum_{i \in s_R} d_i} \sum_{s_R} \hat{y}_i d_i + \frac{1}{\sum_{i \in s_V} 1/\hat{p}_i(\mathbf{x}_i)} \sum_{i \in s_V} (y_i - \hat{y}_i)/\hat{p}_i(\mathbf{x}_i) \tag{15}$$

This estimator follows the idea of the model-assisted generalized difference estimator given in [38] and has the property of being robust to modelling misspecifications either in the propensity estimation or in the matching imputation.

Alternatively, a more direct method has been proposed in [39] to combine SM and PSA. The main idea is to use PSA weights in the predictive models used in Statistical Matching, given that those models use the nonprobability sample as training data. This is a feasible strategy given that most machine learning algorithms allow the weighting of the training data. For example, the previous linear model (13) can minimize a weighted Mean Square Error instead. Let $\hat{y}_{ti}$ the value of y_i imputed by a model trained that uses $1/\hat{p}_i(\mathbf{x}_i),\ i \in s_V$ as training weights. The proposed estimator will be:

$$\hat{\bar{Y}}_{WT} = \frac{1}{\sum_{i \in s_R} d_i} \sum_{s_R} \hat{y}_{ti} d_i. \tag{16}$$

In the next section we introduce a powerful machine learning technique that can be used both for predicting the unknown values in the probability sample (which can be used to obtain the imputed values in the estimators described previously) and also for calculating the propensity scores.

3. XGBoost Estimators

We assume that covariates **x** have been measured on both samples, while the variable of interest y has been measured only in the volunteer sample, s_R.

We will use XGBoost to obtain the imputed values in the matching estimator. XGBoost is a widely known state-of-the-art machine learning system for several problems. For example, it was used in 17 out of 29 winning solutions published during 2015 at Kaggle, a famous machine learning platform for hosting competitions [21].

It works as a decision tree ensemble. Decision trees set split points based on $\mathbf{x}_i$ until reaching a final estimation $\hat{y}_i$ of y_i.

As described in the original paper [21], when they work as an ensemble model the final prediction is defined as follows:

$$\hat{y}_{xgi} = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \tag{17}$$

where K is the number of trees forming the ensemble and $\mathcal{F} = \{f(\mathbf{x}) = \omega_{q(\mathbf{x})}\}$; with $q : \mathbb{R}^m \to T$ representing the structure of each tree which, given $\mathbf{x}_i$, returns its corresponding final node and ω_i the score on the i-th final node. The final prediction is the sum of the scores obtained.

The trees f_k, $k = 1, \ldots, K$, are built aiming to minimize the following regularized objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_{xgi}, y_i) + \sum_k \Omega(f_k) \tag{18}$$

where the first term l is a differentiable convex function which measures the error of the estimations. For example, when estimating a quantitative variable, the squared error can be used:

$$l(\hat{y}, y) = (\hat{y} - y)^2 \tag{19}$$

The second term regularizes the function penalizing complex trees. It penalizes having too many final nodes (T) and returning too high scores:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{20}$$

where γ and λ are hyperparameters which control how much is this regularization prioritized to control overfitting [40] over minimizing the error for the training set.

The objective function is minimized iteratively with the Gradient Tree Boosting method [41]. For the t-th iteration, f_t is added in order to minimize the following objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_{xgi}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{21}$$

where $\hat{y}_{xgi}^{(t)}$ is the estimated value of y for the i-th unit in the t-th iteration. This objective is optimized via second-order approximation [42]:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_{xgi}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{22}$$

where $g_i = \partial_{\hat{y}_{xgi}^{(t-1)}} l(y_i, \hat{y}_{xgi}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}_{xgi}^{(t-1)}} l(y_i, \hat{y}_{xgi}^{(t-1)})$.

In practice, it is impossible to evaluate every possible tree structure q. The loss reduction caused by a potential split point is calculated instead as:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{23}$$

where I_L and I_R are the sets of units corresponding to the left and right side of the split, and $I = I_L \cup I_R$. Split points are added iteratively based on this formula.

XGBoost implements Gradient Tree Boosting with several techniques which improve its efficiency and efficacy. These include shrinkage (in order to limit the influence of each individual tree) and advanced strategies for finding split point candidates, among others [21].

By imputing missing values in the target variable for individuals in the probability sample with their corresponding predicted value, we propose the following SM estimator for the population mean $\overline{Y}$:

$$\hat{\overline{Y}}_{XGM} = \frac{1}{\sum_{i\in s_R} d_i}\sum_{s_R} \hat{y}_{xgi} d_i, \tag{24}$$

where $\hat{y}_{xgi}$ the predicted value of y_i.

Other possibility to make estimators is to consider the idea of generalized difference estimator [43] where an additional term is added to the $\hat{\overline{Y}}_{XGM}$ estimator that takes into account the error made in the estimates given by the model from the nonprobabilistic sample (since in this sample we have the true and the estimated values for y).

Following this idea we propose the estimator:

$$\hat{\overline{Y}}_{XGD} = \frac{1}{\sum_{i\in s_R} d_i}\sum_{s_R} \hat{y}_{xgi} d_i + \frac{1}{\sum_{i\in s_V} 1/\hat{p}_i(\mathbf{x}_i)}\sum_{i\in s_V} (y_i - \hat{y}_{xgi})/\hat{p}_{xgi}(\mathbf{x}_i) \tag{25}$$

where $\hat{p}_i = (e^{-(\hat{\gamma}\prime \mathbf{x}_i)} + 1)^{-1}$. This estimator is similar to the the doubly robust estimator by [32], but they use parametric regression models for estimating y_i.

XGBoost also allows weighting the training data. First we estimate the propensities by logistic regression. Then, the model is trained using the weights $w_i^{\log} = 1/\hat{p}_i; i \in s_V$ in the objective function. Let $\hat{y}_{xgti}$ be the value of y_i imputed by said model. Finally, we make the XGT-estimator:

$$\hat{\overline{Y}}_{XGT} = \frac{1}{\sum_{i\in s_R} d_i}\sum_{s_R} \hat{y}_{xgti} d_i. \tag{26}$$

Finally, a new kernel weighting estimator $\hat{\overline{Y}}_{XKW}$ can be considered, as detailed in (12), but using XGBoost for estimating propensities. That is, the proposed estimator is formulated as:

$$\hat{\overline{Y}}_{XKW} = \frac{1}{\sum_{i\in s_V} w_i^{XKW}}\sum_{i\in s_V} y_i w_i^{XKW}. \tag{27}$$

where $w_i^{XKW} = \sum_{j\in s_R} k_{Wij} d_j$ and k_{Wij} are calculated as in (8) but the propensities p_i are estimated using the XGBoots method as

$$\widehat{p}_{iX} = \varphi(\mathbf{z}_i) = \sum_{k=1}^{K} g_k(\mathbf{z}_i), \;\; g_k \in \mathcal{G} \tag{28}$$

where $\mathcal{G}$ representing the structure of each tree and $\mathbf{z}_i$ the covariates used for modelling the propensities (that may or may not coincide with the variables used to predict the outcome variable y).

The proposed XGBoost estimators (24)–(27) are computationally similar, given that the algorithm does the same work in all of them. However, the XGBoosted kernel weighting variant will be computationally preferable when there are many variables to estimate because only one model has to be trained in order to calculate the weights. Even though XGBoost models are more expensive to train than linear models, training time is insignificant for a single model in any modern processor. However, the difference could be significant when many models have to be trained. The efficiency of each method can be studied by analyzing the variance of the resulting estimator; however, that variance cannot be developed in simple form. Alternatively, resampling methods can be applied to each of the proposed estimators to estimate the variance (see [44]).

3.1. Hyperparameter Optimization

The XGBoost algorithm contains several tuning hyperparameters which determine its functioning for each specific case. Its default values may be used. However, poor results may be obtained due to the fact that said default values are not suitable for some cases. In order to determine its real potential, we will also consider a hyperparameter optimization process for the matching estimator $\hat{\bar{Y}}_{XGM}$ and for the Boosted Kernel Weighting estimator $\hat{\bar{Y}}_{BKW}$. This will also determine how relevant these kind of optimizations can be.

The process will be carried out via the Tree-structured Parzen Estimator (TPE) algorithm [45]. Each tested hyperparameters set will be validated calculating its Rooted Mean Squared Error for several simulations in order to determine the optimal values. In a real case scenario, simulations cannot be carried out and therefore this strategy should be replaced with cross-validation techniques [46].

Among the wide variety of parameters considered by XGBoost, we have selected the most important ones for the search space:

- Number of estimators $\in [10, 400]$: How many trees form the ensemble. The default value is 100.
- Learning rate $\in [0.01, 1]$: How much weight shrinkage is applied after each boosting step. The default value is 0.3.
- Maximum depth $\in [1, 60]$: How many splits can each tree contain. The default value is 6.
- Minimum child weight $\in [1, 6]$: How much instance weight is needed in total to consider a new partition. The default value is 1.

4. Simulation Study

4.1. Simulated Populations

Several simulation experiments are performed in order to demonstrate how much XGBoost can improve the estimations obtained with classic logistic/linear regression.

The first experiment replicates the simulated populations used in the study by [47]. The populations and propensities proposed are replicated, but XGBoost is introduced as the machine learning algorithm used for each estimator proposed. This way, its performance can be compared with the results obtained using logistic/linear regression (the algorithm used in the original paper). The methodological rationale behind the use of this study is to explore the behavior of XGBoost in those situations where the relationship between covariates and target variables is non-linear, and therefore cannot be represented by linear regression if it is not explicitly stated by the practitioner when specifying the model. XGBoost (and other Machine Learning algorithms) are able to represent those non-linearities via boosted decision trees based on learning from data. On the other hand, using artificial data allows us to control the selection mechanisms and the relationships between variables, as well as assess their relevance in the final results. When using real data, these relationships can only be drawn in a conjectural way, although the results might be more representative of real world situations.

Therefore, three finite populations are generated following these models:

$$\xi_1: \ y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + \sigma_a \epsilon_i, \ \ i = 1, 2, ..., N; \tag{29}$$

$$\xi_2: \ y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + 0.2x_{3i}^4 + \sigma_b \epsilon_i, \ \ i = 1, 2, ..., N; \tag{30}$$

$$\xi_3: \ y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + 0.5x_{3i}^4 + \sigma_c \epsilon_i, \ \ i = 1, 2, ..., N; \tag{31}$$

where $N = 20{,}000$, $x_{1i} = z_{1i}$, $x_{2i} = z_{2i} + 0.3x_{1i}$ and $x_{3i} = z_{3i} + 0.3(x_{1i} + x_{2i})$; with $z_{1i} \sim Bernoulli(0.5)$, $z_{2i} \sim Uniform(0, 2)$ and $z_{3i} \sim N(0, 1)$. $\epsilon_i \sim N(0, 1)$ is the error term, controlled by σ_a, σ_b and σ_c. Their values are adjusted in order to set the correlation coefficient, ρ, between y with and without the error term at some desired level.

The propensities π_i^A for the nonprobabilistic samples are generated following these three models:

$$q1: \ \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_a + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i}, \ \ i = 1, 2, \ldots, N; \tag{32}$$

$$q2: \ \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_b + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i} + 0.1x_{3i}^2, \ \ i = 1, 2, \ldots, N; \tag{33}$$

$$q3: \ \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_c + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i} + 0.2x_{3i}^2, \ \ i = 1, 2, \ldots, N; \tag{34}$$

where θ_a, θ_b and θ_c are set such that $\sum_{i=1}^N \pi_i^A = n_V$ for each case, with n_V the target sample size.

The probabilistic samples are obtained using inclusion probabilities proportional to $z_i = c - x_{2i}$, with c such that $\max z_i / \min z_i = 30$.

Using the described probabilities, a nonprobabilistic sample s_V of size $n_V = 500$ and a probabilistic sample s_R of size $n_R = 1000$ are repeatedly drawn from the chosen population. The proposed estimators are applied with said samples so the metrics, relative bias ($\%RB$) and mean square error (MSE), are obtained as follows:

$$\%RB = \frac{1}{B}\sum_{b=1}^{B} \frac{\hat{\mu}^{(b)} - \mu_y}{\mu_y} \times 100, \quad MSE = \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\mu}^{(b)} - \mu_y\right)^2 \tag{35}$$

where $\hat{\mu}^{(b)}$ is the mean estimated from the b-th sample and $B = 2000$.

The estimators considered are: the unweighted sample mean ($\hat{\bar{Y}}$), IPSW with logistic regression ($\hat{\bar{Y}}_{IPSW}$), Tree-Based Inverse Propensity Weighted estimation($\hat{\bar{Y}}_{TrIPW}$), Kernel Weighting ($\hat{\bar{Y}}_{KW}$), Matching with linear regression ($\hat{\bar{Y}}_{SM}$), Doubly Robust with linear regression for Matching and logistic regression for PSA ($\hat{\bar{Y}}_{DR}$), Training with linear regression for Matching and logistic regression for PSA ($\hat{\bar{Y}}_{WT}$), XGBoosted kernel weighting ($\hat{\bar{Y}}_{XKW}$), Matching with XGBoost ($\hat{\bar{Y}}_{XGM}$), Doubly Robust with linear regression for PSA and XGBoost for Matching ($\hat{\bar{Y}}_{XGD}$) and Training with linear regression for PSA and XGBoost for Matching ($\hat{\bar{Y}}_{XGT}$). For those using XGBoost, only its default hyperparameters are considered in this simulation.

The results for every possible population/propensities combination, with different values of the correlation coefficient ρ, can be consulted in Figures 1–6.

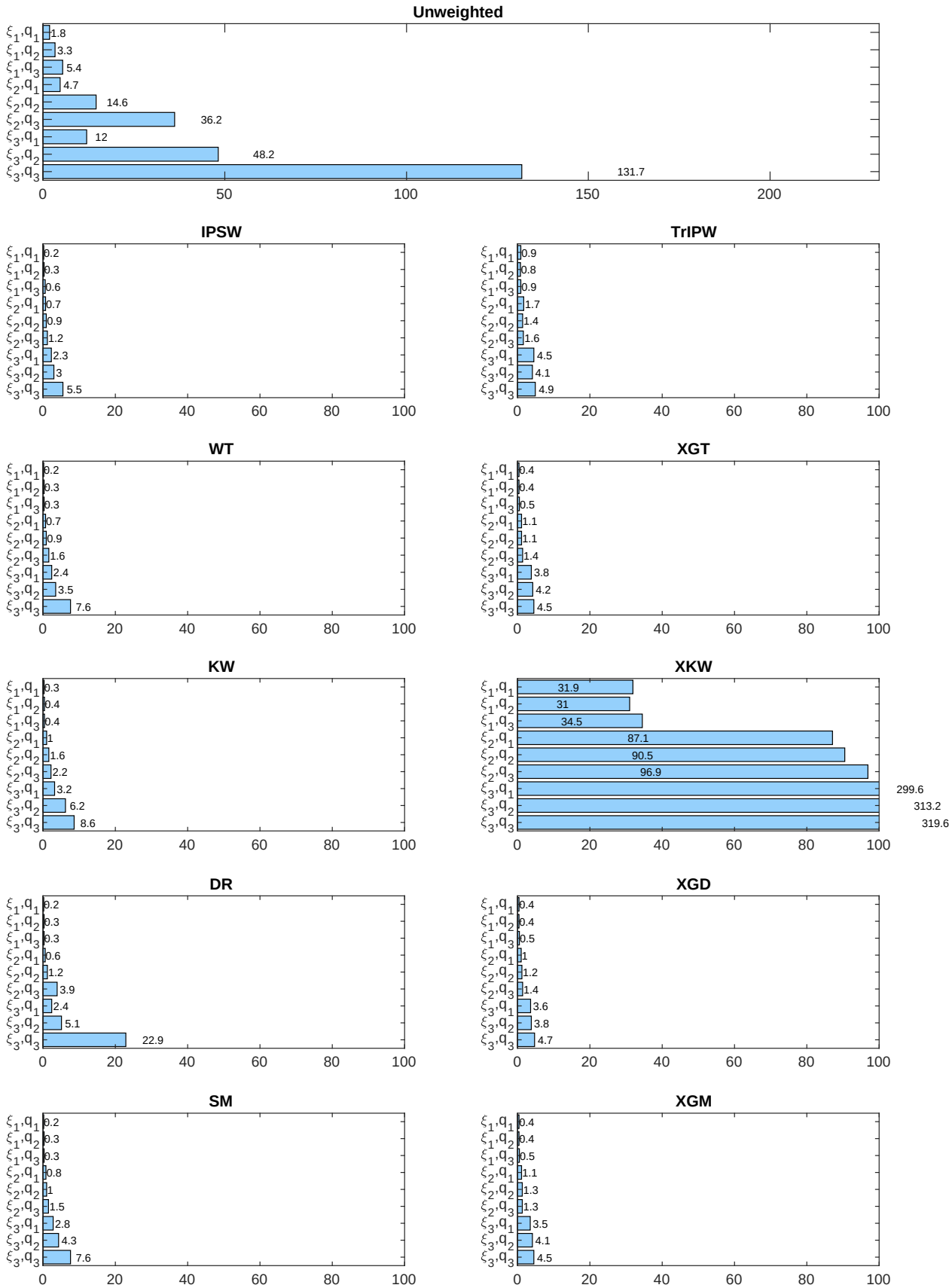


Figure 1. MSE, simulated case, correlation coefficient: 0.3.

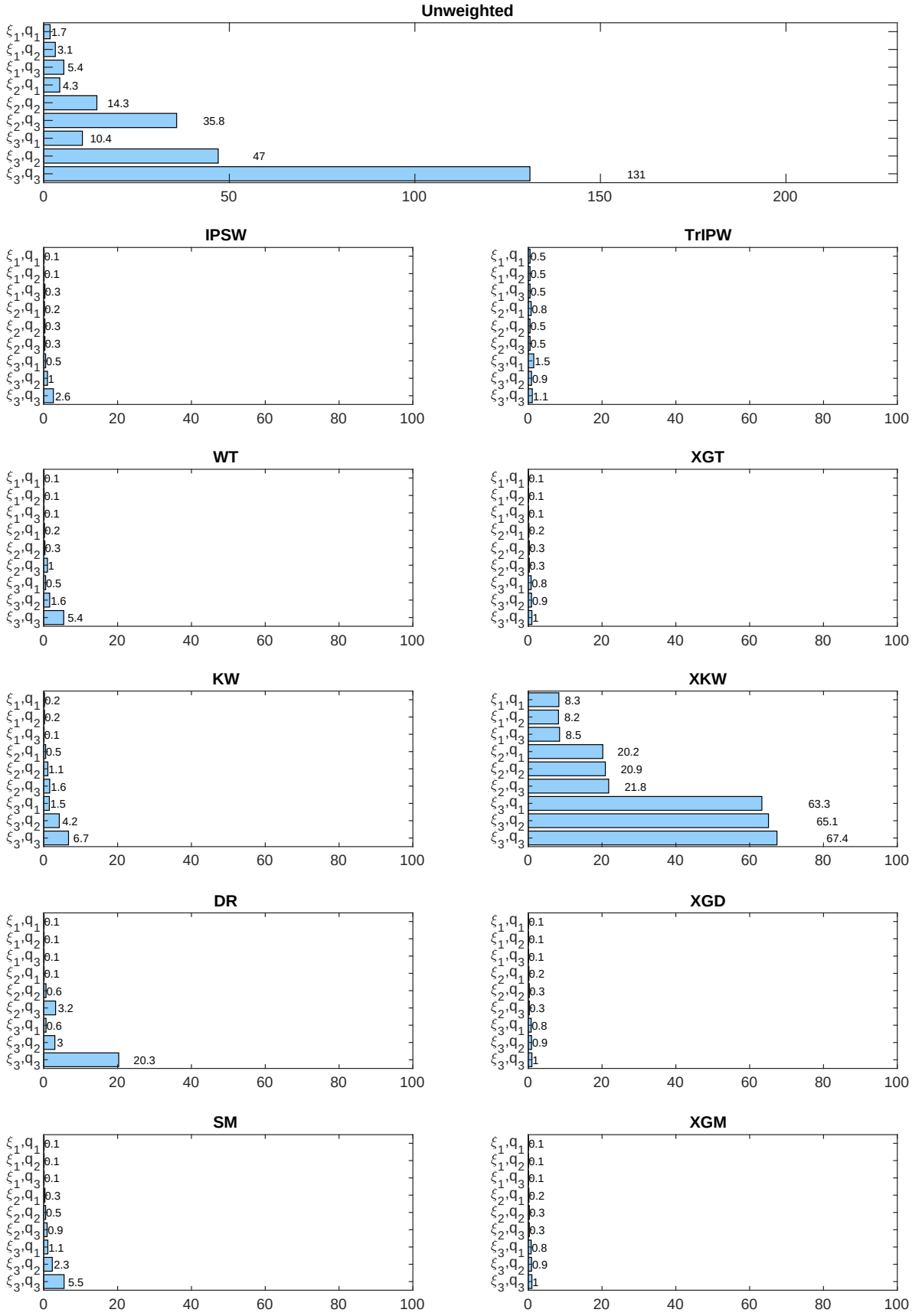


Figure 2. MSE, simulated case, correlation coefficient: 0.6.

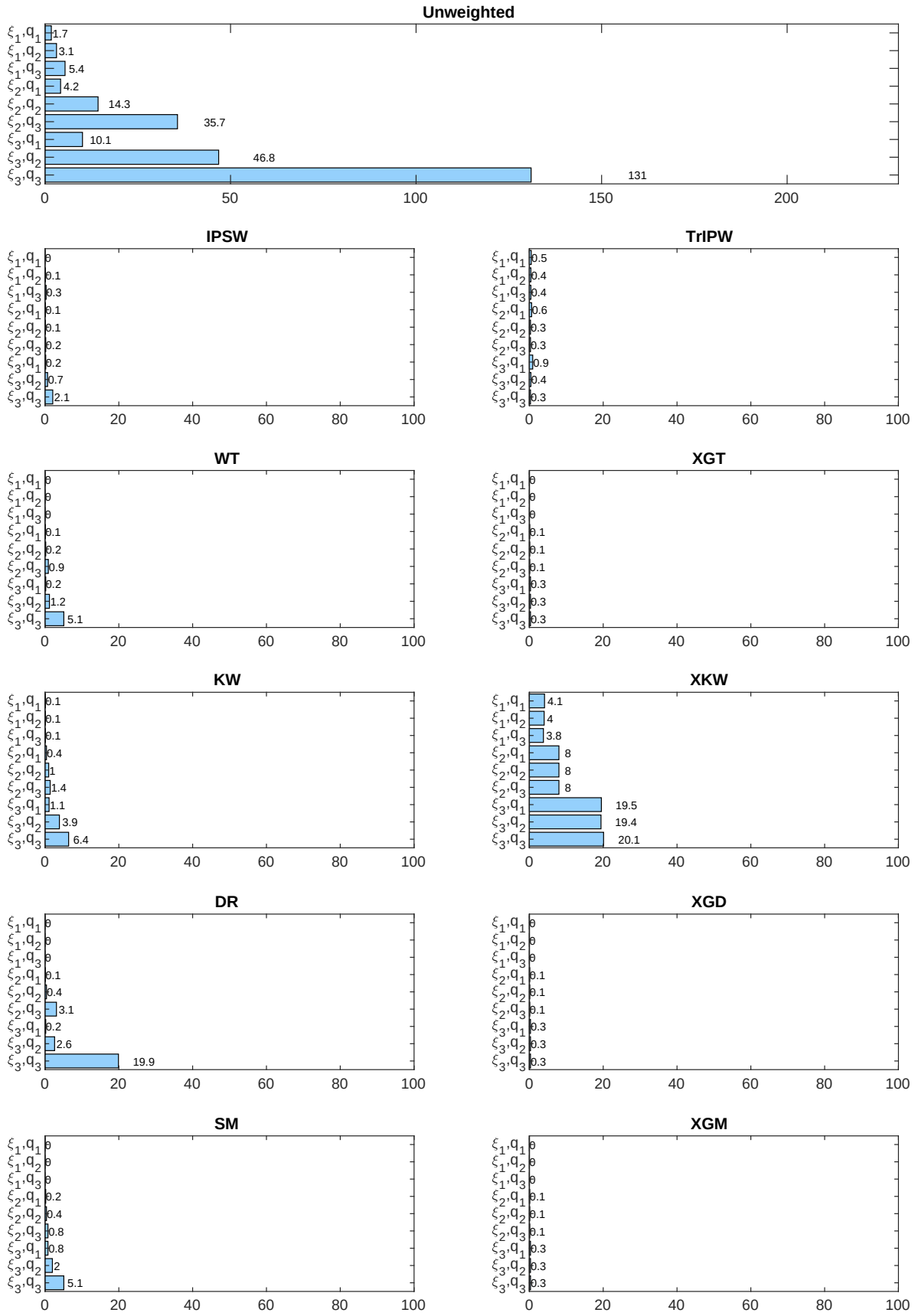


Figure 3. MSE, simulated case, correlation coefficient: 0.9.

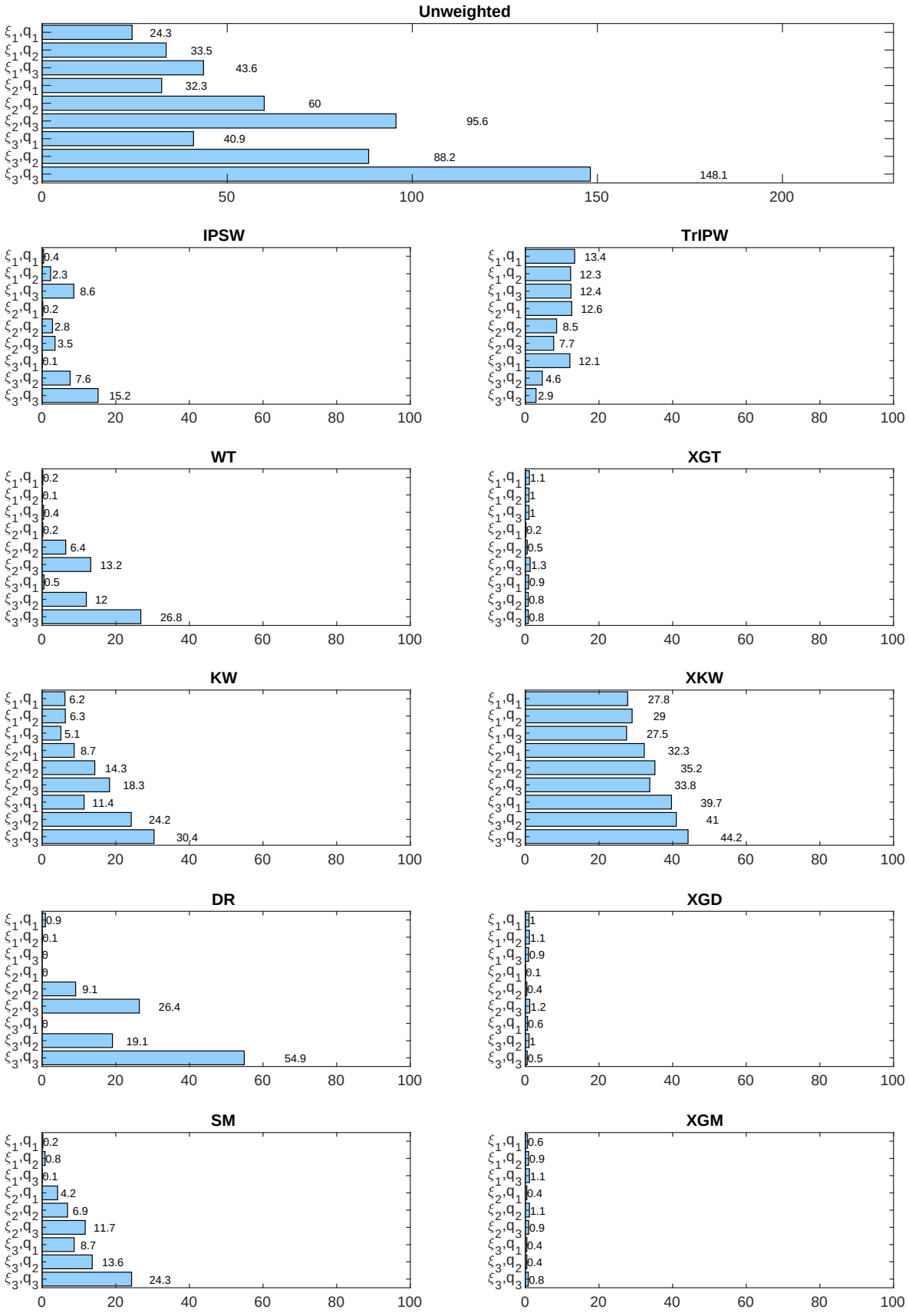


Figure 4. Relative bias (%), simulated case, correlation coefficient: 0.3.

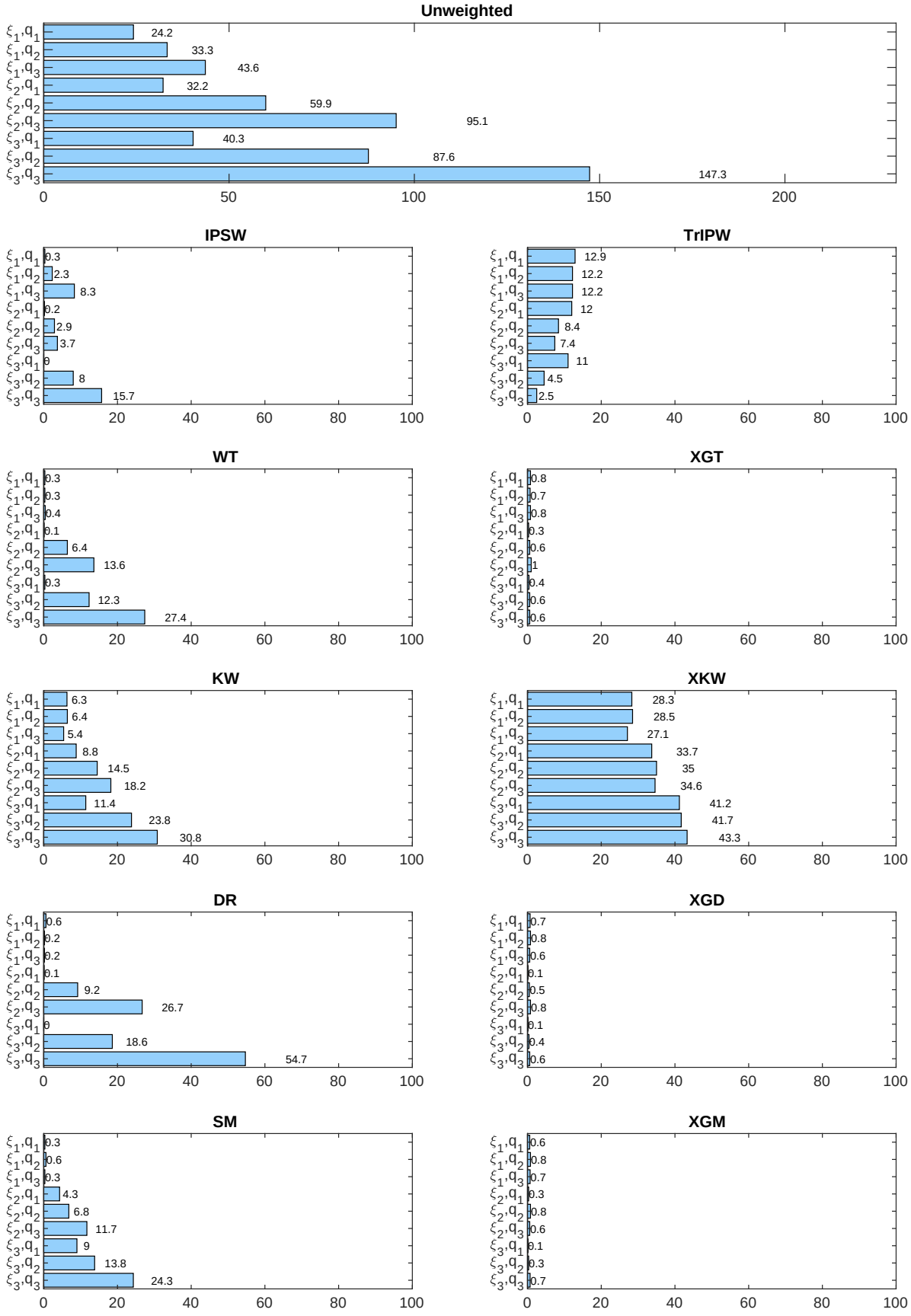


Figure 5. Relative bias (%), simulated case, correlation coefficient: 0.6.

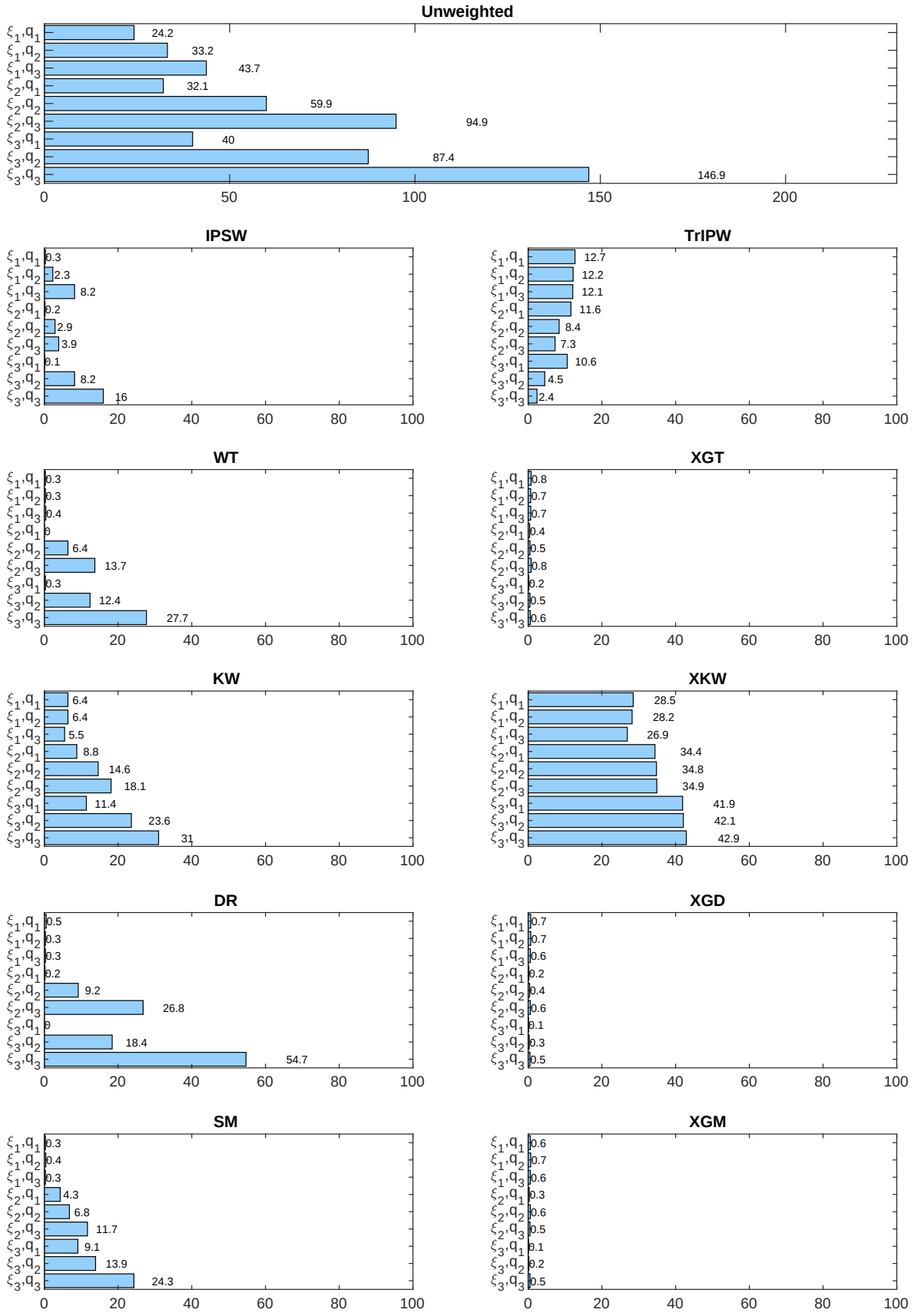


Figure 6. Relative bias (%), simulated case, correlation coefficient: 0.9.

Models ξ_1 and q_1 are linear models. Therefore, linear/logistic regression is theoretically unbeatable for those models. However, it can be observed that XGBoost can also effectively remove the bias in those cases. The difficulties of linear/logistic regression arise as the non-linearity of the models is increased. XGBoost is, however, still able to learn the model in those scenarios. The decrease in bias and MSE of the XGBoost technique with respect to linear/logistic regression is very noticeable in the case of the ξ_3 and q_3 model, and it is observed how this good behavior is accentuated as the correlation between the variables increases.

That is not the case for the $\hat{\bar{Y}}_{TrIPW}$ or $\hat{\bar{Y}}_{XKW}$ estimators. They seem to be suffering from overfitting [40]. Further analysis from simulations considering real populations and hyperparameter optimization will determine if their performance can be fixed.

Regarding doubly robust estimators, again the high learning capacity of Matching with XGBoost causes that combining it with PSA does not necessarily improves the results. In practice, the complexity of real data models may change that fact.

4.2. Real Populations

Following the experiment described in the previous section, the study is repeated with real populations. The same estimators are considered. Default XGBoost hyperparameters are used for an initial simulation. The relative bias is kept as a metric but the mean squared error is replaced by the relative rooted mean squared error ($\%RRMSE$) in order to obtain comparable results.

$$\%RRMSE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}^{(b)} - \mu_y\right)^2} \Big/ \mu_y \times 100 \tag{36}$$

Two datasets are used following two different sampling strategies for each one. In each simulation run, three possibilities for sample sizes, $n_V = n_R = 1000$, $n_V = n_R = 2000$ and $n_V = n_R = 5000$, are considered.

The first population, denoted as P1, corresponds to the Hotel Booking Demand Dataset [48]. It includes the data of bookings for a resort hotel and a city hotel due to arrive between the 1 July 2015 and 31 August 2017. In total, it has 119,390 bookings of which 34% are from the resort hotel and 66% from the city hotel. For the first nonprobability sampling strategy, denoted as S1, resort bookings have 10 times more probability of being chosen than city bookings. For the second nonprobability sampling strategy, denoted as S2, city bookings have five times more probability of being chosen than resort bookings. The target variable is the mean number of weeknights (Friday included) which are booked. In order to estimate it, a probability sample s_R is also obtained via a simple random sampling. The remaining variables included in the dataset are used as covariates, excluding the reservation status and the reservation status date, with a total of 28 covariates.

The second population, denoted as P2, is the Adult Dataset [49]. It includes census income information for 32,561 adult individuals from the 1994 Census database of the United States. For the first nonprobability sampling strategy, denoted as S1, individuals who make over \$50K a year have double the probability of being chosen. For the second nonprobability sampling strategy, denoted as S2, individuals who make over \$50K per year have a propensity to participate multiplied by $Pr(a) = 2a^2$, where a is the individual's age. The target is estimating the proportion of individuals who make over \$50K per year. Therefore, in this case, the target variable in the dataset is binary instead of continuous. Also, in this scenario, the propensities depend on the target variable itself and this dependance may not even be linear. Every other variable in the dataset is used as covariate, for a total of 14 covariates. The probabilistic samples are obtained via simple random sampling.

The bias and relative rooted mean squared error results for each case with each estimator can be viewed in Tables 1 and 2 respectively.

Table 1. Relative bias (%) for each real population case.

	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_{IPSW}$	$\hat{\bar{Y}}_{TrIPW}$	$\hat{\bar{Y}}_{KW}$	$\hat{\bar{Y}}_{SM}$	$\hat{\bar{Y}}_{DR}$	$\hat{\bar{Y}}_{WT}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$	$\hat{\bar{Y}}_{XGD}$	$\hat{\bar{Y}}_{XGT}$
P1S1 1000	18.9	5.5	11.1	3.7	4.5	4.6	4.5	0.2	3.5	3.5	3.3
P1S1 2000	18.9	5.5	10.9	4	4.9	4.9	4.8	−11.9	2.8	2.8	2.5
P1S1 5000	18.6	4.6	10.1	4.2	4.8	4.8	4.7	−7.5	2.2	2	1.7
P1S2 1000	−9.2	−4.1	−5.4	−2.1	−5	−4.1	−4.1	−13.4	−2.6	−2.5	−2.5
P1S2 2000	−9.2	−4.2	−5.5	−2	−4.9	−4.1	−3.9	−7.5	−1.9	−1.8	−1.8
P1S2 5000	−9.1	−3.9	−5.2	−2.4	−4.7	−3.8	−3.6	1.4	−1.4	−1.3	−1.3
P2S1 1000	60	34.4	37	33.5	33.2	33.2	30	8.9	25.9	25.8	24.8
P2S1 2000	58.7	33.3	36	33.1	30.8	30.5	29.2	−12	25	24.7	24
P2S1 5000	54.8	31.3	33.7	30.7	31.1	27.9	27.6	−11.8	23.4	23.2	22.8
P2S2 1000	78.3	34.8	39.8	33	34.9	33.8	31	−5.6	26.4	25.9	24.4
P2S2 2000	76.5	33.9	39.1	32.4	32.2	31.2	30.2	−31.1	25	24.9	23.6
P2S2 5000	71.1	31.7	36.6	30.3	30.6	28.5	28.2	−19.4	23.3	23	22.4

Table 2. Relative RMSE (%) for each real population case.

	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_{IPSW}$	$\hat{\bar{Y}}_{TrIPW}$	$\hat{\bar{Y}}_{KW}$	$\hat{\bar{Y}}_{SM}$	$\hat{\bar{Y}}_{DR}$	$\hat{\bar{Y}}_{WT}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$	$\hat{\bar{Y}}_{XGD}$	$\hat{\bar{Y}}_{XGT}$
P1S1 1000	19.1	6.3	11.7	5.4	5.6	5.5	5.4	17.4	4.7	4.7	4.6
P1S1 2000	18.9	5.9	11.2	4.9	5.4	5.3	5.3	20.6	3.6	3.6	3.4
P1S1 5000	18.7	8.6	10.3	4.4	5	5.6	4.9	8.8	2.5	2.5	2.2
P1S2 1000	9.5	5.7	5.9	5.9	5.9	5.3	5	20	3.9	3.9	3.9
P1S2 2000	9.3	4.8	6	4.2	5.3	4.7	4.4	19.5	2.8	2.7	2.7
P1S2 5000	9.2	4.2	5.4	3	4.8	4	3.8	11	1.9	1.8	1.8
P2S1 1000	60.3	35	37.6	34.2	33.8	33.9	30.7	77	26.9	26.7	25.7
P2S1 2000	58.9	33.5	36.3	33.4	31.1	30.8	29.5	39.6	25.4	25.1	24.4
P2S1 5000	54.9	31.4	33.8	30.9	31.8	28	27.7	15.8	23.5	23.3	22.9
P2S2 1000	78.5	35.4	40.4	33.7	35.4	34.3	31.6	69.4	27.2	26.8	25.3
P2S2 2000	76.6	34.2	39.4	32.7	32.5	31.5	30.5	40.2	25.4	25.4	24.1
P2S2 5000	71.1	31.8	36.7	30.4	30.9	28.7	28.3	20	23.5	23.2	22.6

Again, as it happened with the simulated data, a significant improvement in the estimations can be observed when using XGBoost instead of linear or single tree regressors. This improvement is more relevant now since the datasets are more complex and closer to real scenarios. The results are also better, as more data is avaliable. In the majority of cases, the Matching based variants obtain the best results. However, for some specific cases, XGBoosted Kernel Weighting is better. This probably happens where the algorithm is not overlearning. This assumption is confirmed by later simulations considering hyperparameter optimization in which the methods always behave reliably.

Regarding doubly robust estimators, combining SM with PSA may yield slightly more accurate estimations in these cases with XGBoost as well. This improvement can be more noticeable if a more direct approach like $\hat{\bar{Y}}_{XGT}$ is applied instead of a basic combination like $\hat{\bar{Y}}_{XGD}$.

Some of these results may be improved by applying variable selection, specifically those using linear of logistic regression. Tree based algorithms like XGBoost or CART apply variable selection internally by themselves.

Finally, as explained in Section 3.1, hyperparameter optimization is also considered via the Tree-structured Parzen Estimator (TPE) algorithm [45], as implemented in the software package *Optuna* [50]. The TPE algorithm is able to quickly discard inappropiate settings, so a wide search space may be specified. We have run simulations for the boosted matching estimator $\hat{\bar{Y}}_{XGM}$ and for the XGBoosted kernel weighting estimator $\hat{\bar{Y}}_{XKW}$. The sample size for this scenario is 1000 since it is the hardest case. Each hyperparameter set evaluated by

the algorithm is validated measuring its Mean Squared Error among 50 sub-simulations. Once the best values for each specific case are selected with this procedure, they are used for a new simulation in the same conditions as the one without optimization. Every real population and sampling strategy is considered.

The results can be observed in Tables 3 and 4. The optimization considerably improves the estimations. In some cases, this improvement is so significant that the method which was the worst one without optimization is now the best alternative. Therefore, the importance of applying this kind of procedure is confirmed in order to obtain reliable results, especially for those estimators that have shown to suffer greatly from overlearning.

Table 3. Relative bias (%) for each optimized case.

		Non Optimized		Optimized	
	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$
P1S1 1000	18.9	0.2	3.5	0.4	1.2
P1S2 1000	−9.2	−13.4	−2.6	−1.1	−1.5
P2S1 1000	60.0	8.9	25.9	5.2	25.1
P2S2 1000	78.3	−5.6	26.4	2.0	25.5

Table 4. Relative RMSE (%) for each optimized case.

		Non Optimized		Optimized	
	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$
P1S1 1000	19.1	17.4	4.7	4.0	3.2
P1S2 1000	9.5	20.0	3.9	4.1	3.4
P2S1 1000	60.3	77.0	26.9	10.6	26.2
P2S2 1000	78.5	69.4	27.2	7.8	26.5

5. Application to a Survey on Social Effects of COVID-19 in Spain

This section illustrates the estimation procedures that we have empirically described in a web survey in which respondents were selected by targeting Internet ads at specific profiles.

ESPACOV [51] is a survey that was conducted in Spain in the fourth week of the strict lockdown imposed on 14 March 2020, and provides information on the living conditions of the population, acquired habits, health and consequences of the state of alarm and home confinement. ESPACOV was run by the Institute for Advanced Social Studies (IESA) and the sample was collected via paid advertisements on Google Ads and Facebook/Instagram (nonprobability sampling). A total of 1881 interviews were completed.

Table 5 compares unweighted sample distributions by age group and sex and by education level with Spanish population data [52,53].

Due to coverage and participation bias, people with tertiary education are over-represented, and less educated people vastly under-represented. There are also representation issues in the different age groups for each sex.

We have considered the April 2020 Barometer of the Spanish Center for Sociological Research [54] as the source of auxiliary information. The barometers are probability surveys carried out on a monthly basis, and their main objective is to measure Spanish public opinion at that time. They involve interviews with approximately 2500 randomly-chosen people from all over the country, with extensive social and demographic information on them being gathered for analysis as well as their opinions. The survey follows a multi-stage, stratified cluster sampling, with selection of the primary sampling units (municipalities) and of the secondary units (census sections) randomly with proportional allocation, and of the last units (individuals) by random routes and sex and age quotas. The barometer dataset is often viewed as a reliable source of official statistics and contains a number of common variables with the ESPACOV dataset. More precisely, these include gender, age,

province, municipality size, education level, working status and self-positioning in the ideological scale (10-point Likert, where 1 represents "far left" and 10 "far right").

Table 5. Obtained sample distributions by sex and age group and by education level, and comparison with population parameters.

	ESPACOV Sample	Spanish Population
Age group		
Men		
18–29	9.7	7.6
30–44	9.3	12.9
45–64	11.3	17.6
65+	16.1	10.3
Women		
18–29	10.6	7.3
30–44	13.7	12.9
45–64	17.9	17.9
65+	11.6	13.5
Education		
Obligatory or less	16.2	45.6
Secondary	33.8	21.7
Tertiary	49.6	32.7

We apply the proposed methods to estimate the population mean of the variable "Rate the government action to control the pandemic, from 0 to 10". The values of the estimators $\hat{\bar{Y}}_{IPSW}$, $\hat{\bar{Y}}_{TrIPW}$, $\hat{\bar{Y}}_{KW}$, $\hat{\bar{Y}}_{SM}$, $\hat{\bar{Y}}_{DR}$, $\hat{\bar{Y}}_{WT}$, $\hat{\bar{Y}}_{XKW}$, $\hat{\bar{Y}}_{XGM}$, $\hat{\bar{Y}}_{XGD}$ and $\hat{\bar{Y}}_{XGT}$ are computed for each variable. The unadjusted simple sample mean $\hat{\bar{Y}}$ from the nonprobability sample is also included. Results from using the common set of covariates which are available in both datasets are presented in Table 6.

Table 6. Estimates of the population mean of the variable measuring the rating (1–10) of the Spanish government action to control the COVID-19 pandemic.

Estimator	Mean	S. Deviation
$\hat{\bar{Y}}$	5.52	0.08
$\hat{\bar{Y}}_{IPSW}$	5.04	0.10
$\hat{\bar{Y}}_{TrIPW}$	5.13	0.09
$\hat{\bar{Y}}_{KW}$	4.95	0.12
$\hat{\bar{Y}}_{SM}$	5.18	0.09
$\hat{\bar{Y}}_{DR}$	5.21	0.09
$\hat{\bar{Y}}_{WT}$	5.38	0.09
$\hat{\bar{Y}}_{XKW}$	5.33	0.72
$\hat{\bar{Y}}_{XGM}$	4.91	0.10
$\hat{\bar{Y}}_{XGD}$	4.92	0.10
$\hat{\bar{Y}}_{XGT}$	4.89	0.09

The results generally show that the application of bias correction techniques provides an important shift (towards a lower mean rate) with respect to the unweighted estimate, especially for those which were the most reliable ones during the simulations ($\hat{\bar{Y}}_{XGM}$, $\hat{\bar{Y}}_{XGD}$ and $\hat{\bar{Y}}_{XGT}$). Standard deviations were estimated via bootstraping [44]. 2000 resamples with replacement are obtained in order to calculate the deviation for each method. They show a small and expectable increase in variance from the unweighted case except for the $\hat{\bar{Y}}_{XKW}$ estimator. As seen in the simulations, this behavior is to be expected and should be solved via hyperparameter tuning.

However, the chosen variable is closely related to the ideological scale covariate. We also apply the methods to estimate the population means of the variables, rating, from 1 to 5, the confidence in the following groups/institutions to manage the current health crisis: health workers, the armed forces, the police, the Spanish government and scientists. The results are presented in Table 7. They show that the differences are not as significant when the target variables are not related to the covariates used.

Table 7. Estimates of the population means of the variables measuring the rating (1–5) of the confidence in different groups/institutions to manage the current health crisis.

Variable	$\hat{\bar{Y}}$	$\hat{\bar{Y}}_{IPSW}$	$\hat{\bar{Y}}_{TrIPW}$	$\hat{\bar{Y}}_{KW}$	$\hat{\bar{Y}}_{SM}$	$\hat{\bar{Y}}_{DR}$	$\hat{\bar{Y}}_{WT}$	$\hat{\bar{Y}}_{XKW}$	$\hat{\bar{Y}}_{XGM}$	$\hat{\bar{Y}}_{XGD}$	$\hat{\bar{Y}}_{XGT}$
Health workers	4.48	4.41	4.45	4.4	4.45	4.43	4.43	4.39	4.44	4.43	4.44
Armed forces	4.01	3.99	4.12	3.99	3.99	3.97	3.92	4.1	4.03	4.03	4.03
Police	4.04	4.05	4.14	4.07	4.05	4.04	4	3.92	4.07	4.07	4.04
Spanish government	2.94	2.7	2.77	2.68	2.76	2.78	2.87	2.55	2.61	2.62	2.62
Scientists	4.18	4.12	4.11	4.1	4.13	4.14	4.18	3.95	4.03	4.03	4.04

6. Conclusions

A long and ongoing literature is concerned with the evaluation of selection bias in web surveys. Propensity scorse and matching estimators based on linear models are the established workhorses in this literature. The emerging literature in statistical learning might help to increase the precision of the estimates obtained by these methods.

Although machine learning methods have many well-documented advantages in prediction and classification, it is not obvious that using them for propensity scores and matching estimation in a nonprobability framework will reduce the bias in the estimation of parameters. In this work we present four different methods to estimate parameters based on the use of an important ML technique, the XGBoots method, to predict the values of the target variable in the probability sample and also to determine the propensities of participating in the nonprobability sample.

Our work contributes to the literature in evaluating the performance of classical and machine learning based PSA estimators, matching estimators as well as other methods of estimation from web survey data that are more innovative.

To be as close as possible to other recent estimation works in nonprobability surveys, we have replicated the experiment carried out by [47]. When comparing results from both simulations, we observe that estimators involving XGBoost provide better results overall in certain non-linear situations in comparison to the case where linear models are used. These results are relevant considering that, in practice, models will rarely be linear. In fact, they will likely be much more complex than the ones considered in this simulation. For this reason, we compare the different estimators in two real datasets. We compared performance of XGBoost to a classical regression approach, with the former providing good results in terms of bias and Mean Square Error reduction.

Our findings are mixed. Our evidence suggests the usage of XGBoost is more powerful at removing selection bias in nonprobability samples than traditional linear regression models in scenarios where the propensity model is not linear and the auxiliary variables used for adjustments are related to both the propensity and the variable of interest. In addition, the simulations also show the efficiency of the use of recent training techniques like [34,39] compared to the alternatives of PSA, matching, and double robust [32] techniques.

However, these results can also be unreliable when the algorithms suffer from overfitting. Hyperparameter optimization has shown to be highly effective at controlling this issue. These kind of procedures are therefore important when producing estimations. We will look further into this matter in future works.

The proposed method is also used to analyze a nonprobability survey sample on the social effects of COVID-19. The results of this application show that selection bias

correction techniques have the potential to provide substantial changes in the estimates of population means in nonprobability samples.

In conclusion, the improved learning capacity of XGBoost is capable of significantly reducing bias and MSE in certain scenarios according to our simulations, but it is important to explore its limits with real use cases. Generally speaking, our results illustrate several methods to do inference with nonprobability samples and highlight the importance and usefulness of auxiliary information from probability survey samples. Propensity Score Adjustment and model-based methods are recommended when the sample can be subject to strong selection bias. XGBoost can yield more accurate predictions when the data behavior is more complex, which typically occurs in situations with high dimensionality. Those are the scenarios where we could particularly benefit the most from Xgboost, although it is suitable for most of the situations.

Author Contributions: Conceptualization, resources and methodology, M.d.M.R.; investigation, L.C.-M., R.F.-G. and M.d.M.R.; data curation, L.C.-M. and R.F.-G.; writing—original draft preparation, M.d.M.R., R.F.-G. and L.C.-M.; writing—review and editing, M.d.M.R., R.F.-G., L.C.-M. and C.H.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Ministerio de Economía y Competitividad of Spain [grantPID2019-106861RB-I00] and IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Institute for Advanced Social Studies (IESA-CSIC) for providing data and information about the ESPACOV survey and the Spanish Center for Sociological Studies (CIS) for providing data and information about the April 2020 barometer survey. The authors want to thank Kenneth C. Chu (Statistics Canada) and Jean-François Beaumont (Statistics Canada) for their assessment on the application of TrIPW algorithm, including the R package to perform the simulations.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Neyman, J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.* **1934**, *97*, 558–625. [CrossRef]
2. Neyman, J. Contribution to the theory of sampling human populations. *J. Am. Stat. Assoc.* **1938**, *33*, 101–116. [CrossRef]
3. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
4. Jiang, D.; Zhao, P.; Tang, N. A propensity score adjustment method for regression models with nonignorable missing covariates. *Comput. Stat. Data Anal.* **2016**, *94*, 98–119. [CrossRef]
5. Lee, S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *J. Off. Stat.* **2006**, *22*, 329.
6. Lee, S.; Valliant, R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **2009**, *37*, 319–343. [CrossRef]
7. Rivers, D. Sampling for web surveys. In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA, 1 August 2007; p. 4.
8. Hsu, H.L.; Chang, Y.C.I.; Chen, R.B. Greedy active learning algorithm for logistic regression models. *Comput. Stat. Data Anal.* **2019**, *129*, 119–134. [CrossRef]
9. Yue, M.; Li, J.; Cheng, M.Y. Two-step sparse boosting for high-dimensional longitudinal data with varying coefficients. *Comput. Stat. Data Anal.* **2019**, *131*, 222–234. [CrossRef]
10. Karatzoglou, A.; Feinerer, I. Kernel-based machine learning for fast text mining in R. *Comput. Stat. Data Anal.* **2010**, *54*, 290–297. [CrossRef]
11. Montanari, G.E.; Ranalli, M.G. Nonparametric model calibration estimation in survey sampling. *J. Am. Stat. Assoc.* **2005**, *100*, 1429–1442. [CrossRef]
12. Baffetta, F.; Fattorini, L.; Franceschi, S.; Corona, P. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* **2009**, *113*, 463–475. [CrossRef]
13. Baffetta, F.; Corona, P.; Fattorini, L. Design-based diagnostics for k-NN estimators of forest resources. *Can. J. For. Res.* **2011**, *41*, 59–72. [CrossRef]

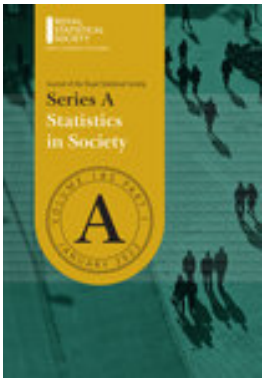
14. Tipton, J.; Opsomer, J.; Moisen, G. Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sens. Environ.* **2013**, *139*, 130–137. [CrossRef]
15. Wang, J.C.; Opsomer, J.D.; Wang, H. Bagging non-differentiable estimators in complex surveys. *Surv. Methodol.* **2014**, *40*, 189–209.
16. Ferri-García, R.; Rueda, M.d.M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE* **2020**, *15*, e0231500. [CrossRef]
17. Buelens, B.; Burger, J.; van den Brakel, J.A. Comparing inference methods for non-probability samples. *Int. Stat. Rev.* **2018**, *86*, 322–343. [CrossRef]
18. Castro-Martín, L.; Rueda, M.d.M.; Ferri-García, R. Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics* **2020**, *8*, 879. [CrossRef]
19. Chu, K.C.K.; Beaumont, J.F. The use uf classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting, Calgary, AB, Canada, 26 May 2019.
20. Kern, C.; Li, Y.; Wang, L. Boosted Kernel Weighting—Using statistical learning to improve inference from nonprobability samples. *J. Surv. Stat. Methodol.* **2020**. [CrossRef]
21. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Lee, B.K.; Lessler, J.; Stuart, E.A. Improving propensity score weighting using machine learning. *Stat. Med.* **2010**, *29*, 337–346. [CrossRef]
23. Lee, B.K.; Lessler, J.; Stuart, E.A. Weight trimming and propensity score weighting. *PLoS ONE* **2011**, *6*, e18174. [CrossRef] [PubMed]
24. McCaffrey, D.F.; Ridgeway, G.; Morral, A.R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **2004**, *9*, 403. [CrossRef]
25. McCaffrey, D.F.; Griffin, B.A.; Almirall, D.; Slaughter, M.E.; Ramchand, R.; Burgette, L.F. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **2013**, *32*, 3388–3414. [CrossRef]
26. Tu, C. Comparison of various machine learning algorithms for estimating generalized propensity score. *J. Stat. Comput. Simul.* **2019**, *89*, 708–719. [CrossRef]
27. Zhu, Y.; Coffman, D.L.; Ghosh, D. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *J. Causal Inference* **2015**, *3*, 25–40. [CrossRef]
28. Couper, M. *Web Survey Methodology: Interface Design, Sampling and Statistical Inference*; Instituto Vasco de Estadística (EUSTAT): Vitoria-Gasteiz, Spain, 2011.
29. Elliott, M.R.; Valliant, R. Inference for nonprobability samples. *Stat. Sci.* **2017**, *32*, 249–264. [CrossRef]
30. Valliant, R. Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **2020**, *8*, 231–263. [CrossRef]
31. Valliant, R.; Dever, J.A. Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **2011**, *40*, 105–137. [CrossRef]
32. Chen, Y.; Li, P.; Wu, C. Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.* **2020**, *115*, 2011–2021. [CrossRef]
33. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. Classification and regression trees. *Biometrics* **1984**, *40*, 358–361.
34. Wang, G.C.; Katki, L. Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *J. R. Stat. Soc.* **2020**, *183*, 1293–1311. [CrossRef]
35. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Routledge: London, UK, 2018.
36. Copas, A.; Burkill, S.; Conrad, F.; Couper, M.P.; Erens, B. An evaluation of whether propensity score adjustment can remove the self-selection bias inherent to web panel surveys addressing sensitive health behaviours. *BMC Med. Res. Methodol.* **2020**, *20*, 1–10. [CrossRef] [PubMed]
37. Beaumont, J.F.; Bissonnette, J. Variance estimation under composite imputation: The methodology behind SEVANI. *Surv. Methodol.* **2011**, *37*, 171–179.
38. Wu, C.; Sitter, R.R. A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.* **2001**, *96*, 185–193. [CrossRef]
39. Castro-Martín, L.; Rueda, M.d.M.; Ferri-García, R. Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *J. Comput. Appl. Math.* **2021**, 113414. [CrossRef]
40. Hawkins, D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12. [CrossRef] [PubMed]
41. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
42. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
43. Särndal, C.E.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2003.
44. Wolter, K.M.; Wolter, K.M. *Introduction to Variance Estimation*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 53.
45. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 2546–2554.
46. Celisse, A. Optimal cross-validation in density estimation with the L^2-loss. *Ann. Stat.* **2014**, *42*, 1879–1910. [CrossRef]

47. Chen, Y. Statistical Analysis with Non-Probability Survey Samples. Doctoral Dissertation, University of Waterloo, Waterloo, ON, Canada, 2020.
48. Antonio, N.; de Almeida, A.; Nunes, L. Hotel booking demand datasets. *Data Brief* **2019**, *22*, 41–49. [CrossRef]
49. Dua, D.; Graff, C. UCI Machine Learning Repository. 2017. Available online: http://archive.ics.uci.edu/ml (accessed on 1 October 2021).
50. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
51. Serrano del Rosal, R.; Biedma Velázquez, L.; Domínguez Álvarez, J.A.; García Rodríguez, M.I.; Lafuente, R.; Sotomayor, R.; Trujillo Carmona, M.; Rinken, S. *Estudio Social sobre la Pandemia del COVID-19 (ESPACOV)*; DIGITAL.CSIC: Madrid, Spain, 2020. [CrossRef]
52. National Institute of Statistics. Resident Population by Date, Sex and Age. Population Figures. 2021. Available online: https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710984 (accessed on 1 October 2021).
53. National Institute of Statistics. Population of 16 Years Old and Over by Educational Level Reached, Sex and Age Group. Economically Active Population Survey. 2021. Available online: https://www.ine.es/jaxiT3/Tabla.htm?t=6347 (accessed on 1 October 2021).
54. Spanish Center for Sociological Research. April Barometer (Study Number 3238). 2020. Available online: http://www.cis.es/cis/opencms/ES/NoticiasNovedades/InfoCIS/2020/Documentacion_3279.html (accessed on 1 October 2021).

Appendix A5

Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

Castro-Martín, Luis; Rueda, María del Mar; Sánchez-Cantalejo, Carmen; Ferri-García, Ramón; Hidalgo Calderón, Jorge; Cabrera, Andrés (2022)
Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.
Submitted to *Journal of the Royal Statistical Society. Series A (Statistics in Society)*



STATISTICS & PROBABILITY			
JCR Year	Impact factor	Rank	Quartile
2020	2.483	28/125	Q1

Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey.

Luis Castro-Martín

Department of Statistics and Operational Research, University of Granada, Granada, Spain.

E-mail: luiscastro193@ugr.es

María del Mar Rueda

Department of Statistics and Operational Research, Math Institute of the University of Granada, University of Granada, Granada, Spain.

E-mail: mrueda@ugr.es

Carmen Sánchez-Cantalejo

Andalusian School of Public Health, Granada, Spain.

E-mail: carmen.sanchezcantalejo.easp@juntadeandalucia.es

Ramón Ferri-García

Department of Statistics and Operational Research, University of Granada, Granada, Spain.

E-mail: rferri@ugr.es

Jorge Hidalgo Calderón

Department of Geometry and Topology, Math Institute of the University of Granada, Granada, Spain.

E-mail: jorgehcal@gmail.com

Andrés Cabrera

Andalusian School of Public Health, Granada, Spain.

E-mail: andres.cabrera.easp@juntadeandalucia.es

Summary. Healthcare statistical services worldwide have used probability surveys to respond to such information needs. The Health Care and Social Survey (ESSOC) research project arises from the need to provide data on the evolution of the COVID-19 impact that can be considered when making decisions to prepare and provide an effective Public Health response in the different affected populations. This survey has an overlapping panel design with 4 measurements throughtout 1 year. The problem of non-response is particularly aggravated in the case of panel surveys, due to the fatigue of the population to be repeatedly surveyed. In this work, we test a new method to reweighting that produces estimators that are suitable for survey data affected by non-response. In each measurement, missing units are substituted by new surveyed units, allowing the obtention of cross-sectional and longitudinal estimates. The weights are the result of two-step process: the original sampling design weights are corrected during a 1st phase by modeling

the non-response with respect to the longitudinal sample obtained in a previous measurement using machine learning techniques. Then, during a 2nd phase, they are calibrated using the auxiliary information available at the population level. The proposed method is applied to the estimation of totals, proportions, differences between measurements as well as gender gaps in the ESSOC.

Keywords: Public health, COVID-19, panel surveys, sampling, machine learning, non-response

1. Introduction

The urgent need to control the expansion rate of COVID-19 requires a quick and efficient assessment of the situation, based on predicting and quantifying the main parameters involved in this phenomenon. Healthcare statistical services worldwide have used probability surveys to respond to information needs concerning the social, economic and health impact of the disease, or on its seroprevalence and evolution or on the characteristics of the infected population, especially those most vulnerable to the virus due to their age, risk of exclusion, health conditions or dependency. These surveys allow valid inferences to be made about the population without having to incorporate hypotheses into the models, which is of great practical benefit.

The Health Care and Social Survey (ESSOC, Encuesta Sanitaria y SOCial) research project arises from the need to provide data on the evolution of the COVID-19 impact that can be considered when making decisions to prepare and provide an effective Public Health response in the different affected populations, especially in the most vulnerable ones, such as, among others, the elderly, the chronically ill, or persons at risk of exclusion (Sánchez-Cantalejo et al., 2021). The objective of this survey is to determine the magnitude, characteristics, and evolution of the impact of COVID-19 on overall health and its socioeconomic, psychosocial, behavioral, occupational, environmental, and clinical determinants in the general population and that with greater socioeconomic vulnerability. The study is based on a Real-World Data design integrating observational data extracted from multiple sources including information obtained from different surveys and clinical, population, and environmental registries. The surveys have an overlapping panel design (Kalton and Citro, 1995) to ensure there are both cross-sectional and longitudinal estimates, and to include population-based probability samples. Thus, the ESSOC is made up of a series of measurements broken down into a new sample and a longitudinal sample for each measurement.

Panel designs are used in practice for studies whose objective is to see the evolution of certain characteristics over time, but have the problem that the lack of response grows with the number of occasions or measurements, due, among others, to the fatigue of the panelist to be repeatedly interviewed. For this reason, partial replacement of units is common to guarantee a minimum number of units in the final sample. Estimation from data obtained with this structure is not easy, especially if one wants to take into account the biases produced both by the lack of response, as well as by the lack of coverage and representativeness of the sample. Some methods of handling wave nonresponse in panels are provided in Kalton et al. (1985), Lepkowski (1989) and Kalton and Brick (1995). Another set of studies focuses on modeling different types of response patterns in panels.

Kern et al. (2019) compare the usage of different Machine Learning (ML) methods for modeling nonresponse in the German Socio-Economic Panel Study (GSOEP) and recently Kern et al. (2021) propose a general framework for building and evaluating nonresponse prediction models with panel data, but this study is focused on model building and evaluation without utilizing the obtained predictions to correct the bias in the estimations.

Nonresponse in panel studies has traditionally been tackled by using nonresponse weights. Although there are reweighting methods to deal with these types of biases, they have been proposed fundamentally for the case of cross-sectional surveys and there are few studies that provide a formal methodology for their treatment in this type of panel. In Rendtel and Harms (2009), the authors discuss adjustments for nonresponse and how calibration can be carried out in panel studies in general and what effects it creates. They consider three possible ways of calibration: initial calibration (at the beginning of the panel, the weights of the units in the panel are calibrated), final calibration (at measurement t the weights of the individuals in the sample are adjusted by calibration) and initial and subsequent final calibration (both, initial as well as final calibration, are carried out). Several approaches are tested in Arcos et al. (2020) to produce calibration estimators that are suitable for survey data affected by non response where auxiliary information exists at both the panel level and the population level. These authors consider non-overlapping panels.

In this work we propose weighting methods for estimating totals, proportions and change or differences of a population characteristic, from overlapping panel survey data, using various combined methods such as Propensity Score Matching, machine learning and calibration. The reweighting methods are formulated based on the ESSOC structure but can be adapted to any other type of overlapping panel design.

The paper is organized as follows. First, in Section 2 we review the estimation in overlapping panels to set the framework and the notation. We present cross-sectional and longitudinal estimators in Sections 3 and 4 and we show how to use machine learning methods to reweighting for non-response based on the data of previous occasions. In Section 5, we apply some of the estimators developed and proposed methods to a specific variable (self-perceived general health) from a real survey about COVID-19: the Health Care and Social Survey. Finally, we highlight the most relevant findings and conclusions in Section 6.

2. Sampling setup in overlapping panels

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. We want to estimate a population parameter of a variable of interest, y.

On the first measurement (t = 1) a sample $s^{(1)}$ of size $n^{(1)}$ is selected from the population U by random stratified sampling. Let h be the stratum to which unit i belongs, $(h = 1, ...L)$ and $s_h^{(1)}$ be the sample corresponding to stratum h on occasion 1.

There is a total lack of response in the sample $s^{(1)}$ which is divided into

$$s_{rh}^{(1)} = \{i \in s^{(1)}/\text{respond in stratum h }\}$$
$$s_{fh}^{(1)} = \{i \in s^{(1)}/\text{missing in stratum h }\},$$

Let $m_h^{(1)}$ denote the number of the observations obtained form the $n_h^{(1)}$ sampled units, that is $\sum_h m_h^{(1)}$ is the size of $s_r^{(1)}$.

In each of the following measurements $t = 2, 3, ..., k$ we denote by $s_{rh}^{(t)}$ the sample of respondents in measurement t in stratum h of the original sample $s^{(1)}$. The size of which we denote by $m_h^{(t)}$. To complete the sample, a new sample $s_{new}^{(t)}$ is selected from the population U by stratified sampling independently of the sample $s^{(1)}$. We verified that the samples $s_{new}^{(t)}$ and the sample $s^{(1)}$ have an empty intersection. Let $n_{hnew}^{(t)}$ be the size of the sample $s_{new}^{(t)}$ in stratum h and denote by $m_{hnew}^{(t)}$ the size of the sample of respondents in this stratum, $s_{rhnew}^{(t)}$. Thus, the total sample of respondents in each stratum and measurement would be $m_{htotal}^{(t)} = m_h^{(t)} + m_{hnew}^{(t)}$.

Let $y_i^{(t)}$ be the value of the target variable associated to the i-th unit in measurement t, and let d_i be the design weight associated to the i-th unit equal to the inverse of the inclusion probability in the initial sample, an estimation of the total of Y in the first occasion is given by:

$$\hat{Y}_{ht}^{(1)} = \sum_h \sum_{i \in s_{rh}^{(1)}} d_{ih} y_{ih}^{(1)} \tag{1}$$

This estimator is called the Horvitz-Thompson (H-T) estimator. In the case of stratified simple random sampling design for unit i belonging to stratum h is $d_{ih} = \frac{N_h}{n_h^{(1)}}$.

Design weights should be adjusted to consider non-response in order to reduce the possible bias of resulting estimates, which may arise when there is a different propensity in answering for different groups. In the first occasion a response rate is determined in each class and a new weight is defined as the product of the design weight and the inverse of the response rate. The response rate in stratum h is evaluated as $r_h = \frac{m_h^{(1)}}{n_h^{(1)}}$. Then the initial weight of unit i in stratum h d_{ih} is replaced with the new weight $d_{ih}^{(1)} = \frac{d_{ih}}{r_h}$ and the estimator is given by

$$\hat{Y}^{(1)} = \sum_h \sum_{i \in s_{rh}^{(1)}} d_{ih}^{(1)} y_{ih}^{(1)} \tag{2}$$

For the following measurements, cross-sectional and longitudinal estimators can be obtained from the new sample obtained in each measurement and from the longitudinal samples of the previous measurements. The process to obtain them is shown below.

3. Cross-sectional estimation

The objective of most cross-sectional surveys is to produce unbiased estimates of totals or means at a given time point, and, in the case of repeated surveys, to produce estimates of the net change that occurred in the population between two time points.

In order to improve the cost-effectiveness of surveys on can derive cross-sectional estimates from longitudinal survey data assuming that the survey design takes this

possibility into account, and that estimation procedures are developed to satisfy cross-sectional as well as longitudinal requirements.

Point estimation of parameters of the cross-sectional population based on data from longitudinal surveys has been studied by Lavallee (1995) among others and the problem of formal comparison of the estimates from two years, which requires variance estimation for the difference of the estimates, is considered in Kovacevic (2001). We will follow a methodology similar to that used in these works. We will elaborate a cross-sectional weighting scheme that includes a non-response adjustment, an optimal combination of the two samples, and a calibration for completing representativeness of the population at a given time. This proposal is described below.

First, we adjust the basic weights of the H-T estimator by the fraction of non-response $\frac{n_{hnew}^{(t)}}{m_{hnew}^{(t)}}$ obtaining the total estimator for the new sample in measurement t:

$$\hat{Y}_n^{(t)} = \sum_h \sum_{s_{rhnew}^{(t)}} \frac{N_h}{n_{hnew}^{(t)}} \frac{n_{hnew}^{(t)}}{m_{hnew}^{(t)}} y_{ih}^{(t)} = \sum_h \sum_{s_{rhnew}^{(t)}} d_{ihn}^{(t)} y_{ih}^{(t)} \tag{3}$$

In a similar way, from the sample $s_r^{(t)}$ we can estimate the total as:

$$\hat{Y}_r^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} \frac{N_h}{n_h^{(1)}} \frac{n_h^{(1)}}{m_h^{(t)}} y_{ih}^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ihr}^{(t)} y_{ih}^{(t)} \tag{4}$$

Combining these estimators we considered the following estimator

$$\hat{Y}^{(t)} = \alpha_1 \hat{Y}_r^{(t)} + \alpha_2 \hat{Y}_n^{(t)} \tag{5}$$

where α_1 and α_2 are nonnegative constants such that $\alpha_1 + \alpha_2 = 1$.

Next, we consider the problem of selection of the best coefficients.

We denote the values $V(\hat{Y}_r^{(t)})$, $V(\hat{Y}_n^{(t)})$ by V_1, V_2 respectively. Thus, the variance of $\hat{Y}^{(t)}$ is:

$$V(\hat{Y}^{(t)}) = \alpha_1^2 V_1 + (1-\alpha_1)^2 V_2$$

$$= \Big[V_1 + V_2\Big] \cdot \left[\alpha_1^2 - \frac{2\alpha_1 V_2}{V_1 + V_2}\right] + V_2 =$$

$$= \Big[V_1 + V_2\Big] \cdot \left[\alpha_1 - \frac{V_2}{V_1 + V_2}\right]^2 + \frac{V_2 \cdot V_1}{V_1 + V_2} \geq$$

$$\geq \frac{V_2 \cdot V_1}{V_1 + V_2} = V_{min}(\hat{Y}^{(t)}) \tag{6}$$

Because $V_1 + V_2 \geq 0$, equality holds if and only if

$$\alpha_1 = 1 - \alpha_2 = \frac{V_2}{V_1 + V_2} \tag{7}$$

But the values V_1 and V_2 are unknown, one possibility is to estimate them from the sample and substitute them in the previous expression to calculate the coefficients α but that does not ensure their optimality. A simple solution is to weight each estimator by the weight that sample has in the total sample available at the time t. In this way we consider the self-weighted total estimator

$$\hat{Y}_{sw}^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} \frac{m_h^{(t)}}{m_h^{(t)} + m_{hnew}^{(t)}} \frac{N_h}{m_h^{(t)}} y_{ih}^{(t)} + \sum_h \sum_{i \in s_{rhnew}^{(t)}} \frac{m_{hnew}^{(t)}}{m_h^{(t)} + m_{hnew}^{(t)}} \frac{N_h}{m_{hnew}^{(t)}} y_{ih}^{(t)}$$

$$= \sum_h \frac{N_h}{m_h^{(t)} + m_{hnew}^{(t)}} \left(\sum_{i \in s_{rh}^{(t)}} y_{ih}^{(t)} + \sum_{i \in s_{rhnew}^{(t)}} y_{ih}^{(t)} \right) = \sum_h \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} d_{ich}^{(t)} y_{ih}^{(t)} \tag{8}$$

The weights $d_{ich}^{(t)}$ are the same for all units included in stratum h, so the sample within each stratum is self-weighted.

Besides the modification of weights for handling non-response, weights adjustment may also be carried out to take into account of auxiliary information. Calibration Deville and Särndal (1992) is the most used technique for weights adjustment and can have the aim to insure consistency among estimates of different sample surveys and some improvement in the precision of estimators may be achieved (Rueda et al. (2006), Kott and Liao (2015), Cabrera-León et al. (2015), Devaud and Tillé (2019)).

Let $\mathbf{x}^{*(t)}$ be a set of auxiliary variables related to y such that their population totals at the stratum level are known at measurement t, $\mathbf{X}_h^{*(t)} = \sum_{\mathcal{U}_h} \mathbf{x}_{kh}^{*(t)}$.

The calibration total estimator is obtained as:

$$\hat{Y}_{\text{CAL}}^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} y_{ih}^{(t)} \tag{9}$$

where the weights $w_{ih}^{(t)}$, are as close as possible, with respect to a given distance G, to the weights $d_{ich}^{(t)}$:

$$\min_{\omega_k} \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} G\left(w_{ih}^{(t)}, d_{ich}^{(t)}\right) \tag{10}$$

fulfilling the calibration condition

$$\sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^t \mathbf{x}_{ih}^{*(t)} = \sum_{\mathcal{U}_h} \mathbf{x}_{ih}^{*(t)} \tag{11}$$

for all stratum h.

A parameter of interest is the absolute change from one measurement to the first measurement of the variable and we denote by $\theta^{(t)} = Y^{(t)} - Y^{(1)}$ this parameter. Variations over time are measured more accurately with overlapping samples with respect to the case where samples on different occasions do not overlap (see Särndal et al. (2003)). An estimator of this parameter for measurement t based on the previous calibration total estimators can be obtained as follows:

$$\hat{\theta}_{abs}^{(t)} = \hat{Y}_{\text{CAL}}^{(t)} - \hat{Y}_{\text{CAL}}^{(1)} \tag{12}$$

Other parameter of interest in panel surveys is the relative change $\theta_{rel}^{(t)} = \frac{Y^{(t)} - Y^{(1)}}{Y^{(1)}}$ between measurement 1 and measurement t, which is estimated as:

$$\hat{\theta}_{rel}^{(t)} = \frac{\hat{\theta}^{(t)}}{\hat{Y}^{(1)}} \tag{13}$$

The estimator is a quotient of two estimators of the total based on two different samples, meaning that its properties are not equivalent to those of the ratio estimator commonly used in survey sampling, but its theoretical properties can be derived by using Taylor linear approximation.

The impact of the COVID-19 in the social determinants of health might have been widely different between genders. For this reason, it is of great interest to define the estimators of the gender gaps observed in the absolute and relative changes defined in previous sections, both in absolute and relative terms as well, in order to observe if the changes were significantly larger among people of a given gender in comparison to their counterpart.

Let $Gen = \{M, W\}$ be the variable measured in $s^{(t)}, t = 1, 2, 3, ..., k$ which reflects whether a respondent is a man (M) or a woman (W). We define the two indicator variables: $I_{ih}^{M} = 1$ if the unit i in stratum h is a man and 0 elsewhere, and I_{ih}^{W} in a similar way.

We start by defining the absolute gender gap estimator in the absolute change as follows:

$$\begin{gathered} \hat{GGabs}_{abs}^{(t)} = \hat{\theta}_{W}^{(t)} - \hat{\theta}_{M}^{(t)} = \\ = (\textstyle\sum_h \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} y_{ih}^{(t)} I_{ih}^{W} - \sum_h \sum_{i \in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^{W}) - \\ (\textstyle\sum_h \sum_{i \in s_{rh}^{(t)} \cup s_{rhnew}^{(t)}} w_{ih}^{(t)} y_{ih}^{(t)} I_{ih}^{M} - \sum_h \sum_{i \in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^{M}) \end{gathered} \tag{14}$$

The estimator $\hat{GGabs}_{abs}^{(t)}$ is defined as the linear combination of two estimators in certain domains, hence its theoretical properties can be easily derived (see section 5.4 in Särndal et al. (2003). This estimator is the most simple one can build on the gender gap and can tell the difference in the absolute change between men and women between measurement t and measurement 1. However, this estimator is subject to the base rate on each variable. For this reason, we define the relative gender gap estimator in the absolute change as follows:

$$\hat{GGabs}_{rel}^{(t)} = \frac{\hat{GGabs}_{abs}^{(t)}}{\hat{\theta}_{M}^{(t)}} = \frac{\hat{\theta}_{W}^{(t)} - \hat{\theta}_{M}^{(t)}}{\hat{\theta}_{M}^{(t)}} \tag{15}$$

The estimator $G\hat{G}abs_{rel}^{(t)}$ allows us to observe the gender gap in the growth between measurement 1 and measurement t taking into account the base rate of the given target variable, but its theoretical properties are more difficult to develop as it is a nonlinear combination of two estimators from non-overlapping samples.

We define the absolute gender gap in the relative change as follows:

$$G\hat{G}rel_{abs}^{(t)} = \hat{\theta}_{relW}^{(t)} - \hat{\theta}_{relM}^{(t)} = \frac{\hat{\theta}_{W}^{(t)}}{\sum_h \sum_{i \in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^{W}} - \frac{\hat{\theta}_{M}^{(t)}}{\sum_h \sum_{i \in s_{rh}^{(1)}} w_{ih}^{(1)} y_{ih}^{(1)} I_{ih}^{M}} \tag{16}$$

The estimator $G\hat{G}rel_{abs}^{(t)}$ allows us to observe the difference in percentage points in the relative growth of a given variable between women and men. We define the relative gender gap in the relative change as follows:

$$G\hat{G}rel_{rel}^{(t)} = \frac{G\hat{G}rel_{abs}^{(t)}}{\hat{\theta}_{relM}^{(t)}} \tag{17}$$

Thus, for the study variables of each ESSOC measurement, we start from the H-T estimator (1) that is adjusted for non-response (4), combined from the cross-sectional and longitudinal samples (8) and, finally, calibrated to increase the representativeness of the sample (9). This estimator serves as the basis for calculating the absolute (12) and relative (13) change estimators between measurement t and 1, which are also used to obtain the different estimators to measure the absolute and relative gender gap in the absolute and relative changes of a measurement with respect to the first (14 and 15, and 16 and 17, respectively).

4. Longitudinal estimation

The primary objective of panel surveys is the production of longitudinal data series that are appropriate for studying the gross change in the population between collection dates, and for research on causal relationships among variables.

Definition of weights for each measurement may require specific computation when using panel surveys. Evaluation of weights for each occasion of the survey follows the standard steps: determination of a design weight equal to the inverse of the inclusion probability and subsequent adjustment for non-response and for improving estimators. We are going to detail these reweighting procedures below.

After the first measurement, determination of weights for the following measurements should take attrition into account. Then, at each subsequent measurement the first operation should consist in adjusting the weights for non-response due to attrition. In the sample $s_r^{(t)}$ we have the values of the variables in the previous measurements and we can use these values to model the lack of response and adjust the weights in a more efficient way. For this we are going to use the popular Propensity Score Adjustment (PSA) method (Rosenbaum and Rubin (1983),Ferri-García and Rueda (2018), Ferri-García and Rueda (2020)) to model the probability that a unit of the sample $s_r^{(1)}$ responds on occasion t.

For each sample unit $s_r^{(1)}$ let be $\delta_k^{(t)} = 1$ if $k \in s_r^{(t)}$ and $\delta_k^{(t)} = 0$ if $k \in s_r^{(1)} - s_r^{(t)}$. We assume that the selection mechanism of response is ignorable, this is:

$$\pi_k^{(t)} = P(\delta_k^{(t)} = i|y_k, \mathbf{x}_k) = P(\delta_k = i|\mathbf{x}_k), i = 0, 1; k \in s_r^{(t)} \tag{18}$$

We also assume that the mechanism follows a parametric model:

$$P(\delta_k^{(t)} = 1|y_k, \mathbf{x}_k) = f_t(\mathbf{x}_k) \tag{19}$$

We use a state-of-the-art machine learning method: XGBoost Chen and Guestrin (2016) for estimating $\pi_k^{(t)}$. This technique builds decision trees ensembles which optimize an objective function via Gradient Tree Bosting Friedman et al. (2000). Kern et al. (2019) has shown the effectiveness of this technique when studying nonresponse in the GSOEP panel. Ferri-García and Rueda (2020) showed that Gradient Tree Bosting can lead to selection bias reductions in situations of high dimensionality or where the selection mechanism is Missing At Random (MAR). Lee et al. (2010, 2011); McCaffrey et al. (2004, 2013); Tu (2019); Zhu et al. (2015) have applied boosting algorithms in propensity score weighting showing better results than conventional parametric models.

In order to obtain the estimated propensities $\hat{\pi}_k^{(t)}$, we train a model with $s_r^{(1)}$ where $\mathbf{x}_k$ includes every available variable observed in $s_r^{(1)}$. Said model minimizes the logistic loss for $\delta_k^{(t)}; k \in s_r^{(1)}$. This logistic loss is measured as:

$$l(\hat{\pi}^{(t)}) = \sum_{k \in s_r^{(1)}} -\delta_k^{(t)} log(\hat{\pi}_k^{(t)}) - (1 - \delta_k^{(t)}) log(1 - \hat{\pi}_k^{(t)})$$

Since the values we are interested in, $\hat{\pi}_k^{(t)}$ for $k \in s_r^{(1)} \cap s_r^{(t)}$, are a subset of the values used for training, $\delta_k^{(t)}$ for $k \in s_r^{(1)}$, overfitting is likely to happen. This means we will obtain values extremely close to 1 instead of real propensities. Hyperparameter optimization is essential in order to avoid this issue.

The estimated propensities for each unit i of sample $s_{rh}^{(t)}$, $\hat{\pi}_{ih}^{(t)}$, are used to reweighting for nonresponse, and we define an estimator for $\theta^{(t)}$ from the sample of respondents on occasion t by:

$$\hat{\theta}_l^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ih}^{(1)} \frac{1}{\hat{\pi}_{ih}^{(t)}} (y_{ih}^{(t)} - y_{ih}^{(1)}) = \sum_h \sum_{i \in s_{rh}^{(t)}} d_{ihPSA}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \tag{20}$$

As population is subject to changes, it is important to modify weights to reflect these changes, as well. If updated totals are available then calibration to new totals can reduce presence of bias. Thus, in the next phase, calibration is applied to change the weights. So we get some weights $v_{ih}^{(t)}$, minimizing:

$$\sum_{i \in s_{rh}^{(t)}} G\left(v_{ih}^{(t)}, d_{ihPSA}^{(t)}\right) \tag{21}$$

subject to

$$\sum_{i \in s_{rh}^{(t)}} v_{ih}^{t} \mathbf{x}_{ih}^{*(t)} = \sum_{\mathcal{U}_h} \mathbf{x}_{ih}^{*(t)} \tag{22}$$

for all stratum h. The final calibrate estimator is given by

$$\hat{\theta}_c^{(t)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t)} (y_{ih}^{(t)} - y_{ih}^{(1)}) \tag{23}$$

The researcher may also be interested in estimating the change from occasion t to the occasion $t-1$, $\theta^{(t,t-1)} = Y^{(t)} - Y^{(t-1)}$. In this situation the estimator is calculated in the same way but modeling the non-response with respect to the sample obtained in the previous occasion, that is, we estimate the new propensities:

$$P(\delta_k^{(t,t-1)} = 1 | y_k, \mathbf{x}_k) = g_t(\mathbf{x}_k) \tag{24}$$

being $\delta_k^{(t,t-1)} = 1$ if $k \in s_r^{(t)}$ and $\delta_k^{(t)} = 0$ if $k \in s_r^{(t-1)} - s_r^{(t)}$.

The estimated propensities for each unit i of sample $s_{rh}^{(t)}$, $\hat{\pi}_{ih}^{(t,t-1)}$, are used in the first stage to reweighting for adjusting the nonresponse, and in the second stage, calibration is applied to reweight these weights and obtain new ones, $v_{ih}^{(t,t-1)}$, so as to obtain better representativeness of the population. The longitudinal estimator of $\theta^{(t,t-1)} = Y^{(t)} - Y^{(t-1)}$ can be defined as follows:

$$\hat{\theta}_c^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) \tag{25}$$

The longitudinal nature of the estimator allows us to define new estimators on the number of population individuals whose value of y increases, decreases or remains the same between $t-1$ and t. Let A be a subset of interest ($\mathbb{R}^+$, $\mathbb{R}^-$ or 0 if we are interested in the units whose value of y increases, decreases or remains the same respectively); the estimator of the number of population individuals for which $y^{(t)} - y^{(t-1)} \in A$ can be estimated as follows:

$$\hat{\theta}_{cA}^{(t,t-1)} = \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A, I_A = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \in A \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \notin A \end{cases} \tag{26}$$

We can also obtain the estimator of the rate of people whose value in y has decreased between $t-1$ and t, in reference to the people whose value in y has increased between $t-1$ and t. If the variable y measures health status, this rate can be considered a deterioration/improvement rate, $DIRate$. The formula can be defined as follows:

$$\widehat{DIRate}_c^{(t,t-1)} = \frac{\hat{\theta}_{cA_{R^-}}^{(t,t-1)} - \hat{\theta}_{cA_{R^+}}^{(t,t-1)}}{\hat{\theta}_{cA_{R^+}}^{(t,t-1)}} = \frac{\sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R^-}} - \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R^+}}}{\sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_{A_{R^+}}}, \tag{27}$$

where

$$I_{A_{R^+}} = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} > 0 \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \leq 0 \end{cases}$$

and

$$I_{A_{R^-}} = \begin{cases} 1 & y_{ih}^{(t)} - y_{ih}^{(t-1)} < 0 \\ 0 & y_{ih}^{(t)} - y_{ih}^{(t-1)} \geq 0 \end{cases}$$

Based on previous estimators, we can define estimators of the gender gap of the change between $t-1$ and t can be defined as follows:

$$\hat{GGlong}_{abs}^{(t)} = \hat{\theta}_{cW}^{(t,t-1)} - \hat{\theta}_{cM}^{(t,t-1)} = \sum_h \sum_{1 \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) I_{ih}^W - \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} (y_{ih}^{(t)} - y_{ih}^{(t-1)}) I_{ih}^M \tag{28}$$

$$\hat{GGlong}_{absA}^{(t)} = \hat{\theta}_{cAW}^{(t,t-1)} - \hat{\theta}_{cAM}^{(t,t-1)} = \sum_h \sum_{1 \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A I_{ih}^W - \sum_h \sum_h \sum_{i \in s_{rh}^{(t)}} v_{ih}^{(t,t-1)} I_A I_{ih}^M \tag{29}$$

$$\hat{GGlong}_{rel}^{(t)} = \frac{\hat{GGlong}_{abs}^{(t)}}{\hat{\theta}_{cM}^{(t,t-1)}} \tag{30}$$

$$\hat{GGlong}_{relA}^{(t)} = \frac{\hat{GGlong}_{absA}^{(t)}}{\hat{\theta}_{cAM}^{(t,t-1)}} \tag{31}$$

Several techniques can be used for obtaining variance estimates of these proposed estimators. The variance estimation problem in longitudinal surveys is addressed in Kovacevic (2001) for Canada's Survey of Labour and Income Dynamics within a Taylor linearization approach and a bootstrap method (named pseudo-ordinated bootstrap method). We have used the first of these techniques for calculating the errors of the estimates in the ESSOC survey described of the next section.

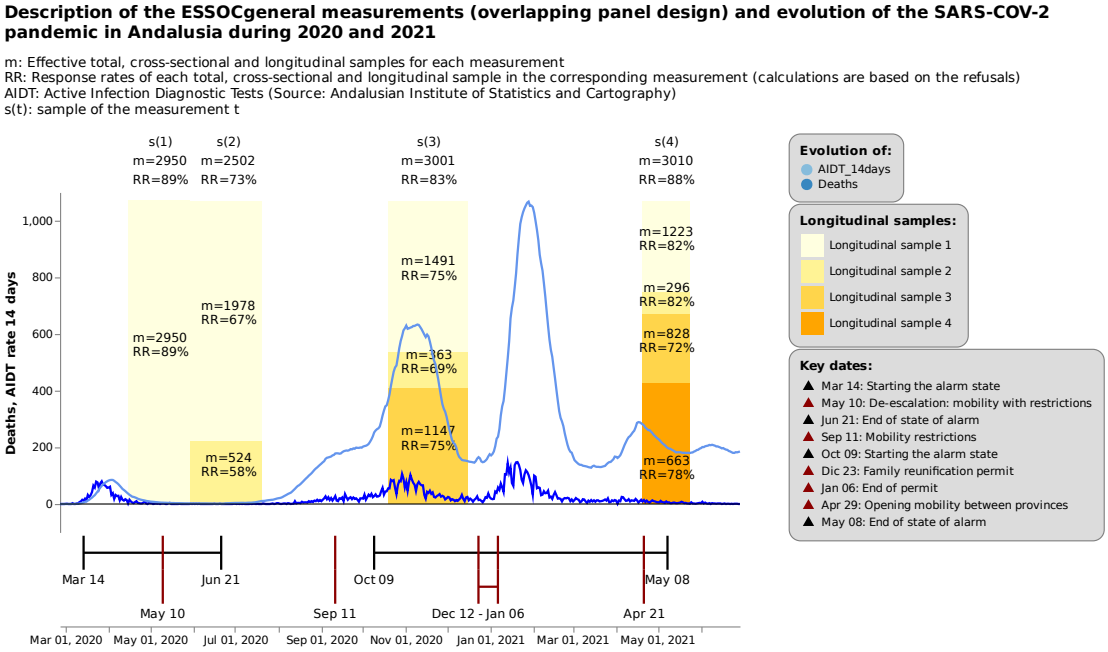
5. Application to the Health Care and Social Survey (ESSOC)

5.1. The ESSOC Study Framework

In this section, we apply the proposed methods to a real survey about COVID-19: the Health Care and Social Survey (ESSOC, Encuesta Sanitaria y SOCial). It provides a follow-up over time on the impact of the pandemic, and its resulting lockdown, on the population of Andalusia (Spain) over the age of 16.

As shown in Figure 1, the ESSOC study includes four measurements. The first one, $s^{(1)}$, coincides with the beginning of the Spanish State of Alarm on April 2020, while the 2nd measurement $s^{(2)}$ was carried out in June and July (a month after the 1st interview, coinciding with the de-escalation), the 3rd measurement $s^{(3)}$ in November and December (6 months after the 1st interview and coinciding with the 2nd wave of the pandemic), and the 4th measurement $s^{(4)}$ in April and May 2021 (12 months after the

Fig. 1. Temporal scope, response rates (RR) and effective sample size for each measurement in ESSOC

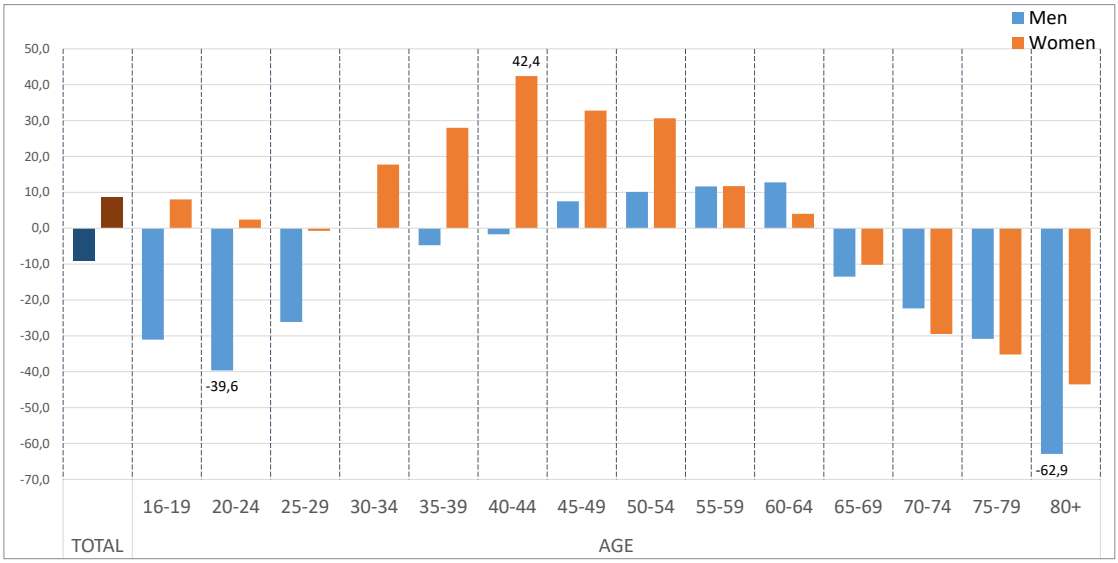


1st interview, coinciding with the opening of mobility and the end of the state of alarm). All the theoretical samples have a size of 5000 people. They were obtained using an overlapping panel design so the individuals from the previous measurement are sampled again. However, the non-response is compensated with another sample including new individuals. The details of this non-response and the effective sample size for each measurement can be consulted at Figure 1. That figure also provides a description of the evolution of the SARS-COV-2 pandemic in Andalusia during 2020 and 2021 in terms of active infection diagnostic tests and deaths. The sampling method is stratified by province and degree of urbanization. A detailed description of the protocol followed for this survey can be seen in Sánchez-Cantalejo et al. (2021).

5.2. Calibrating the representativeness of the sample

In relation to the observed bias, Figures 2 and 3 depict the differences between the sample and the population at measurement 4 of the ESSOC according to the cross of the sex variable with age, province, degree of urbanization and nationality. Thus, with respect to age, the largest differences between the values observed from the sample and those from the population are found in youngest men (under 30 years old), in middle-aged women (between 35 and 54 years old) and in oldest women and men (over 70 years old), with higher differences as age increases. As for the other segmentation variables, the largest differences were found among people with a nationality other than Spanish, especially among men. These results show a lower participation of these population

Fig. 2. Observed biases for the calibration variables in measurement 4 (age and sex)



groups in the ESSOC and therefore justify the need to adjust the sample weights. Thus, the weights are calibrated using truncated linear calibration with 0.1 to 10 limits and the total population size for the cross of the sex variable with province, age, urbanization grades and nationality as auxiliary information. The data for said totals are obtained from the 2020 Municipal Register of Inhabitants (Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym), 2020).

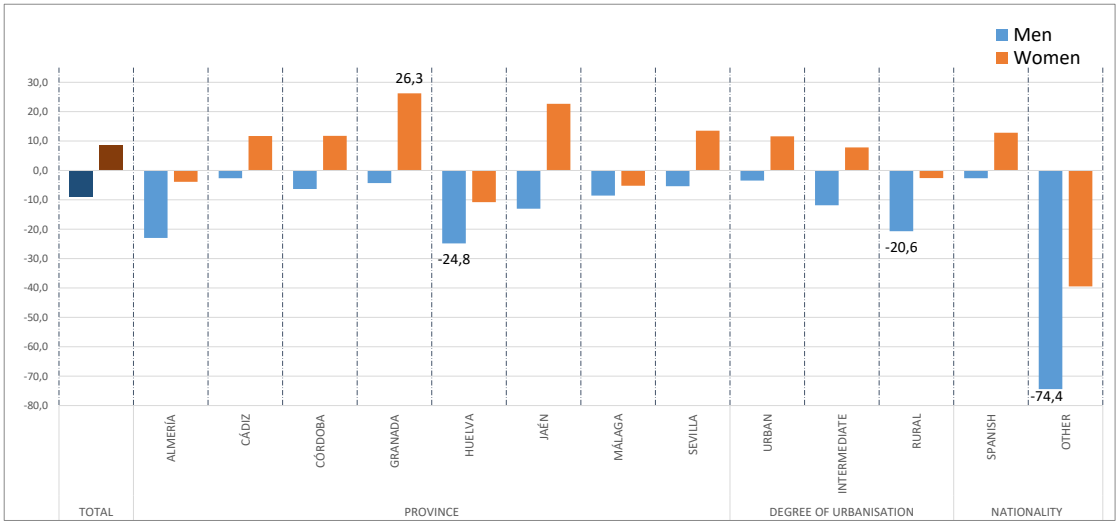
5.3. Modeling the non-response

The non-response at $s_r^{(t)}$, $\pi_k^{(t)}$ for $k \in s_r^{(t)}$, is modeled with PSA considering every variable of $s_r^{(1)}$. In order to ensure that the XGBoost model is learning properly, we have considered the following hyperparameters:

- Number of estimators $\in [10, 400]$: The number of trees forming the ensemble.
- Learning rate $\in [0.01, 1]$: the weight shrinkage applied after each boosting step.
- Maximum depth $\in [1, 60]$: The maximum number of splits that each tree can contain.
- Minimum child weight $\in [1, 6]$: The minimum total of instance weights needed to consider a new partition.

The accuracy of the algorithm is tested with cross-validation. Therefore, training data is partitioned into 5 complementary subsets so that each one has the same proportion of $\delta_k^{(t)} = 1$ and $\delta_k^{(t)} = 0$ as the total. Then 5 models are trained leaving each one of the

Fig. 3. Observed biases for the calibration variables in measurement 4 (sex-province, sex-urbanization and sex-nationality)



subsets out of the training data. For each model, the logistic loss is calculated for its corresponding remaining subset. The mean logistic loss is the estimated error.

The values for the hyperparameters minimizing this estimated error are obtained using the Tree-structured Parzen Estimator (TPE) algorithm Bergstra et al. (2011) Bergstra et al. (2013). TPE is implemented in Optuna Akiba et al. (2019), an optimization library for Python, as its default method.

5.4. Cross-sectional results

Table 1 shows, for measurement 4, the percentages with corresponding confidence intervals at 95% as well as the sample size and population estimations for each original category of the self-perceived general health variable grouped by sex and age. It may be observed from the chart that the percentages for the 'excellent' or 'very good' categories do not follow a clear pattern throughout measurements for the population between 16 and 34 years old, neither for men nor for women. However, the excellent or very good self-perceived health decreases for the population older than 35 years as the pandemic advances. This can be observed more as age increases, especially in women. This reduction results in an increment for the 'fair' and 'bad' categories. However, the 'good' general health category stays stable throughout the pandemic for each sex and age group.

Based on these results, we dichotomized that variable with the categories 'excellent and very good' and 'good, fair and bad'. Figure 4 shows the percentages and confidence intervals given in table 1 not only for measurement 4, but also for all other ESSOC measurements. Table 2 shows, for each measurement of the ESSOC, the percentages and confidence intervals at 95% of the dichotomized self-perceived general health variable as described in the previous paragraph. These results can be seen at figure 5, where it can be observed that the excellent and very good self-perceived health decreased in mea-

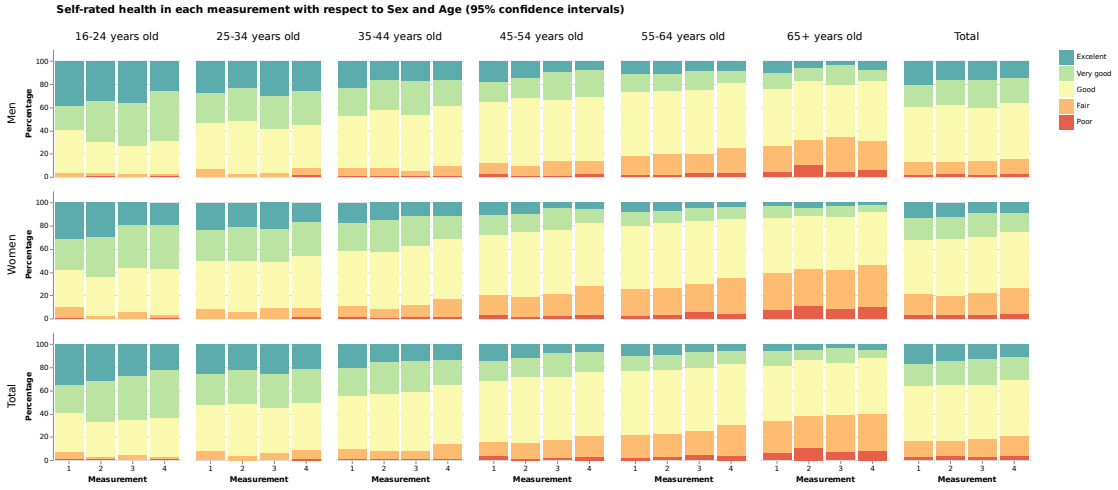
Table 1. Estimations grouped by sex and age for the original categories of self-perceived general health at measurement 4

Age group (years)	Self-perceived general	Total					Men					Women				
		sample size	Population	Percentage	Confidence Interval 95% lower limit	upper limit	sample size	Population	Percentage	Confidence Interval 95% lower limit	upper limit	sample size	Population	Percentage	Confidence Interval 95% lower limit	upper limit
Total	Excelent	339	815847	11,6%	10,3%	12,9%	184	503835	14,6%	12,4%	16,8%	155	312012	8,7%	7,3%	10,0%
	Very good	586	1377920	19,6%	18 %	21,1%	287	758203	22,0%	19,5%	24,5%	299	619717	17,2%	15,3%	19,1%
	Good	1499	3392315	48,1%	46,2%	50,1%	671	1649144	47,8%	44,8%	50,8%	828	1743171	48,4%	45,9%	51,0%
	Fair	501	1246884	17,7%	16,1%	19,3%	166	459783	13,3%	11,0%	15,7%	335	787101	21,9%	19,7%	24,0%
	Bad	80	216887	3,1%	2,4%	3,8%	31	78536	2,3%	1,4%	3,1%	49	138351	3,8%	2,7%	5,0%
	Total	3005	7049853	100 %			1339	3449501	100 %			1666	3600351	100 %		
16-24	Excelent	65	184634	22,9%	17,7%	28,0%	29	107514	25,8%	17,7%	34,0%	36	77119	19,7%	13,7%	25,7%
	Very good	117	329657	40,8%	34,9%	46,8%	51	182659	43,9%	34,7%	53,0%	66	146998	37,6%	30,1%	45,0%
	Good	99	273063	33,8%	28,1%	39,5%	31	116424	28,0%	19,6%	36,3%	68	156640	40,0%	32,4%	47,7%
	Fair	6	15088	1,9%	0,3%	3,4%	2	6615	1,6%	-0,6%	3,8%	4	8473	2,2%	0,1%	4,3%
	Bad	2	5325	0,7%	-0,3%	1,6%	1	3264	0,8%	-0,8%	2,3%	1	2061	0,5%	-0,5%	1,6%
	Total	289	807767	100 %			114	416475	100 %			175	391292	100 %		
25-34	Excelent	91	217937	21,7%	17,3%	26,1%	49	132814	26,0%	18,8%	33,2%	42	85123	17,2%	12,3%	22,1%
	Very good	116	295926	29,4%	24,6%	34,3%	53	151772	29,7%	22,3%	37,1%	63	144154	29,1%	22,8%	35,4%
	Good	177	405789	40,3%	35,3%	45,4%	71	185695	36,4%	28,7%	44,0%	106	220094	44,4%	37,8%	51,1%
	Fair	33	76831	7,6%	5,0%	10,3%	15	35201	6,9%	3,4%	10,4%	18	41630	8,4%	4,5%	12,3%
	Bad	3	9508	1,0%	-0,2%	2,1%	2	5201	1,0%	-0,4%	2,5%	1	4307	0,9%	-0,8%	2,6%
	Total	420	1005991	100 %			190	510684	100 %			230	495307	100 %		
35-44	Excelent	85	182522	13,8%	10,7%	16,9%	42	107701	16,2%	11,0%	21,3%	43	74821	11,4%	8,1%	14,8%
	Very good	145	285541	21,6%	18,2%	25,1%	66	153681	23,1%	17,6%	28,6%	79	131860	20,2%	16,1%	24,2%
	Good	338	677020	51,3%	47,0%	55,6%	142	339453	51,0%	44,3%	57,7%	196	337567	51,6%	46,5%	56,8%
	Fair	79	164400	12,5%	9,6%	15,3%	24	62944	9,5%	5,2%	13,7%	55	101456	15,5%	11,6%	19,4%
	Bad	6	10379	0,8%	0,2%	1,4%	1	2142	0,3%	-0,3%	1,0%	5	8237	1,3%	0,2%	2,4%
	Total	653	1319862	100 %			275	665920	100 %			378	653942	100 %		
45-54	Excelent	48	91677	6,8%	4,9%	8,8%	28	55012	8,2%	5,2%	11,2%	20	36666	5,5%	3,1%	7,9%
	Very good	117	239303	17,9%	14,7%	21,1%	70	156650	23,4%	18,0%	28,7%	47	82653	12,3%	9,0%	15,7%
	Good	377	731205	54,6%	50,6%	58,5%	170	367600	54,8%	48,7%	60,9%	207	363605	54,3%	49,1%	59,4%
	Fair	126	242342	18,1%	15,1%	21,1%	35	75747	11,3%	7,5%	15,1%	91	166595	24,9%	20,4%	29,4%
	Bad	19	35966	2,7%	1,5%	3,9%	8	15673	2,3%	0,7%	4,0%	11	20293	3,0%	1,3%	4,8%
	Total	687	1340494	100 %			311	670682	100 %			376	669812	100 %		
55-64	Excelent	30	67275	6,1%	3,8%	8,3%	21	46155	8,5%	4,7%	12,3%	9	21120	3,7%	1,2%	6,2%
	Very good	58	121195	10,9%	8,2%	13,6%	29	57984	10,7%	6,9%	14,4%	29	63211	11,1%	7,2%	15,1%
	Good	281	590830	53,1%	48,7%	57,6%	149	305410	56,2%	49,8%	62,5%	132	285419	50,2%	44,0%	56,4%
	Fair	134	293014	26,4%	22,4%	30,3%	52	116069	21,3%	15,7%	27,0%	82	176946	31,1%	25,5%	36,8%
	Bad	19	39865	3,6%	2,0%	5,2%	9	18328	3,4%	1,2%	5,6%	10	21537	3,8%	1,5%	6,1%
	Total	522	1112179	100 %			260	543946	100 %			262	568233	100 %		
>=65	Excelent	20	71801	4,9%	2,6%	7,2%	15	54638	8,5%	3,9%	13,2%	5	17163	2,1%	0,3%	3,9%
	Very good	33	106299	7,3%	4,8%	9,8%	18	55458	8,6%	4,6%	12,7%	15	50841	6,2%	3,0%	9,4%
	Good	227	714407	48,8%	43,7%	53,9%	108	334561	52,1%	43,8%	60,5%	119	379846	46,2%	39,8%	52,7%
	Fair	123	455208	31,1%	26,0%	36,2%	38	163207	25,4%	16,8%	34,1%	85	292001	35,5%	29,3%	41,7%
	Bad	31	115845	7,9%	5,1%	10,8%	10	33929	5,3%	2,0%	8,6%	21	81915	10,0%	5,7%	14,3%
	Total	434	1463560	100 %			189	641794	100 %			245	821766	100 %		

Table 2. Evolution in each measurement (M) of percentages, gender gaps and confidence intervals at 95% of people with excellent or very good self-perceived general health.

General health self-perception: excelent or very good (Age group)	Total (percentage and confidence interval 95%)				Men (percentage and confidence interval 95%)				Women (percentage and confidence interval 95%)				Absolute Gender gap: ($Women^{Mt}$ - $Women^{M1}$) - (Men^{Mt} - Men^{M1}) (percentage points and confidence interval 95%)		
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M2 - M1	M3 - M1	M4 - M1
Total	83,1%	83,8%	82,2%	79,2%	87,2%	87,3%	86,3%	84,4%	79,1%	80,5%	78,2%	74,3%	1,3	-0,01	-2,01
	(81,5-84,5)	(82,0-85,4)	(80,59-83,69)	(77,5-80,9)	(85,0-89,2)	(84,5-89,6)	(84,0-88,4)	(81,8-86,7)	(76,9-81,2)	(78,0-82,7)	(76,0-80,3)	(71,9-76,6)	(-3,28; 5,88)	(-4,3; 4,29)	(-6,53; 2,52)
16-24	93,4%	97,4%	96,3%	97,5%	96,6%	97,1%	97,8%	97,6%	89,9%	97,8%	94,7%	97,3%	7,49	3,65	6,42
	(90,0-95,7)	(94,7-98,8)	(93,6-97,9)	(94,9-98,8)	(92,6-98,5)	(92,4-98,9)	(93,2-99,3)	(92,8-99,2)	(83,9-93,8)	(93,3-99,3)	(90,3-97,2)	(93,6-98,9)	(0,78; 14,2)	(-3,25; 10,54)	(-0,16; 13)
25-34	92,4%	96,4%	94,2%	91,4%	93,2%	98,1%	97,0%	92,1%	91,6%	94,6%	91,4%	90,7%	-1,99	-4,08	0,2
	(88,9-94,9)	(93,5-98,0)	(91,5-96,1)	(88,1-93,9)	(87,2-96,5)	(95,0-99,3)	(93,6-98,6)	(87,4-95,1)	(86,8-94,8)	(89,1-97,4)	(86,7-94,5)	(85,6-94,2)	(-9,28; 5,3)	(-11,46; 3,29)	(-7,95; 8,36)
35-44	90,6%	92,6%	92,1%	86,8%	92,3%	93,0%	95,2%	90,2%	89,0%	92,2%	89,0%	83,2%	2,49	-2,93	-3,66
	(87,8-92,9)	(89,9-94,6)	(89,7-94,0)	(83,5-89,4)	(87,8-95,2)	(88,9-95,7)	(92,0-97,2)	(85,0-93,8)	(84,9-92,0)	(88,1-94,9)	(85,1-91,9)	(78,9-86,9)	(-4,37; 9,36)	(-9,47; 3,61)	(-11,37; 4,05)
45-54	84,3%	85,9%	83,1%	79,2%	88,5%	90,6%	86,8%	86,4%	80,1%	81,3%	79,4%	72,1%	-0,84	1,06	-5,83
	(81,1-87,1)	(82,4-88,8)	(79,8-86,0)	(75,9-82,2)	(83,7-92,1)	(85,3-94,1)	(81,6-90,7)	(81,7-90,0)	(75,4-84,1)	(76,2-85,5)	(74,8-83,4)	(67,2-76,5)	(-9,53; 7,84)	(-7,56; 9,69)	(-14,46; 2,8)
55-64	78,4%	77,1%	75,3%	70,1%	82,3%	80,1%	80,5%	75,3%	74,6%	74,3%	70,3%	65,1%	1,99	-2,54	-2,5
	(73,8-82,3)	(72,4-81,3)	(71,2-78,9)	(65,8-74,0)	(75,0-87,8)	(72,5-85,9)	(74,9-85,2)	(69,0-80,7)	(69,0-79,9)	(67,9-79,8)	(64,3-75,6)	(59,0-70,7)	(-10,33; 14,32)	(-13,96; 8,89)	(-14,35; 9,35)
>=65	66,6%	62,7%	61,6%	61,0%	73,8%	68,8%	65,6%	69,3%	61,0%	57,8%	58,5%	54,5%	1,8	5,73	-1,97
	(62,2-70,8)	(57,1-68,0)	(56,6-66,4)	(55,7-66,1)	(67,0-79,7)	(59,1-77,1)	(57,4-73,0)	(59,9-77,3)	(55,0-66,7)	(50,8-64,6)	(52,2-64,6)	(48,0-60,9)	(-12,42; 16,03)	(-7,44; 18,89)	(-15,8; 11,86)

Fig. 4. Estimated percentages grouped by sex and age for the original categories of self-perceived general health at measurement 4



surements 3 and 4, with the decrease being slightly larger among women than among men. Regarding age groups, the evolution has been stable throughout the pandemic since the lockdown for the population between 16 and 34 years old, for men as well as for women. However, for the population above 35 years old, the evolution worsens as the age increases and the pandemic advances, especially in women. Therefore, this subpopulation got the lowest 'excellent or very good' general health values at the beginning of the lockdown for every age group above 35 years old and, also, it was when the difference with respect to men was bigger.

Fig. 5. Evolution of percentages and confidence intervals at 95% level of people with excellent or very good self-perceived general health with respect to age and sex

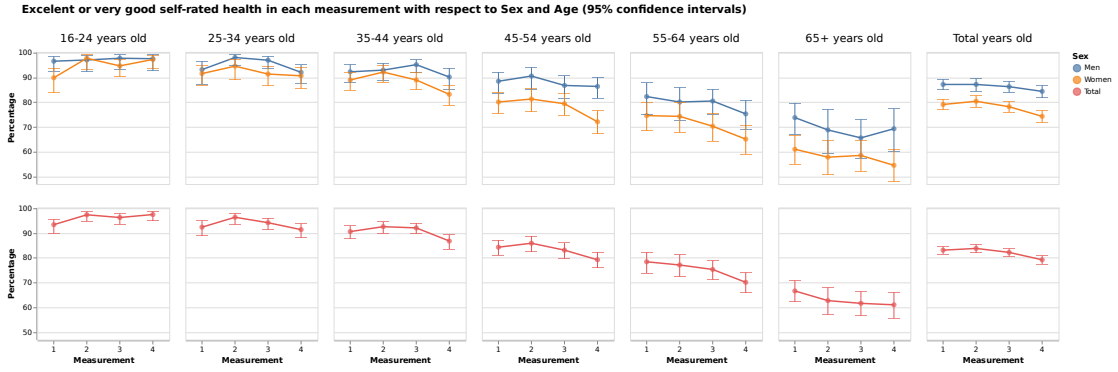
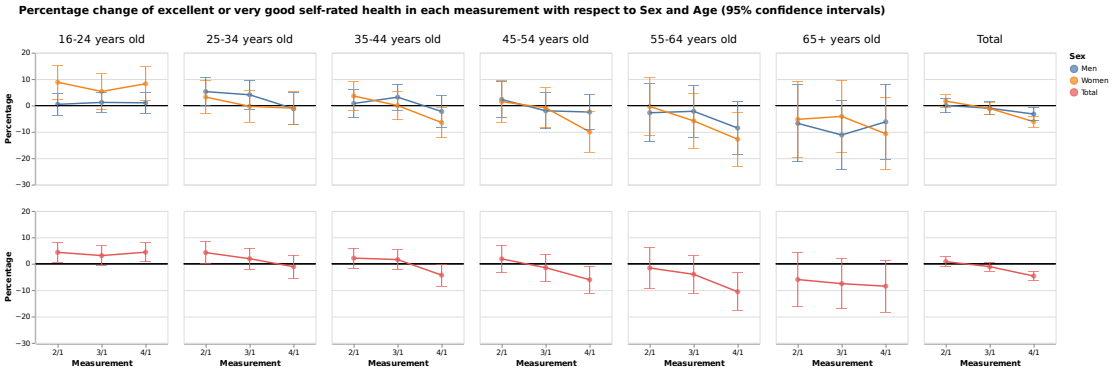


Table 3 and Figure 6 shows the relative percentage changes and 95% confidence intervals for each measurement with respect to measurement 1 for the 'excellent or very

Table 3. Evolution in each measurement (M) with respect to measurement 1 of relative percentage changes, gender gaps and 95% confidence intervals for people with excellent or very good self-perceived general health.

General health self-perception: excelent or very good (Age group)	Total (percentage and confidence interval 95%)			Men (percentage and confidence interval 95%)			Women (percentage and confidence interval 95%)			Relative Gender gap: ($Women^{Mt}$ / $Women^{M1}$) - (Men^{Mt} / Men^{M1}) (percentage points and confidence interval 95%)		
	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1	M2 / M1	M3 / M1	M4 / M1
Total	0,8% (-1,0; 2,7)	-1,1% (-2,8; 0,6)	-4,6% (-6,3; -3,0)	0,0% (-2,8; 2,8)	-1,0% (-3,5; 1,5)	-3,2% (-5,8; -0,6)	1,7% (-0,7; 4,1)	-1,1% (-3,3; 1,2)	-6,1% (-8,2; -4,0)	1,65 (-2,06; 5,36)	-0,11 (-3,47; 3,24)	-2,87 (-6,18; 0,45)
16-24	4,4% (0,7; 8,1)	3,1% (-0,6; 6,9)	4,4% (0,8; 8,0)	0,5% (-3,6; 4,5)	1,2% (-2,6; 5,0)	1,0% (-2,9; 5,0)	8,8% (2,3; 15,3)	5,3% (-1,4; 12,1)	8,3% (1,9; 14,7)	8,37 (0,71; 16,03)	4,15 (-3,6; 11,89)	7,22 (-0,29; 14,73)
25-34	4,3% (0,2; 8,3)	1,9% (-2,1; 6)	-1,1% (-5,5; 3,3)	5,3% (-0,1; 10,6)	4,1% (-1,4; 9,6)	-1,2% (-7,4; 5,0)	3,2% (-2,9; 9,3)	-0,3% (-6,3; 5,7)	-1% (-7,2; 5,3)	-2,08 (-10,2; 6,04)	-4,39 (-12,48; 3,71)	0,2 (-8,57; 8,97)
35-44	2,2% (-1,7; 6)	1,6% (-2,0; 5,3)	-4,3% (-8,5; -0,1)	0,8% (-4,5; 6,1)	3,2% (-1,7; 8,0)	-2,3% (-8,3; 3,8)	3,6% (-1,9; 9,1)	0,0% (-5,5; 5,4)	-6,5% (-12,2; -0,7)	2,83 (-4,83; 10,5)	-3,18 (-10,48; 4,13)	-4,2 (-12,54; 4,15)
45-54	1,9% (-3,4; 7,1)	-1,5% (-6,6; 3,6)	-6,0% (-11,1; -1,0)	2,3% (-4,5; 9,1)	-2,0% (-8,8; 4,8)	-2,5% (-8,9; 4,0)	1,5% (-6,5; 9,4)	-0,9% (-8,5; 6,7)	-10,0% (-17,6; -2,4)	-0,82 (-11,28; 9,65)	1,12 (-9,09; 11,32)	-7,54 (-17,53; 2,46)
55-64	-1,6% (-9,3; 6,2)	-4,0% (-11,1; 3,2)	-10,6% (-17,7; -3,5)	-2,7% (-13,8; 8,3)	-2,2% (-12,0; 7,6)	-8,5% (-18,5; 1,5)	-0,4% (-11,3; 10,6)	-5,8% (-16,2; 4,6)	-12,8% (-23,1; -2,5)	2,39 (-13,18; 17,96)	-3,63 (-17,92; 10,67)	-4,24 (-18,56; 10,09)
>=65	-6,0% (-16,2; 4,3)	-7,5% (-17,0; 1,9)	-8,5% (-18,3; 1,3)	-6,8% (-21,4; 7,8)	-11,2% (-24,2; 1,9)	-6,2% (-20,4; 8,1)	-5,2% (-19,7; 9,2)	-4,1% (-17,8; 9,6)	-10,7% (-24,3; 2,9)	1,53 (-18,99; 22,06)	7,04 (-11,86; 25,94)	-4,53 (-24,24; 15,18)

Fig. 6. Evolution of relative percentage changes and 95% confidence intervals for people with excellent or very good self-perceived general health in each measurement (M) with respect to measurement 1.



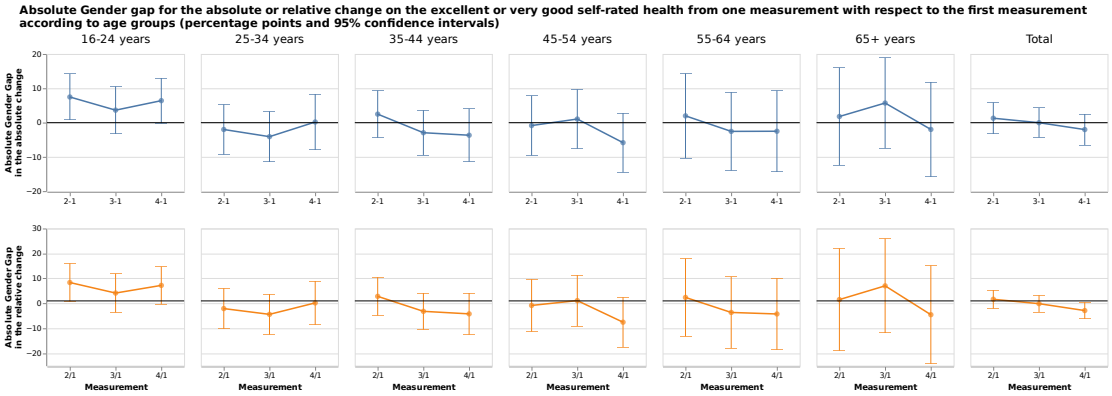
good' self-perceived general health variable. It could be observed that excellent or very good general health decreased in the general population by a 4.6% (CI95%=[-6.3;-3] in measurement 4 with respect to measurement 1. In addition, the decrease among women was of double the decrease among men (-6.1 and -3.2% respectively) and this gap was larger in older people, specially for women, except for women between 16 and 24 years old, where an increase was observed in all measurements with respect to measurement 1.

Tables 2 and 3 incorporate absolute gender gaps on absolute and relative changes respectively, i.e. the absolute difference (in percentage points) between men and women of absolute and relative changes of a given measurement with respect to measurement 1. Their interpretation would be that a positive value indicates that women showed a larger change (absolute or relative) in comparison to men in their 'excellent or very good' self-perceived general health. Therefore, this result could be seen as a positive gender gap (i.e., better result or favorable to women) in the corresponding measurement compared to the first one. On the contrary, a negative value indicates that women showed a smaller change (absolute or relative) in comparison to men in their 'excellent or very good' self-perceived general health, which could be seen as a negative (or unfavourable to women) gender gap. These results are visualized in Figure 7; it can be observed that, for example, the gender gap was negative along the pandemic, confirming an increasingly negative impact on women with respect to men in relation to self-perceived general health. The gender gap went from positive in measurement 2 with respect to measurement 1, by 1.65 percentage points, to negative and statistically significant in measurement 4 by -2.87 percentage points. Results by age reveal that negative gender gaps were observed in people over 35 years old in measurement 4 (with respect to measurement 1), with the group between 45 and 54 years old showing the largest gender gap, of almost 8 percentage points (relative change).

Table 4. Percentage of people whose self-perceived general health improves, deteriorates or remains the same between measurements, and absolute and relative gender gap in the percentage

Age group (years)	Self-perceived general health	Total			Women			Men			Absolute gender gap (change in percentage points: Women - Men)			Relative gender gap (percentages change rate: Women / Men)		
		M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)	M2-M1 Percentage (CI95%)	M3-M2 Percentage (CI95%)	M4-M3 Percentage (CI95%)
Total	Health improves	20,75% (18,75-22,76)	20,58% (18,36-22,8)	18,32% (16,56-20,07)	19,47% (16,44-22,51)	24,21% (20,59-27,82)	19,92% (17,1-22,73)	21,98% (19,35-24,61)	17,12% (14,54-19,7)	16,78% (14,66-18,9)	2.51 (-1.51; 6.53)	-7.09 (-11.53; -2.65)	-3.14 (-6.66; 0.38)	12.73% (-12.15; 37.61)	-29.39% (-46.52; -12.27)	-16.19% (-33.81; 1.44)
	Remains the same	54,96% (52,5-57,42)	54,04% (51,46-56,62)	56,97% (54,74-59,2)	55,21% (51,44-58,98)	53,24% (49,24-57,25)	56,31% (52,86-59,75)	54,72% (51,54-57,89)	54,80% (51,51-58,08)	57,61% (54,75-60,47)	-0.49 (-5.43; 4.44)	1.55 (-3.63; 6.73)	1.3 (-3.17; 5.78)	-1.03% (-14.19; 12.13)	2.75% (-11.26; 16.77)	1.79% (-10.12; 13.7)
	Health deteriorates	24,29% (22,16-26,42)	25,38% (23,15-27,61)	24,71% (22,74-26,68)	25,32% (22,03-28,61)	22,55% (19,22-25,88)	23,78% (20,75-26,81)	23,30% (20,57-26,03)	28,09% (25,12-31,06)	25,61% (23,07-28,15)	-2.02 (-6.29; 2.26)	5.54 (1.07; 10)	1.83 (-2.12; 5.79)	-8.08% (-26.6; 10.43)	24.37% (-1.67; 50.42)	7.16% (-12.97; 27.29)
16-24	Health improves	29,90% (23,88-35,93)	21,48% (15,6-27,37)	24,10% (18,41-29,8)	25,30% (17,35-33,24)	28,37% (18,86-37,88)	20,92% (12,43-29,41)	34,80% (25,83-43,77)	14,16% (7,76-20,55)	27,49% (20-34,98)	9.5 (-2.48; 21.49)	-14.22 (-25.67; -2.76)	6.58 (-4.75; 17.9)	37.57% (-29.22; 104.35)	-50.1% (-81.14; -19.06)	31.44% (-41.35; 104.22)
	Remains the same	44,26% (37,76-50,76)	53,92% (46,73-61,12)	57,08% (50,39-63,77)	45,83% (36,62-55,03)	53,79% (43,12-64,46)	59,93% (49,77-70,08)	42,60% (33,45-51,75)	54,06% (44,49-63,63)	54,05% (45,47-62,62)	-3.23 (-16.21; 9.75)	0.27 (-14.07; 14.6)	-5.88 (-19.17; 7.41)	-7.04% (-43.39; 29.3)	0.5% (-38.85; 39.85)	-9.81% (-42.8; 23.17)
	Health deteriorates	25,84% (20,09-31,59)	24,59% (18,34-30,85)	18,82% (13,64-24)	28,88% (20,44-37,32)	17,84% (9,55-26,12)	19,16% (11,24-27,07)	22,60% (14,92-30,28)	31,78% (22,66-40,91)	18,46% (11,86-25,06)	-6.27 (-17.69; 5.14)	13.95 (1.62; 26.28)	-0.7 (-11; 9.61)	-21.73% (-62.34; 18.88)	78.18% (-33.05; 189.42)	-3.64% (-61.56; 54.29)
25-34	Health improves	20,39% (14,01-26,77)	24,09% (17,85-30,34)	20,28% (15,3-25,26)	17,93% (7,67-28,2)	29,62% (19,52-39,71)	23,06% (14,85-31,26)	22,92% (15,32-30,52)	18,39% (11,46-25,33)	17,42% (12,01-22,83)	4.99 (-7.79; 17.76)	-11.22 (-23.47; 1.02)	-5.64 (-15.47; 4.19)	27.8% (-69.74; 125.35)	-37.9% (-74.51; -1.29)	-24.45% (-64.79; 15.9)
	Remains the same	55,39% (48,35-62,42)	48,07% (41,14-55)	54,24% (48,29-60,2)	58,48% (47,55-69,41)	43,25% (33,01-53,49)	52,19% (42,97-61,41)	52,19% (43,44-60,95)	53,03% (44,12-61,94)	56,36% (48,92-63,81)	-6.29 (-20.29; 7.72)	9.78 (-3.8; 23.36)	4.18 (-7.67; 16.03)	-10.75% (-43.25; 21.75)	22.61% (-22.76; 67.99)	8% (-26.98; 42.99)
	Health deteriorates	24,23% (18,62-29,84)	27,84% (21,27-34,41)	25,47% (20,36-30,58)	23,58% (15,43-31,74)	27,13% (16,81-37,45)	24,75% (17,04-32,47)	24,89% (17,24-32,54)	28,57% (20,48-36,67)	26,21% (19,56-32,87)	1.3 (-9.88; 12.49)	1.44 (-11.67; 14.56)	1.46 (-8.73; 11.65)	5.53% (-48.55; 59.6)	5.32% (-55.13; 65.77)	5.9% (-42.95; 54.75)
35-44	Health improves	20,39% (15,85-24,93)	20,19% (15,29-25,09)	16,42% (12,85-19,98)	17,78% (10,74-24,83)	22,66% (14,44-30,88)	15% (9,51-20,49)	23,04% (17,41-28,67)	17,68% (12,47-22,9)	17,87% (13,35-22,38)	5.26 (-3.77; 14.28)	-4.98 (-14.71; 4.76)	2.86 (-4.25; 9.98)	29.55% (-36.93; 96.04)	-21.96% (-63.49; 19.56)	18.77% (-38.83; 76.37)
	Remains the same	56,87% (51,07-62,68)	51,37% (45,6-57,14)	56,09% (51,27-60,91)	57,66% (48,2-67,13)	53,14% (43,8-62,48)	59,28% (51,71-66,84)	56,08% (49,43-62,72)	49,57% (42,87-56,27)	52,85% (46,93-58,76)	-1.59 (-13.15; 9.98)	-3.56 (-15.06; 7.93)	-6.43 (-16.03; 3.17)	-2.75% (-33.51; 28)	-6.71% (-36.49; 23.08)	-11.1% (-34.09; 11.9)
	Health deteriorates	22,73% (17,78-27,68)	28,44% (23,44-33,44)	27,49% (23,17-31,81)	24,55% (16,33-32,77)	24,20% (16,6-31,81)	25,72% (18,99-32,45)	20,89% (15,43-26,34)	32,74% (26,44-39,05)	29,29% (23,9-34,68)	-3.67 (-13.54; 6.2)	8.54 (-1.34; 18.42)	3.57 (-5.06; 12.19)	-14.94% (-56.03; 26.15)	35.28% (-21.88; 92.45)	13.54% (-28.92; 56.01)
45-54	Health improves	20,83% (16,97-24,69)	17,21% (13,52-20,91)	17,64% (13,8-21,49)	19,26% (13,75-24,78)	21,34% (15,37-27,31)	21,80% (15,31-28,3)	22,38% (17-27,76)	13,13% (8,63-17,63)	13,49% (9,59-17,39)	3.12 (-4.59; 10.82)	-8.21 (-15.68; -0.73)	-8.32 (-15.89; -0.74)	16.78% (-31.81; 65.36)	-38.08% (-67.22; -8.93)	-38.33% (-67.16; -9.49)
	Remains the same	58,96% (54,22-63,7)	60,44% (55,2-65,68)	55,65% (51,1-60,2)	58% (50,92-65,08)	57,10% (48,76-65,45)	53,57% (46,54-60,6)	59,92% (53,58-66,26)	63,74% (57,38-70,1)	57,73% (52,01-63,46)	1.92 (-7.58; 11.42)	6.64 (-3.86; 17.13)	4.17 (-4.9; 13.23)	3.83% (-23.15; 30.81)	12.32% (-19.48; 44.12)	7.46% (-18.02; 32.94)
	Health deteriorates	20,21% (16,39-24,03)	22,35% (17,66-27,03)	26,70% (22,77-30,64)	22,74% (16,94-28,54)	21,56% (13,99-29,13)	24,63% (18,79-30,46)	17,70% (12,7-22,71)	23,13% (17,56-28,7)	28,78% (23,53-34,02)	-5.03 (-12.69; 2.62)	1.57 (-7.83; 10.97)	4.15 (-3.7; 12)	-21.75% (-54.76; 11.27)	7.95% (-45.11; 61.01)	16.52% (-23.63; 56.67)
55-64	Health improves	17,21% (12,7-21,72)	18,76% (14,3-23,22)	16,31% (12,69-19,93)	20,02% (12,74-27,3)	22,26% (15,71-28,8)	15,78% (10,76-20,81)	14,54% (9,17-19,91)	15,43% (9,36-21,51)	16,82% (11,61-22,02)	-5.48 (-14.53; 3.56)	-6.82 (-15.75; 2.11)	1.03 (-6.2; 8.27)	-26.99% (-68.66; 14.69)	-30.28% (-68.06; 7.5)	7.03% (-44.98; 59.05)
	Remains the same	56,22% (50,32-62,13)	56,01% (50,29-61,73)	61,52% (56,77-66,27)	52,50% (43,49-61,5)	55,93% (47,91-63,95)	60,37% (53,63-67,11)	59,77% (52,18-67,36)	56,09% (47,93-64,24)	62,62% (55,93-69,3)	7.28 (-4.5; 19.05)	0.16 (-11.28; 11.59)	2.24 (-7.25; 11.74)	14.5% (-20.37; 49.37)	0.83% (-30.22; 31.88)	4.2% (-21.6; 30)
	Health deteriorates	26,57% (21,14-32)	25,23% (20,19-30,27)	22,17% (18,11-26,22)	27,48% (19,02-35,94)	21,82% (15,27-28,36)	23,84% (17,96-29,72)	25,69% (18,82-32,56)	28,48% (20,94-36,02)	20,57% (14,98-26,15)	-1.79 (-12.69; 9.11)	6.66 (-3.32; 16.65)	-3.28 (-11.38; 4.83)	-5.99% (-51.82; 39.83)	31.26% (-29.42; 91.94)	-13.34% (-49.38; 22.7)
>=65	Health improves	18,90% (14,3-23,5)	22,50% (16,06-28,94)	17,61% (13,21-22,01)	18,43% (11,5-25,35)	23,43% (12,04-34,81)	23,38% (15,81-30,96)	19,27% (13,11-25,43)	21,77% (14,6-28,94)	13,05% (7,95-18,14)	0.85 (-8.42; 10.11)	-1.66 (-15.11; 11.8)	-10.34 (-19.47; -1.21)	3.13% (-52.75; 59.01)	-8.52% (-73.11; 56.06)	-45.38% (-75.57; -15.2)
	Remains the same	54,21% (48,16-60,25)	53,26% (46,53-59,98)	57,34% (51,36-63,32)	55,54% (46,22-64,86)	54,68% (43,61-65,76)	53,58% (43,97-63,19)	53,16% (45,22-61,1)	52,13% (43,81-60,46)	60,31% (52,9-67,73)	-2.38 (-14.62; 9.87)	-2.55 (-16.41; 11.31)	6.73 (-5.4; 18.87)	-5.62% (-36.55; 25.32)	-6.15% (-38.36; 26.06)	10.21% (-23.1; 43.51)
	Health deteriorates	26,89% (21,39-32,4)	24,24% (18,79-29,7)	25,05% (19,42-30,68)	26,03% (17,3-34,77)	21,89% (13,77-30,01)	23,04% (13,53-32,55)	27,57% (20,51-34,62)	26,10% (18,8-33,39)	26,64% (19,86-33,42)	1.53 (-9.7; 12.76)	4.21 (-6.7; 15.12)	3.6 (-8.07; 15.28)	4.4% (-48; 56.8)	17.36% (-43.15; 77.88)	13.22% (-52.72; 79.16)

Fig. 7. Gender gap for the change on the excellent or very good self-perceived health from one measurement with respect to the first measurement



5.5. Longitudinal results

Table 4 shows estimates of better, equal or worse self-perception of health in the population for a given measurement with respect to the same population in the previous measurement. 20.75% of the andalusian population below 16 years old improved their self-perceived general health in measurement 2 with respect to measurement 1, but this percentage was slightly smaller in the following measurements, specially in measurement 4 with respect to measurement 3. On the contrary, 24.29% of this population group presented a worse self-perceived general health in measurement 2 with respect to measurement 1, with this percentage being slightly greater in next measurements. When we analyze these results by sex, it can be observed that it is the women that experiment that decrease in the improvement of general health, as well as the increase in the deterioration of self-perceived general health. Regarding the age, the decreases in the improvement of general health along the pandemic are observed among women between 25 and 54 years old, and the increase in deterioration percentages are observed in those women between 45 and 54 years old. On the other hand, the percentage of people that remained with the same self-perceived general health status, in a given measurement with respect to the previous one, did not vary along the pandemic, except for the population below 24 years old that did experiment increases in the aforementioned percentage, going from 44.26% in measurement 2 to 57.08% in measurement 4. These results are visualized in Figure 8.

If we calculate the ratio of the population that worsens their general health (in a given measurement with respect to the previous one) and the population that improves it, a positive value means that there are more people whose self-perceived general health has deteriorated than people whose health has improved, as seen at Figure 9. In relative terms, it could be observed that, in measurement 2 with respect to measurement 1, there was 17.06% more population with worse health than with better health; this percentage increased to 23.32% and 34.88% in measurements 3 and 4 with respect to measurements 2 and 3 respectively. These differences are greater in women, reaching values of 64.08% and 52.32% in measurements 3 and 4 respectively. If the ratio is analyzed according

Fig. 8. Percentage of population whose self-perceived general health improves, deteriorates or remains the same between a measurement and the previous one

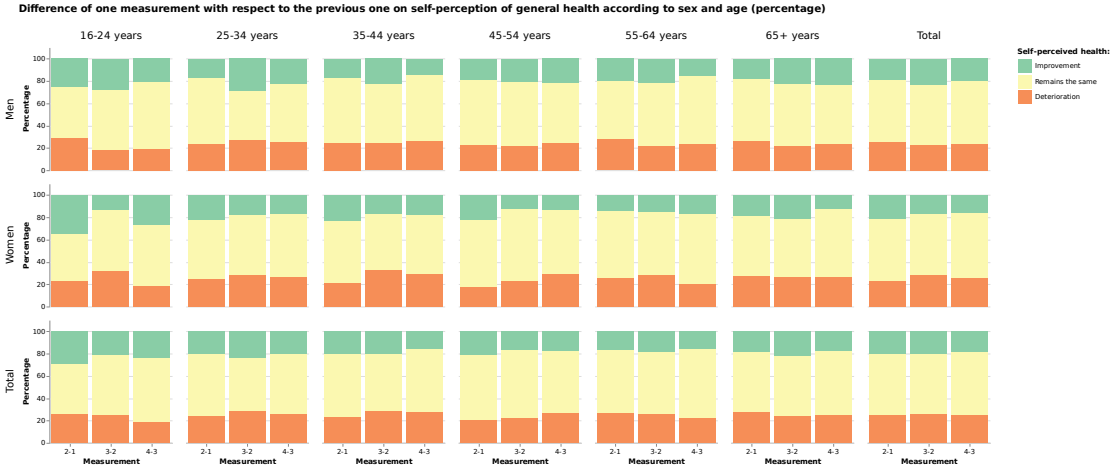
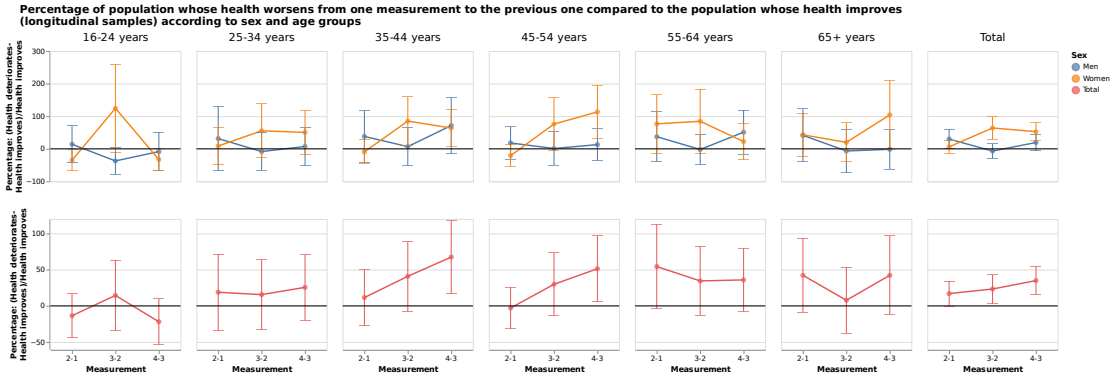


Fig. 9. Percentage of population whose health worsens from one measurement to the previous one compared to the population whose health improves

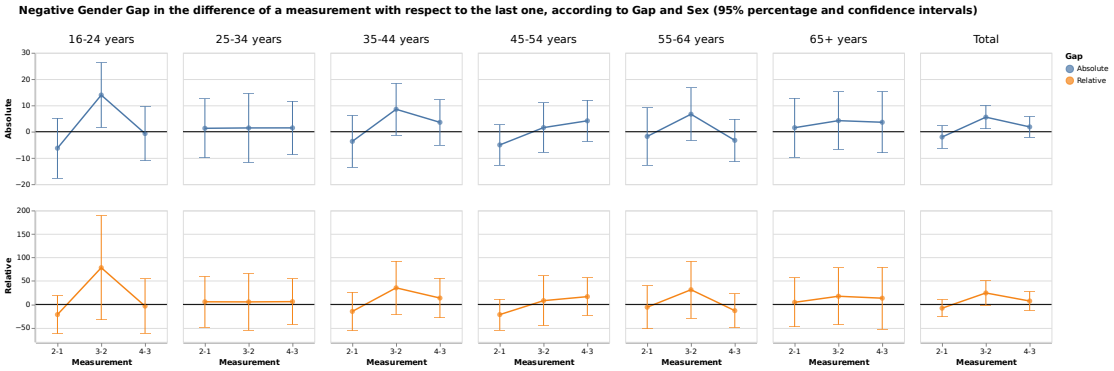


to the age of the individuals, we can spot different patterns in men and women. In this sense, more men population under 55 years old perceived a deterioration in their health in measurement 2 with respect to measurement 1, while this was only observed in women above that age in the women population. Regarding measurement 3 with respect to measurement 2, deterioration of health was more frequently observed in women of any age, with almost no changes between age groups in men. Finally, deterioration of health in measurement 4 with respect to measurement 3 was more frequently observed in men population between 25 and 64 years old, something that was also observed in women population above 25 years old.

Table 4 also shows absolute and relative gender gaps in the improvement, in staying the same or in the deterioration of self-perceived general health in a measurement with

respect to the previous one in the same population. On the one hand, absolute gender gap is the absolute difference (in percentage points) between women and men with a better, equal or worse perceived health in a measurement with respect to the previous one, and on the other hand relative gender gap is the relative difference (in percentage) between women and men with a better, equal or worse perceived health. This means that a positive value in the gap (absolute or relative) indicates that the population percentage of women that improved, stayed the same or worsened their self-perceived general health was greater than the corresponding percentage in men population. A negative value would indicate that the percentage was smaller in women. Regarding deterioration of health , we observe at Figure 10 that the percentage of women population whose self-perceived health was worse in measurement 2 than in measurement 1 was 8% lower than its men population counterpart. However, this relative gender gap on health deterioration became positive in next measurements, i.e. the deteriorating percentages were greater among women in measurement 3 and in measurement 4. This result was observed across all age groups, except for the population younger than 24 years old and the population between 55 and 64 years old.

Fig. 10. Negative Gender Gap in the difference of a measurement with respect to the last one



6. Conclusions

The rapid evolution of the COVID-19 pandemic has forced researchers to provide timely estimates on the impact of the disease in the population. This has often lead to the establishment of survey studies which did not meet the criteria to be considered probabilistic, entailing many sources of error that may have affect the final estimates obtained from them. For this reason, the ESSOC survey is particularly valuable in the sense that its overlapping probability panel design offers the possibility of obtaining reliable estimates, both cross-sectional and longitudinal, on the impact of COVID-19 on health and its determinants. However, the analysis of the survey has not been exempt from statistical adjustments to correct for attrition and survey nonresponse.

The two-step adjustment procedure has been established in this study to remove the two main sources of error in the sampling design: the population nonresponse,

understood as people who did not take part in the survey despite having been selected in the sample, which was treated in the calibration step, and panel nonresponse, understood as people who participated in some of the measurements but did not follow up in further ones. Panel nonresponse has been treated using PSA, which is a technique often used for addressing selection bias in online surveys (Lee, 2006) but which can also be used for nonresponse; in fact, it was originally adapted from Rosenbaum and Rubin (1983) for that matter (Little, 1986).

In our study, the XGB technique has been used to model the lack of response from one measurement to another. Other ML methods (as logistic regression, decision trees, random forests, ...) could be used, but several papers show (Kern et al. (2021), Castro-Martín et al. (2020), Ferri-García and Rueda (2020)) that the set of predictor variables used in general mattered more than the type of ML technique. The application developed in this work is one example where techniques of the machine learning field have to be combined with other important techniques in survey research as calibration and PSA, when studying nonresponse in a panel setting.

Tables 5 and 6 summarise the name, table, figure, formula and interpretation related to the estimators developed throughout this paper for the cross-sectional and longitudinal samples, respectively.

The results observed in the different estimates obtained from the survey show that the impact of the pandemic has hit differently across age groups and genders. More precisely, the self-perceived general health seems to have decreased more notably in older age groups and women, both according to the evolution of cross-sectional estimates and longitudinal estimates. The gender gap, both in absolute and relative terms, has mostly grown as the pandemic advanced, meaning that the changes (mostly decreases in self-perceived general health) have been larger and worse in women in comparison to men. The variable of interest has been the self-perceived general health. It is a well-known fact that subjective variables usually entail measurement errors, as the response given in such questions by the interviewee may depend on many unmeasurable factors unrelated to the matter of study but that move the final response away from the objective value that should be given. Further studies should consider the measurement of such variables using validated instruments for a more objective understanding of the matter.

The descriptive results for the general health self-perception variable are an example applied to this paper in order to show the different estimators, tables and figures developed. All these are extended to the more than 400 ESSOC variables through the web platform developed at `www.researchprojects.com/ESSOC`. On this website, after selecting the set of variables to be described, the estimators to be shown and the segmentation variables to be considered (sex and age or sex and degree of urbanisation), the user obtains the corresponding interactive figures to help the interpretations for the selected variables. This will allow the scientific community not only to access the descriptive results for all the variables of the ESSOC, but also to carry out their own analyses.

Funding

Funding has been obtained from the competitive calls of the SUPERA COVID-19 Fund

Table 5. Name, table, figure, formula and interpretation of each estimator developed for the cross-sectional samples

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Original variables	1	4	(9)	Percentages, confidence intervals at 95%, sample size and population estimations at measurement 4, grouped by sex and age, for the original categories of self-perceived general health.
Dichotomized variables	2	5	(9)	Evolution of percentages and confidence intervals at 95%, grouped by sex and age, of people with excellent or very good self-perceived general health. If the confidence intervals for the same measurement do not overlap, it can be said that there are statistically significant differences between women and men. Similarly, if the confidence intervals of two different measurements do not overlap, it can be said that there are statistically significant differences between them.
Absolute/Relative change	No/3	No/6	(12)/(13)	Evolution in each measurement with respect to measurement 1 of absolute/relative changes and confidence intervals at 95%, grouped by sex and age, of people with excellent or very good self-perceived general health. A positive value indicates an increase, in percentage points/terms, in the excellent or very good self-perception of overall health of the corresponding measure compared to the first measure. Conversely, a negative value indicates a decrease, in percentage points/terms, in the excellent or very good self-perception of overall health of the corresponding measure compared to the first one. If the confidence interval does not include the value 0, this increase or decrease can be said to be statistically significant. Similarly, if the confidence intervals for the same measurement do not overlap, it can be said that there are statistically significant differences between women and men.
Absolute/Relative gender gap in the absolute change	2/No	7/No	(14)/(15)	Evolution in each measurement (M) with respect to measurement 1 of absolute/relative gender gaps (women versus men) in the absolute/relative change and confidence intervals at 95%, grouped by age, of people with excellent or very good self-perceived general health. A positive value indicates that women show, in percentage points/terms, a larger absolute/relative change in comparison to men in their 'excellent or very good' self-perceived general health of the corresponding measure compared to the first one. Therefore, this result could be seen as a positive gender gap (i.e., better result or favorable to women) in the corresponding measurement compared to the first one. On the contrary, a negative value indicates that women showed, in percentage points/terms, a smaller absolute/relative change in comparison to men in their 'excellent or very good' self-perceived general health. It could be seen as a negative gender gap (i.e., worse result or unfavorable to women) in the corresponding measurement compared to the first one. If the confidence interval does not include the value 0, the corresponding gender gap can be said to be statistically significant.
Absolute/Relative gender gap in the relative change	3/No	7/No	(16)/(17)	

Table 6. Name, table, figure, formula and interpretation of each estimator developed for the longitudinal samples

NAME	TABLE	FIGURE	FORMULA	INTERPRETATION
Longitudinal difference	4	8	(26)	Percentage of population and confidence intervals at 95% whose self-perceived general health increases/improves, decreases/deteriorates, or remains the same between a measurement and the previous one
Decrease Increase Rate	No	9	(27)	Percentage of the population and confidence intervals at 95% that worsens their general health (in a given measurement with respect to the previous one) and the population that improves it. A positive value means that there are more people whose self-perceived general health has deteriorated than people whose health has improved.
Absolute/Relative gender gap in the absolute difference	4	10	(28)/(30)	Absolute/Relative difference (in percentage points/terms) and confidence intervals at 95% between women and men with a better, equal, or worse self-perceived health in a measurement with respect to the previous one. A positive value indicates that the percentage of women that improved, stayed the same or worsened their self-perceived general health was greater than the corresponding percentage in men population. A negative value would indicate that the percentage was smaller in women.

of Santander Universities (SAUN), the Conference of Rectors of Spanish Universities (CRUE), and the Higher Council for Scientific Research (CSIC), in addition to the COVID Competitive Aid Program -19 from Pfizer Global Medical Grants. This work is also partially supported by Ministerio de Economía y Competitividad and Ministerio de Ciencia y Educación of Spain [grant PID2019-106861RB-I00 and PI20/00855, respectively] and IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. Anchorage AK USA: ACM. URL: https://dl.acm.org/doi/10.1145/3292500.3330701.

Andalusian Institute of Statistics and Cartography (IECA, Spanish acronym) (2020) Population and housing census. URL: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/padron/index.htm.

Arcos, A., Rueda, M. d. M. and Pasadas-del Amo, S. (2020) Treating Nonresponse in Probability-Based Online Panels through Calibration: Empirical Evidence from a Survey of Political Decision-Making Procedures. *Mathematics*, **8**, 423. URL: https://www.mdpi.com/2227-7390/8/3/423.

Bergstra, J., Bardenet, R., Bengio, Y. and Kégl, B. (2011) Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Bergstra, J., Yamins, D. and Cox, D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, 115–123. PMLR. URL: https://proceedings.mlr.press/v28/bergstra13.html.

Cabrera-León, A., Lopez-Villaverde, V., Rueda, M. and Moya-Garrido, M. (2015) Calibrated prevalence of infertility in 30- to 49-year-old women according to different approaches: a cross-sectional population-based study. *Human Reproduction*, **30**, 2677–2685. URL: https://academic.oup.com/humrep/article-lookup/doi/10.1093/humrep/dev226.

Castro-Martín, L., Rueda, M. d. M. and Ferri-García, R. (2020) Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment. *Mathematics*, **8**, 2096. URL: https://www.mdpi.com/2227-7390/8/11/2096.

Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. San Francisco California USA: ACM. URL: https://dl.acm.org/doi/10.1145/2939672.2939785.

Devaud, D. and Tillé, Y. (2019) Rejoinder on: Deville and Särndal's calibration: revisiting a 25-year-old successful optimization problem. *TEST*, **28**, 1087–1091. URL: http://link.springer.com/10.1007/s11749-019-00685-z.

Deville, J.-C. and Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 376–382. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217.

Ferri-García, R. and Rueda, M. d. M. (2018) Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Statistics and Operations Research Transactions*, 159–162. URL: https://raco.cat/index.php/SORT/article/view/347847.

— (2020) Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLOS ONE*, **15**, e0231500. URL: https://dx.plos.org/10.1371/journal.pone.0231500.

Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, **28**. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-2/Additive-logistic-regression--a-statistical-view-of-boosting-With/10.1214/aos/1016218223.full.

Kalton, G. and Brick, J. M. (1995) Weighting schemes for household panel surveys. *Survey Methodology*, **21**, 33–44.

Kalton, G. and Citro, C. F. (1995) Panel surveys: Adding the fourth dimension. *Innovation: The European Journal of Social Science Research*, **8**, 25–39. URL: http://www.tandfonline.com/doi/abs/10.1080/13511610.1995.9968429.

Kalton, G., Lepkowski, J. and Lin, T. K. (1985) Compensating for wave nonresponse in the 1979 ISDP research panel. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, vol. 372, 377.

Kern, C., Klausch, T. and Kreuter, F. (2019) Tree-based Machine Learning Methods for Survey Research. *Survey research methods*, **13**, 73–93. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7425836/.

Kern, C., Weiß, B. and Kolb, J.-P. (2021) Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning. *Journal of Survey Statistics and Methodology*, smab009. URL: https://academic.oup.com/jssam/advance-article/doi/10.1093/jssam/smab009/6364780.

Kott, P. S. and Liao, D. (2015) One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, **41**, 165–182.

Kovacevic, M. S. (2001) Cross-sectional inference based on longitudinal surveys: Some experiences with statistics Canada surveys. In *Federal Committee on Statistical Methodology Conference*.

Lavallee, P. (1995) Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, **21**, 25–32.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine*, **29**, 337–346. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.3782.

— (2011) Weight Trimming and Propensity Score Weighting. *PLoS ONE*, **6**, e18174. URL: https://dx.plos.org/10.1371/journal.pone.0018174.

Lee, S. (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, **22**, 329.

Lepkowski, J. M. (1989) Treatment of wave nonresponse in panel surveys. *Panel surveys*.

Little, R. J. (1986) Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, 139–157.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013) A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, **32**, 3388–3414. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.5753.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004) Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, **9**, 403–425. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.9.4.403.

Rendtel, U. and Harms, T. (2009) Weighting and Calibration for Household Panels. In *Methodology of Longitudinal Surveys* (ed. P. Lynn), 265–286. Chichester, UK: John Wiley & Sons, Ltd. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470743874.ch15.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41.

Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2006) Mean estimation with calibration techniques in presence of missing data. *Computational Statistics & Data Analysis*, **50**, 3263–3277. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947305001349.

Sánchez-Cantalejo, C., Rueda, M. d. M., Saez, M., Enrique, I., Ferri, R., Fuente, M. d. L., Villegas, R., Castro, L., Barceló, M. A., Daponte-Codina, A., Lorusso, N. and Cabrera-León, A. (2021) Impact of COVID-19 on the Health of the General and More Vulnerable Population and Its Determinants: Health Care and Social Survey–ESSOC, Study Protocol. *International Journal of Environmental Research and Public Health*, **18**, 8120. URL: https://www.mdpi.com/1660-4601/18/15/8120.

Särndal, C.-E., Swensson, B. and Wretman, J. H. (2003) *Model assisted survey sampling*. Springer series in statistics. New York Berlin Heidelberg: Springer, 1. softcover print edn.

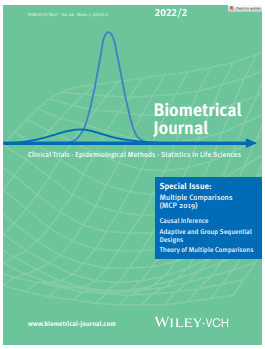
Tu, C. (2019) Comparison of various machine learning algorithms for estimating generalized propensity score. *Journal of Statistical Computation and Simulation*, **89**, 708–719. URL: https://www.tandfonline.com/doi/full/10.1080/00949655.2019.1571059.

Zhu, Y., Coffman, D. L. and Ghosh, D. (2015) A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. *Journal of Causal Inference*, **3**, 25–40. URL: https://www.degruyter.com/document/doi/10.1515/jci-2014-0022/html.

Appendix A6

Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain.

Rueda, María del Mar; Pasadas-del-Amo, Sara; Cobo, Beatriz; Castro-Martín, Luis; Ferri-García, Ramón (2022)
Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain.
Biometrical Journal (in revision)



STATISTICS & PROBABILITY			
JCR Year	Impact factor	Rank	Quartile
2020	2.207	39/125	Q2

Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Enhancing estimation methods for integrating probability and non-probability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain.

Rueda María del Mar[*1], **Pasadas-del-Amo Sara** [2], **Cobo Beatriz** [3], **Castro Luis** [1], and **Ferri Ramón** [1]

[1] *Department of Statistics and Operational Research, University of Granada. Avda. Fuentenueva s/n, 18071, Granada, Spain*

[2] *Institute for Advanced Social Studies/Spanish Research Council (IESA-CSIC), Campo Santo de los Mártires 7, 14004 Córdoba, Spain*

[3] *Department of Quantitative Methods for Economics and Business, University of Granada. Paseo de Cartuja, 7, 18011 Granada, Spain*

Received zzz, revised zzz, accepted zzz

Abstract

Web surveys have replaced Face-to-Face and CATI as the main mode of data collection in most countries. This trend was reinforced as a consequence of Covid-19 pandemic related restrictions. However, this mode still face significant limitations in obtaining probability-based samples of the general population. For this reason, most web surveys rely on non-probability survey designs. Whereas probability-based designs continue to be the gold standard in survey sampling, non-probability web surveys may still prove useful in some situations. For instance, when small sub-populations are the group under study and probability sampling is unlikely to meet sample size requirements, complementing a small probability sample with a larger non-probability one may improve the efficiency of the estimates. Non-probability samples may also be designed as a mean for compensating for known biases in probability-based web survey samples by purposely targeting respondent profiles that tend to be underrepresented in these surveys. This is the case in the Survey on the impact of the COVID-19 pandemic in Spain (ESPACOV) that motivates this paper.

In this article we propose a methodology for combining probability and non-probability web-based survey samples with the help of machine learning techniques. We then assess the efficiency of the resulting estimates by comparing them with other strategies that have been used before. Our simulation study and the application of the proposed estimation method to the second wave of the ESPACOV survey allow us to conclude that this is the best option for reducing the biases observed in our data.

Key words: Machine learning techniques; Non-probability surveys; Propensity score adjustment; Survey sampling; COVID19.

1 Introduction

Ten years have passed since the American Association for Public Opinion Research (AAPOR) appointed a taskforce to evaluate non-probability survey sampling methods that were more and more frequently used in applied research contexts at the time (11). During the last decade, non-probability sampling designs have continued to grow as the use of Big Data and web surveys spread (27). The lockdowns that followed the onset of the COVID-19, together with the need of quick data to grasp the impacts of the pandemic and to inform policymakers decisions, further bolstered non-probability sampling designs (22). Web surveys mushroomed during COVID-19 related confinements and have further displaced traditional survey modes,

*Corresponding author: e-mail: mrueda@ugr.es, Phone: +34958240494

such as Face to Face and CATI, that still face important restrictions due to social distancing rules and population fear to COVID-19 infection.

Web surveys conducted during this time have relied mostly on online convenience samples using social media and/or river sampling to recruit participants or on quota samples selected from online access panels (36). However, the use of these sampling methods comes at the cost of adding non-coverage and self-selection to the array of potential sources of sampling bias and applying complex estimation procedures to the resulting data (2), (38).

In spite of these challenges, non-probability web surveys may provide useful information that would not be available otherwise (27). In other cases, where small sub-populations are the group under study and probability sampling is unlikely to meet sample size requirements (14), (33), complementing a small probability sample with a larger non-probability one may improve the efficiency of the estimates (40). Non-probability samples may also be designed as a mean for compensating for known biases in probability-based web survey samples by purposely targeting respondent profiles that tend to be underrepresented in these surveys. This is the case in the Survey on the impact of the COVID-19 pandemic in Spain (ESPA-COV) that motivates this paper (31).

Survey statisticians have provided different methods for combining information from multiple data sources. Current reviews of statistical methods of data integration for finite population inference can be seen in (37), (3), (29). Among the most important methods we could mention inverse probability weighting (23; 24), mass imputation (32), doubly robust methods (8), kernel smoothing methods (39) or statistical matching combined with PSA (5). (42) provide a good review of some of these techniques. Most of these works assume that the variable of interest is only available in the non-probability sample, whereas other auxiliary variables are present in both data sources. However, as described above, there are other scenarios where both the probability and non-probability-based samples share the same questionnaire and measures, meaning that it is possible to combine both of them in order to maximize the efficiency of the estimates.

Most surveys that integrate probability and non-probability samples simply pool the samples and make inference using the Horvitz-Thompson or Hàjek estimator assuming the entire sample is probabilistic (31). This method is rarely appropriate because usually non-probability samples are not distributed proportionally with respect to demographic or other relevant subgroups in the population. Some efforts have been undertaken to combine both probability and non-probability samples to make inference while dealing with the different sources of bias. (15) studies a composite estimator that is a linear combination of an unbiased sample mean estimate from a probability sample and a biased sample mean estimate from a convenience sample. The weight of the mean estimator based on the probability sample is determined by the ratio of its Mean Squared Error (MSE) to the sum of that term and the MSE of the convenience sample mean. (14) propose an alternative procedure using calibration. These authors combine the previously calibrated probability sample with the non-probability sample and then recalibrate overall to the probability samples benchmarks from the previous step. Their simulation study shows that calibrating non-probability samples with probability samples using early adopter questions minimizes bias in the resulting estimates in the larger combined sample. Recently (33) proposed weighting techniques that enable the two datasets to be analysed as a single one (i.e., a blended sample) by assuming four conditions for the probability and non-probability samples. Authors consider four separate methods for blending based on propensity score methods or on calibration weighting and warn on the challenges of integrating both kind of samples. Finally, (35) propose a Bayesian approach to combine information from probability and non-probability samples. Data from the non-probability sample is used to build an informative prior distribution that is subsequently used to inform the estimates from the probability sample. The simulation study and the application with real data suggests that resulting Bayesian estimates are more efficient than estimates exclusively based in probability samples, even when their sample sizes are quite small.

In this paper we explore other alternatives that combine some of these ideas with the help of machine learning methods. Our main contributions to this area of research are the development of a new estimation method for integrating data from probability and non-probability samples in those situations where the variables of interest are observed in both samples. We then assess the efficiency of the resulting estimates

by comparing them with other strategies that have been used before. The application of this method to the second wave of the Survey on the impact of the COVID-19 pandemic in Spain allows us to conclude that the estimation method that we propose is the best option for reducing observed biases in our data.

This article is structured as follows. Section 2 introduces the ESPACOV II survey that is our motivating case study. Section 3 establishes notation and describes the proposed methods for integrating probability and non-probability samples. Section 4 reports the results of an extensive simulation study run on a set of synthetic populations in which the performance of the proposed estimators is analysed for finite size samples. The proposed methods are applied in a real-world scenario in Section 5. Finally, the implications of our findings are discussed in Section 6.

2 Motivating case study

This new estimation technique was designed to analyse the data obtained in a web survey on the effects of the COVID-19 pandemic in Spain (ESPACOV Survey) that used a mixed multiphase sampling design inspired by the responsive approach (18). This survey was designed, implemented and funded by the Institute for Advanced Social Studies at the Spanish National Research Council (IESA-CSIC) (31). There were two editions of the survey: the first one was fielded from April 4th through April 11th in the fourth week of the lockdown, that in Spain began on March 14th. The second edition was conducted from January 18th to 25th, almost one year into the pandemic. Both questionnaires addressed the opinions and attitudes of the Spanish population regarding the COVID-19 crisis, as well as the assessments of its management and its consequences, either anticipated (ESPACOV I) or endured (ESPACOV II)[1]

Both editions of the ESPACOV Survey were web based and followed a sampling design that combined the use of SMS invitations to take part in the survey -sent to a list of randomly generated mobile phone numbers- (probability-based sample) with the publication of Facebook, Instagram and Google Ads segmented to purposely oversample the socio-demographic profiles that were underrepresented in the probability-based sample (non-probability sample). In the first edition of the ESPACOV survey both sampling procedures were applied sequentially so that the outcomes of the probability-based sample informed the design of the purposive sample. In the second edition both samples were fielded simultaneously taking advantage of the knowledge acquired in the previous edition. An in-depth explanation and justification of this methodology is provided in (31).

This article focuses on the measurement of the direct impact of the COVID-19 pandemic in terms of infection and severity of the disease and the consequences of the pandemic on the overall physical and mental health self-perception as well as the economic situation in the respondents households. For that reason, we use data from the second edition of the survey that allows to assess the situation almost one year after the beginning of this major health crisis.

66,439 SMS invitations with links to the questionnaire were sent in January 2021 for the probability-based sample in this edition, of which 51.3% were delivered. The effective sample size after eight days in fieldwork was n=973 (2.97% of delivered SMS). Invitations to complete the survey were advertised via Facebook, Instagram and Google ads from January, 18th to 22nd. The invitation reached 1,054,301 impressions and 7,647 clicks for a total number of 671 completed interviews.

As shown in table 1, this survey design partially accomplished its aim of maximizing representativeness of the Spanish resident population aged 18 and more. The distribution of the unweighted blended survey is more similar to the population than those of the individual samples (with the exception of gender). Moreover, the profiling of ads worked as intended oversampling respondents aged 65 and more and reducing,

[1] The research data and related documentation of both editions of the survey can be retrieved at the Spanish Research Council institutional repository: https://digital.csic.es/handle/10261/211271 (ESPACOV I) and https://digital.csic.es/handle/10261/233224 (ESPACOV II).

although less than needed, the proportion of employed respondents and those with higher education. Contrary to expectations, the profiling resulted in a significant overrepresentation of women in the blended sample.

Table 1 Population Data Sources: [1]Continuous population register, official population data as of 1^{st} January 2021 and [2]Economically active population survey (EAPS), 1^{st} quarter 2021. National Statistics Institute of Spain (INE).

		Probability	Non-probability	Blended Sample (Unweighted)	Population
Gender[1]	Male	48.4%	40.7%	45.3%	48.5%
	Female	51.6%	59.3%	54.7%	51.5%
Age[1]	18 to 29	18.2%	3.3%	12.1%	15.0%
	30 to 44	33.0%	15.8%	26.0%	25.4%
	45 to 64	41.3%	37.4%	39.7%	35.9%
	65 or more	7.5%	43.5%	22.2%	23.7%
Age (mean)		45	61	52	51
Education level[2]	First degree	21.0%	20.7%	20.9%	17.1%
	Second degree	18.7%	26.1%	21.7%	49.1%
	Higher ED	60.3%	53.2%	57.4%	33.8%
Labour Status[2]	Employed	69.2%	41.3%	57.8%	48.5%
	Unemployed	9.1%	6.4%	8.0%	9.2%
	Inactive	21.7%	52.3%	34.2%	42.3%

The next section develops the methods followed for correcting biases in both probability and non-probability samples and blending the data so that they can be analysed as a single dataset.

3 Methods

3.1 Context and survey design

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. Let s_r be a probability sample of size n_r selected from U under a probability sampling design (s_r, p_r) with $\pi_i = \sum_{s_r \ni i} p_r(s_r)$ the first order inclusion probability for individual i. Let s_v be a non-probability (volunteer) sample of size n_v, self-selected from U. Let y be the variable of interest in the survey estimation and let $\mathbf{x}_i$ be the values presented by individual i for a vector of covariates $\mathbf{x}$. The variable of interest and the covariates have been measured in both samples.

The population total, Y, can be estimated via the Horvitz-Thompson estimator:

$$\hat{Y}_R = \sum_{i \in s_r} \frac{y_i}{\pi_i} = \sum_{i \in s_r} d_i y_i \tag{1}$$

which is design-unbiased of the population total if there is not lack of response. The design-based variance of this estimator is given by:

$$V_p(\hat{Y}_R) = \sum_{i,j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \left(\pi_{ij} - \pi_i \pi_j\right). \tag{2}$$

where π_{ij} are the second order probabilities of the sampling design p_r.

Y can be also estimated with the naive estimator based on the sample mean of y in s_v:

$$\hat{\overline{Y}}_v = N \sum_{i \in s_v} \frac{y_i}{n_v} \tag{3}$$

If the convenience sample s_V suffers from selection bias, this estimator will provide biased results. This can happen if there is an important fraction of the population with zero chance of being included in the sample (coverage bias) and if there are significant differences in the inclusion probabilities among the different members of the population (selection) (10; 16).

3.2 Estimating propensities in the non probability sample

In this context, Propensity Score Adjustment (PSA) can be used to reduce the selection bias that would affect the unweighted estimates. This approach aims to estimate the propensity of an individual to be included in the non-probability sample by combining the data from both samples, s_r and s_v.

Propensity scores, π_{vi}, can be defined as the propensity of the i-th individual of participating in the survey, this is, the probability that $I_{vi} = 1$, being I_{vi} the indicator variable for unit i being included in the sample s_v

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \in U - s_v \end{cases} \quad , \quad i = 1, ..., N. \tag{4}$$

PSA assumes that the selection mechanism of s_V is ignorable and follows a parametric model:

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

for some known function $m(\cdot)$ with second continuous derivatives with respect to an unknown parameter γ.

The procedure is to estimate the propensity scores by using data of both, the volunteer and the probability sample. The maximum likelihood estimator (MLE) of π_{vi} is $m(\hat{\gamma}, \mathbf{x}_i)$ where $\hat{\gamma}$ maximizes the log-likelihood function:

$$l(\gamma) = \sum_U (I_{vi} log(m(\gamma, \mathbf{x}_i)) + (1 - I_{vi}) log(1 - m(\gamma, \mathbf{x}_i)) =$$

$$\sum_{s_v} log \frac{m(\gamma, \mathbf{x}_i)}{1 - m(\gamma, \mathbf{x}_i)} + \sum_U log(1 - m(\gamma, \mathbf{x}_i)). \tag{6}$$

As is usual in survey sampling, we consider the pseudo-likelihood since we do not observe all units in the finite population:

$$\tilde{l}(\gamma) = \sum_{s_v} log \frac{m(\gamma, \mathbf{x}_i)}{1 - m(\gamma, \mathbf{x}_i)} + \sum_{s_p} \frac{1}{\pi_i} log(1 - m(\gamma, \mathbf{x}_i)). \tag{7}$$

Once the MLE of π_{vi} has been obtained, we transform the estimated propensities $\hat{\pi}_{vi} = m(\hat{\gamma}, \mathbf{x}_i)$ to weights by inverting them (37) and obtain the inverse probability weighted IPW estimator:

$$\hat{Y}_{IPW} = \sum_{i \in s_V} y_i / \hat{\pi}_{vi} = \sum_{i \in s_V} y_i d_{vi} \tag{8}$$

The properties of the IPW estimators (under both the model for the propensity scores and the survey design for the probability sample) are developed in (8). These authors prove that under certain regularity

conditions and assuming the logistic regression model for the propensity scores, the IPW estimator $\hat{Y}_{IPW}$ is asymptotically unbiased for the population total ($\hat{Y}_{IPW} - Y = O_p(n_v^{-1/2})$) and they obtain an asymptotic expression for its variance:

$$V(\hat{Y}_{IPW}) = \sum_{i=1}^{N} (y_i/\hat{\pi}_{vi} - \mathbf{b}_1^T \mathbf{x}_i)^2 (1 - \hat{\pi}_{vi})\hat{\pi}_{vi} + \mathbf{b}_1^T D \mathbf{b}_1 \tag{9}$$

where $\mathbf{b}_1^T = \sum_{i=1}^{N} (1 - \hat{\pi}_{vi}) y_i \mathbf{x}_i^T / \sum_{i=1}^{N} \hat{\pi}_{vi}(1 - \hat{\pi}_{vi}) \mathbf{x}_i \mathbf{x}_i^T$, and $D = V_p(\sum_{i \in s_r} d_i \hat{\pi}_{vi} \mathbf{x}_i)$ where V_p denotes the design-based variance under the sampling design p.

3.3 Combining the probability and the non-probability samples

We are going to consider the situation in which there are no coverage biases in either the probability or the non-probability sample. Let U_r and U_v be two sampling-frames, in this situation U_r and U_v coincide with the population under study U.

A simple estimator is calculated by weighting the estimators obtained from each sample:

$$\hat{Y}_{com} = \alpha \hat{Y}_R + (1 - \alpha)\hat{Y}_{IPW} \tag{10}$$

where α is a nonnegative constant such that $0 \leq \alpha \leq 1$.

We denote the values of the variance of $\hat{Y}_R$ and the mean squared error of the estimator of $\hat{Y}_{IPW}$ by V_1, V_2 respectively. Since frames U_r and U_v are sampled independently, the MSE of $\hat{Y}_{com}$ is given by

$$MSE(\hat{Y}_{com}) = \alpha^2 V_1 + (1 - \alpha)^2) V_2 \tag{11}$$

where the first component of the right hand side is computed under the sampling design p_R and the second one under the selection mechanism model.

Next, we consider the problem of selection of the best coefficients. The value of α that minimizes the variance in (11) is given by

$$\alpha = \frac{V_2}{V_1 + V_2} \tag{12}$$

but the values V_1 and V_2 are unknown. One possibility is to estimate them from the sample and substitute them in the previous expression to calculate the coefficients α_o. Other solutions are to weight each estimator by the weight that sample has in the total sample $\alpha_n = n_r/(n_r + n_v)$ or $\alpha_e = 0.5$.

The resulting estimator 10 can be rewritten as:

$$\hat{Y}_{com} = \sum_{i \in s} y_i d_i^{\star} \tag{13}$$

being $s = s_r \bigcup s_v$ and

$$d_i^{\star} = \begin{cases} \alpha d_i & \text{if } i \in s_r \\ (1 - \alpha) d_{vi} & \text{if } i \in s_v \end{cases} \tag{14}$$

Besides the modification of weights for handling selection bias, other adjustments may also be carried out to take into account auxiliary information. Calibration (12) is the most used technique for weights adjustment, aiming at ensuring consistency among estimates of different sample surveys and some improving

the precision of estimators ((34), (13)). Calibration weighting was previously used in this context by (14) who proposes calibrating auxiliary information in the non-probability sample with that in the probability sample, so that after calibration the weighted distribution of the non-probability sample is similar to that of the target population.

Using the calibration paradigm, we wish to modify, as little as possible, basic weights $d_i^\star$ to obtain new weights $w_i^\star$, for $i \in s$ to account for auxiliary information and derive a more accurate estimation of the total Y. Let $\underline{z}_i = (z_{1i}, \ldots, z_{pi})$ be the value taken on unit i by a vector of auxiliary variables $\underline{z}$ of which we assume to know the population total $\underline{t}_z = \sum_{k=1}^{N} \underline{z}_k$ and that is available for the units of each sample. The vector of calibration variables $\underline{z}_i$ does not have to match the vector $\mathbf{x}$ used in the propensity model.

A general calibration estimator can be defined as

$$\hat{Y}_{CAL} = \sum_{k \in s} w_i^\star y_k \tag{15}$$

where $w_i^\star$ is such that

$$\min \sum_{k \in s} G(w_i^\star, d_i^\star) \qquad \text{s.t.} \qquad \sum_{k \in s} w_i^\star \underline{z}_k = \underline{t}_z, \tag{16}$$

where $G(w, d)$ is a distance measure satisfying the usual conditions required in the calibration paradigm. Given the set of constraints, different calibration estimators are obtained by using alternative distance measures. If we take the Euclidean type of distance function $G(w_i^\star, d_i^\star) = (w_i^\star - d_i^\star)^2/2d_i^\star$, we can obtain an analytic solution that produces the linear calibration estimator:

$$\hat{Y}_{\text{CAL}} = \sum_{i \in s} w_i^\star y_i. \tag{17}$$

The asymptotic properties of this calibration estimator can be obtain by adapting the asymptotic framework of (20), to the case of the dual-frame finite population as in (28).

3.4 Using Machine Learning Techniques

Logistic models are often used to estimate the propensity to participate in the survey of each individual. In recent decades, numerous machine learning (ML) methods have been considered in the literature for the treatment of non-probability samples and have proved to be more suitable for regression and classification than linear regression methods (17; 4; 9; 21).

Among the most important ML methods are boosting algorithms. Boosting algorithms have been applied in propensity score weighting (25; 26) showing on average better results than conventional parametric regression models. A common machine learning algorithm under the Gradient Boosting framework is XGBoost (7). Given its theoretical advantage over Gradient Boosting, which could lead to even better results in a broader range of situations (6), we propose the use of this method for estimating propensities that will be used to define the estimators previously proposed.

However, the importance of choosing the right hyperparameters has also been underlined for the proper functioning of the algorithm. Therefore, a grid search of the optimal parameters is also performed before training. Each considered set is validated with cross-validation (30). The grid includes the following hyperparameters:

(i) Maximum depth: The depth limit which is applied to each tree forming the ensemble. The considered values are 1, 2 and 3.

(ii) Number of rounds: The number of boosting iterations which are computed. The considered values are 50, 100 and 150.

(iii) Learning rate: A step size shrinkage rate used in order to avoid overfitting. The considered values are 0.3 and 0.4.

(iv) Colsample by tree: The ratio of variables considered when training the trees. The variables are chosen by simple random sampling independently for each tree. The considered values are 0.6 and 0.8.

(v) Subsample: The ratio of training data considered by simple random sampling at each boosting iteration. The considered values are 0.5, 0.75 and 1.

4 Simulation study

We carry out a simulation study to see which of the proposed estimators works best.
We simulate a population of size 500,000 in which we have two target variables y_1, and y_2 and eight auxiliary variables to perform the PSA algorithms and the calibration, $x_1, \ldots, x_8$. Four variables (x_1, x_3, x_5, x_7) follow a Bernouilli distribution with $p = 0.5$ and four others (x_2, x_4, x_6, x_8) follow Normal distributions with a standard deviation of one and a mean parameter dependent on the value of the previous Bernoulli variable for each individual

$$\begin{aligned} x_{1i}, x_{3i}, x_{5i}, x_{7i} &\sim B(0.5), \quad i \in U \\ x_{ji} &\sim N(\mu_{ji}, 1), \quad i \in U, j = 2, 4, 6, 8 \\ \mu_{ji} = \begin{cases} 2, & \text{if } x_{(j-1)i} = 1 \\ 0, & \text{if } x_{(j-1)i} = 0 \end{cases}, &\quad i \in U, j = 2, 4, 6, 8 \end{aligned} \tag{18}$$

The target variables were simulated as follows:

$$\begin{aligned} y_{1i} &= N(10, 4) + 5\pi_i, \quad i \in U \\ y_{2i} &= N(10, 4) + 2(x_{7i} = 1) - 2(x_{7i} = 0) + x_{8i} + 5\pi_i, \quad i \in U \end{aligned} \tag{19}$$

2,000 iterations are carried out and in each one of them we draw a probability sample of size $n_P = 250$ and a non-probability sample of sizes $n_{NP} = 500; 1,000; 2,000$. The probability sample is drawn by simple random sampling without replacement (SRSWOR) from the full population. The non-probability sample is drawn according to an unequal probability sampling design where π is the vector of inclusion probabilities. This probability was made dependent on x_5, x_6, x_7 and x_8 (which allowed the experiment to cover Missing At Random situations) as:

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = -0.5 + 2.5(x_{5i} = 1) + \sqrt{2\pi}x_{6i}x_{8i} - 2.5(x_{7i} = 1), \quad i \in U. \tag{20}$$

The evaluated estimators are the following:

(i) Reference estimator ($\hat{Y}_{REF}$): the two samples are joined and calibration is performed to obtain the final estimator.

(ii) Elliott and Haviland estimator ($\hat{Y}_{EH}$): we join the probabilistic and non-probabilistic sample and obtain the final estimator using the formulas proposed in the article by Elliott and Haviland (2007).

(iii) Based on the article by (33) we calculate four estimators:

  (i) the disjoint propensity score (DPS) weights estimator (section 2.1.1. of (33)): $\hat{Y}_{RDR1}$

(ii) the simultaneous weights estimator (section 2.1.2. of (33)): $\hat{Y}_{RDR2}$

(iii) the disjoint calibration (DC) weights estimators (section 2.2 of (33)):$\hat{Y}_{RDR3}$

(iv) the combined calibration estimator (section 2.2 of (33)) $\hat{Y}_{RDR4}$

(iv) Propensities estimator ($\hat{Y}_{PPSA}$): the probability and non-probability sample propensities are obtained, both samples are merged and calibration is performed to obtain the final estimator using the inverse of propensities as initial weights.

(v) Calibration - PSA estimator ($\hat{Y}_{CPSA}$): calibration is performed in the probability sample and in the non-probability sample we calculate the propensities. To obtain the final estimator, we combine them in several ways, considering $\alpha_{0.5}$, α_n and α_0 and them we do the calibration. We will denote these estimators $\hat{Y}_{CPSA-0.5}$, $\hat{Y}_{CPSA-n}$ and $\hat{Y}_{CPSA-\alpha_0}$.

In all the estimators in which the propensities are calculated we use both XGBoost and logistic regression methods to see if there are differences in the results derived from the classification method used. We use x for XGBoost and l for logistic regression in the subscripts to distinguish among methods.

The procedure is repeated across 500 iterations, and finally the Relative Bias (RB) and Root Mean Square Error (RMSE) is obtained for each method.

$$\begin{aligned} RB &= \frac{1}{B}\frac{\sum_{i=1}^{B}\hat{Y}_i - Y}{Y} * 100 \\ RMSE &= \sqrt{\frac{1}{B}\left(\frac{\sum_{i=1}^{B}\hat{Y}_i - Y}{Y}\right)^2} * 100 \end{aligned} \tag{21}$$

where B is the number of iterations.

In Tables 2 and 5 values of RB and RMSE can be seen for each of the proposed estimators.

It can be observed that calibration in both samples is not enough to completely remove selection bias, although this approach provides smaller RB and RMSE than other methods. The method proposed by (15) is vastly efficient at removing part of the selection bias that exists in the simulation data, where the selection mechanism of the non-probability sample could be considered Missing At Random.

The combination of calibration and PSA (propensity weights are used as base weights in calibration) reduces bias and RMSE, particularly when the algorithm used in PSA is XGBoost, although the advantage of this algorithm vanishes as the non-probability sample size increases.

The behavior of the estimators considered in (33) is very diverse. In some cases, particularly $\hat{Y}_{RDR1}$, the relative bias is even larger than the case where only calibration is used. This could happen because some of the assumptions made for these estimators do not apply in our simulation study. On the other hand, $\hat{Y}_{RDR2}$ and $\hat{Y}_{RDR3}$ are able to reduce RB and RMSE in comparison to $\hat{Y}_{REF}$, as long as XGBoost is used; in fact, they seem to be particularly sensitive to the algorithm used for propensity estimation.

Finally, the behavior of the proposed estimators $\hat{Y}_{CPSA}$ depends on the factor used in weighting. The best estimator in our simulator has been, by a huge margin, $\hat{Y}_{CPSA-\alpha_0}$, which is the estimator that weights the samples by the MSE.

It is worth mentioning that the results on RB and RMSE are very similar between methods and sample sizes. This can be explained by the fact that the target variables y_1 and y_2 have a very similar behavior, only varying because of the relationship between y_2 and the variables x_7 and x_8. The Missing At Random nature of both variables, which means that the ignorability assumption of PSA applies in this study, explains why the application of adjustment methods can lead to substantial reductions in bias and RMSE.

Table 2 Target variable y_1

	n_P=250, n_{NP}=500		n_P=250, n_{NP}=1000		n_P=250, n_{NP}=2000	
	RB	RMSE	RB	RMSE	RB	RMSE
$\hat{Y}_{REF}$	4.0089	4.2325	3.9975	4.1746	4.0150	4.1862
$\hat{Y}_{EH}$	1.8829	2.3656	1.7167	2.2035	1.8847	2.3331
$\hat{Y}_{PPSA-x}$	2.0657	2.5444	2.6950	3.3292	3.6646	4.4649
$\hat{Y}_{PPSA-l}$	3.6414	3.9105	3.6438	3.8654	3.6862	3.8695
$\hat{Y}_{RDR1-x}$	4.9785	5.1669	7.6796	7.7483	9.4612	9.4882
$\hat{Y}_{RDR1-l}$	4.9785	5.1669	7.6796	7.7483	9.4612	9.4882
$\hat{Y}_{RDR2-x}$	2.2063	2.6275	2.6684	3.1008	2.9175	3.2960
$\hat{Y}_{RDR2-l}$	5.3830	5.5579	7.2005	7.2753	8.3981	8.4328
$\hat{Y}_{RDR3-x}$	3.0664	3.4212	2.6377	2.9739	1.9837	2.3745
$\hat{Y}_{RDR3-l}$	4.3508	4.5603	4.3532	4.5194	4.3759	4.5263
$\hat{Y}_{RDR4-x}$	3.3407	3.7062	3.1919	3.5298	2.2949	2.7083
$\hat{Y}_{RDR4-l}$	5.2964	5.4696	6.6016	6.6844	7.5067	7.5478
$\hat{Y}_{CPSA-0.5-x}$	3.2861	3.6344	2.8010	3.1304	2.1006	2.4880
$\hat{Y}_{CPSA-0.5-l}$	4.3928	4.6010	4.3621	4.5244	4.3960	4.5426
$\hat{Y}_{CPSA-n-x}$	4.3127	4.6285	4.4216	4.7102	3.4650	3.8981
$\hat{Y}_{CPSA-n-l}$	5.8100	5.9651	6.9818	7.0578	7.7466	7.7861
$\hat{Y}_{CPSA-\alpha_0-x}$	1.7371	2.1937	1.6253	2.0568	1.6617	2.0761
$\hat{Y}_{CPSA-\alpha_0-l}$	1.7677	2.2334	1.6677	2.1159	1.7663	2.2183

5 Application to a survey on the social effects of COVID-19 in Spain

In this section, we apply the estimation method proposed in section 3, $\hat{Y}_{CPSA}$, to several variables that assess the impact of the COVID-19 in Spain and compare the results with the measurements obtained when both probability and non-probability-based samples are merged and calibrated to correct observed deviations from population benchmarks in relevant socio-demographic variables, $\hat{Y}_{REF}$.

Tables 2 and 5 show the outcomes of several variables measuring these direct and indirect effects of the pandemic in Spain considering probability and non probability-based samples separately as well as the integrated file using the estimation methods described above. The variables analyzed are the following:

(i) COVID-19 infection (respondent) (V1)

(ii) COVID-19 infection (close relatives) (V2)

(iii) Severity of infection No symptoms (V3)

(iv) Severity of infection Mild symptoms (V4)

(v) Severity of infection Serious symptoms (V5)

(vi) Severity of infection Hospital admission (V6)

(vii) Self-assessed health status (V7)

(viii) Mood self-assessment (V8)

(ix) Household income decreased as a result of COVID-19 pandemic (V9)

Table 3 Target variable y_2

	n_P=250, n_{NP}=500		n_P=250, n_{NP}=1000		n_P=250, n_{NP}=2000	
	RB	RMSE	RB	RMSE	RB	RMSE
$\hat{Y}_{REF}$	4.4896	4.6836	4.4295	4.5954	4.4039	4.5376
$\hat{Y}_{EH}$	2.2969	2.8396	2.2355	2.8294	2.1261	2.6716
$\hat{Y}_{PPSA-x}$	2.1424	2.6954	2.7771	3.4387	3.6646	4.4232
$\hat{Y}_{PPSA-l}$	3.0467	3.4685	2.9805	3.4032	2.9749	3.3381
$\hat{Y}_{RDR1-x}$	6.4868	6.6735	9.3444	9.4275	11.1084	11.1396
$\hat{Y}_{RDR1-l}$	6.4868	6.6735	9.3444	9.4275	11.1085	11.1397
$\hat{Y}_{RDR2-x}$	2.1913	2.6276	2.5367	3.0326	2.7151	3.1543
$\hat{Y}_{RDR2-l}$	5.4169	5.7445	7.6562	7.8112	9.0623	9.1459
$\hat{Y}_{RDR3-x}$	3.2944	3.6588	2.8563	3.2279	2.0788	2.5324
$\hat{Y}_{RDR3-l}$	4.0923	4.3647	4.0599	4.3128	3.9509	4.1784
$\hat{Y}_{RDR4-x}$	3.5544	3.8934	3.3697	3.7363	2.2891	2.6838
$\hat{Y}_{RDR4-l}$	4.9899	5.2172	6.1594	6.2886	6.9170	6.9871
$\hat{Y}_{CPSA-0.5-x}$	3.4837	3.8161	2.9433	3.3008	2.0841	2.4774
$\hat{Y}_{CPSA-0.5-l}$	4.1404	4.4017	4.0628	4.2905	4.0024	4.1855
$\hat{Y}_{CPSA-n-x}$	4.5550	4.8456	4.6472	4.9614	3.4805	3.9333
$\hat{Y}_{CPSA-n-l}$	5.4627	5.6568	6.5150	6.6299	7.1407	7.2063
$\hat{Y}_{CPSA-\alpha_0-x}$	1.7518	2.2293	1.7182	2.1520	1.6122	2.0024
$\hat{Y}_{CPSA-\alpha_0-l}$	1.7717	2.2576	1.7495	2.1984	1.7052	2.1117

The main differences between both samples in the survey are the infection rate and the severity of the disease. The proportion of respondents that have suffered the infection is more than three points higher in the probability-based sample. Also, hospitalisation seems to be less likely for COVID-19 patients in this sample, although the difference is not statistically significant. Both trends may be explained by the differences in the age structure of both samples, with the non-probability sample being considerably older than the probability-based (61 vs. 45). Those differences are consistent with what we already know regarding disease severity, with elders more at risk of developing serious illness, and compliance with COVID-19 preventative measures (41).

Similarly, the age distribution of samples would explain the difference, statistically significant, on the assessment of the impact of the pandemic on household income. This evaluation is considerably worse in the probability-based sample where the weight of employment incomes is most important. In all these cases, the estimator that seems to correct best the impact of the differences in age structure between both samples is the estimator that we develop in section 3. As shown in the simulation presented in the preceding section, this estimator is very efficient at reducing biases in the integrated dataset.

6 Conclusions

With more than 8 million official cases and almost 91,000 casualties as of mid-January 2022, Spain is one of the EU countries that has been worst affected by COVID-19. Spanish GDP declined by 10.8% in 2020 and working hours for the equivalent of 2 million of jobs were lost according to ILO data. Using a design that combines probability and non-probability-based sampling methods and proper estimation techniques, the second edition of the ESPACOV survey fully reflects the relevance of this impact. According to main survey estimators, 11% of the Spanish population had had COVID-19 and 30% had witnessed the infection

Table 4 Estimates of selected variables on the direct impact of COVID-19 in Spain from integrated data using a new estimation method based on calibration and XGBoost PSA ($\hat{Y}_{CAL-PSA}$) and direct calibration of the integrated sample ($\hat{Y}_{REF}$).

	Individual samples				Integrated sample			
	Probability		Non-probability		$Y_{CAL-PSA}$		Y_{REF}	
Variable	Estimation	CI	Estimation	CI	Estimation	CI	Estimation	CI
V1	0.1264	0.1066-0.1484	0.0939	0.0736-0.1177	0.1101	0.0862-0.1344	0.1220	0.1018-0.1422
V2	0.3001	0.2719-0.3295	0.2846	0.2515-0.3197	0.3038	0.2641-0.3436	0.2852	0.2577-0.3128
V3	0.1474	0.0871-0.2288	0.1458	0.0677-0.2651	0.2024	0.1197-0.2851	0.1860	0.1130-0.2589
V4	0.7158	0.6196-0.7990	0.6875	0.5484-0.8048	0.6745	0.5483-0.8006	0.6605	0.4838-0.8372
V5	0.1158	0.0631-0.1914	0.1042	0.0409-0.2133	0.0983	0.0340-0.1625	0.1041	0.0188-0.1895
V6	0.0211	0.0044-0.0658	0.0417	0.0088-0.1270	0.0194	0.0000-0.0579	0.0367	0.0000-0.0880

Table 5 Estimates of selected variables on indirect effects of COVID-19 in Spain from integrated data using a new estimation method based on calibration and XGBoost PSA ($\hat{Y}_{CAL-PSA}$) and direct calibration of the integrated sample ($\hat{Y}_{REF}$).

	Individual samples				Integrated sample			
	Probability		Non-probability		$Y_{CAL-PSA}$		Y_{REF}	
Variable	Estimation	CI	Estimation	CI	Estimation	CI	Estimation	CI
V7	0.0668	0.0524-0.0838	0.0760	0.0578-0.0979	0.0657	0.0449-0.0865	0.0750	0.0581-0.0919
V8	0.2754	0.2481-0.3041	0.2340	0.2031-0.2671	0.2705	0.2290-0.3122	0.2607	0.2336-0.2879
V9	0.4162	0.3856-0.4474	0.2996	0.2658-0.3350	0.3950	0.3583-0.4318	0.4026	0.3738-0.4314

of close relatives until January 2021, 10 months after the World Health Organization (WHO) declared the novel coronavirus (COVID-19) outbreak a global pandemic. Although the majority of those infections were asymptomatic or endured with mild symptoms (87.7%), the pandemic was taking a huge toll on the economy of families (39.5% declared that household income had decreased) and on mental wellbeing, with more than one in four (27%) assessing their mood as very bad or bad.

The estimates suggested in literature that could be applied to the data from this survey were based on the simple union of both samples. In this article we address the problem of how to improve these estimates. We introduce four methods for calculating weights that blend probability and convenience samples; these methods combine calibration and propensity score adjustment using machine learning techniques for those situations where the variables of interest are observed in both samples.

Before their application to the survey, we evaluate the behavior of the proposed estimators against other techniques for integrating probability and non-probability samples used in the literature. As in many simulation studies, the number of simulation conditions we have generated is limited. However, we considered a simulation study with several sample sizes to cover different Missing At Random situations and we compared the performance of standard logistic regression model with a machine learning algorithm (XGBoost) when estimating the propensity score. Our simulation study shows that the proposed estimator based on calibration and PSA techniques is very efficient at reducing self-selection bias and RMSE with this kind of data. In our simulations, the best performing techniques for the estimation of the propensity scores were those based on boosting, which guaranteed considerably lower bias and RMSE in comparison to a similar estimator based on logistic regression and other techniques considered in the study.

Before applying ML techniques we have considered hyperparameter tuning. The simulation proved that, in the context of integrating probability and non-probability data, tuning is data-dependent and therefore we strongly suggest that researchers consider tuning parameters before using ML techniques in this context.

Finally, based on the simulation results, we consider the use of the proposed estimator $\hat{Y}_{CAL-PSA-\alpha_0}$ (which is the estimator that weights the samples by the MSE) as an alternative to the usual estimators for the estimation of the effects of the COVID-19 pandemic in Spain. The application of this method to ESPACOV II Survey provides the best correction of the impact of the deviations from population parameters in both samples.

Acknowledgments

The authors would like to thank the Institute for Advanced Social Studies (IESA-CSIC) for providing data and information about the ESPACOV survey. This study was partially supported by Ministerio de Educación y Ciencia (PID2019-106861RB-I00, Spain), IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Econiómica, Industria, Conocimiento y Universidades (FQM170-UGR20, A-SEJ-154-UGR20).

Conflict of interest

The authors declare no potential conflict of interests.

References

[1] Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... and Tourangeau, R. (2013). Non-probability sampling: report of the AAPOR task force on nonprobability sampling. *American Association for Public Opinion Research*. Available from https://www.aapor.org/Education-Resources/Reports/Non-Probability-Sampling.aspx.

[2] Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, **78**(2), 161-188.

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

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

[5] Castro-Martín, L., Rueda, M.d.M. and Ferri-García, R. (2021) Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *J. Comput. Appl. Math.*, 113414. https://doi.org/10.1016/j.cam.2021.113414.

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

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

[8] Chen, Y., Li, P., and Wu, C. (2019). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, **115**(532), 2011-2021.

[9] Chu, K.C.K. and Beaumont, J.F. (2019) *The use uf classification trees to reduce selection bias for a non-probability sample with help from a probability sample.* In Proceedings of the Survey Methods Section: SSC Annual Meeting, Calgary, AB, Canada.

[10] Couper, M. (2011) *Web Survey Methodology: Interface Design, Sampling and Statistical Inference.* Instituto Vasco de Estadística (EUSTAT): Vitoria-Gasteiz, Spain.

[11] Couper, M. P., Dever, J. A., and Gile, K. J. (2013). Report of the AAPOR task force on non-probability sampling. *Retrieved November*, **8**.

[12] Deville J.C. and Särndal C.E.. (1992). Calibration estimators in survey sampling. *J Am Stat Assoc*, **87**(418), 376-382.

[13] Devaud, D. and Tillé, Y. (2019). Deville and Särndal's calibration: revisiting a 25-year-old successful optimization problem. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, Springer; Sociedad de Estadística e Investigación Operativa, **28**(4), 1033-1065.

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

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

[16] Elliott, M.R. and Valliant, R. (2017) Inference for nonprobability samples. *Stat. Sci.* **32** , 249–264.

[17] Ferri-García, R. and Rueda, M.d.M. (2020) Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE* **15** , e0231500.

[18] Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(3), 439-457.

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

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

[21] Kern, C., Li, Y. and Wang, L. (2020) Boosted Kernel Weighting—Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, **9**(5),10881113. https://doi.org/10.1093/jssam/smaa028.

[22] Kohler, U. (2020). Survey Research Methods during the COVID-19 Crisis. *In Survey Research Methods*, **14**(2), 93-94.

[23] Lee, S. (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *J. Off. Stat.* **22**, 329-349.

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

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

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

[27] Lenau, S., Marchetti, S., Mnnich, R., Pratesi, M., Salvati, N., Shlomo, N., Schirripa Spagnolo, F. and Zhang, L.C. (2021) Methods for sampling and inference with non-probability samples. Deliverable D11.8, Leuven, InGRID-2 project 730998 H2020.

[28] Ranalli M.G., Arcos A., Rueda M.M. and Teodoro A. (2016) Calibration estimation in dual-frame surveys. *Stat Method Appl* **25**(3),321-349.

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

[30] Refaeilzadeh, P., Tang, L. and Liu, H. (2009). *Cross-validation.* Encyclopedia of database systems. Springer. 532-538.

[31] Rinken, S., Domnguez-lvarez, J. A., Trujillo, M., Lafuente, R., Sotomayor, R., and Serrano-del-Rosal, R. (2020). Combined mobile-phone and social-media sampling for web survey on social effects of COVID-19 in Spain. *In Survey Research Methods* **14**(2), 165-170. https://doi.org/10.18148/srm/2020.v14i2.7733

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

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

[34] Rueda M.M., Martínez S., Martínez H. and Arcos A. (2006). Mean estimation with calibration techniques in presence of missing data. *Comput Stat Data An*, **50**(11),3263-3277.

[35] Sakshaug, J. W., Wisniowski, A., Ruiz, D. A. P. and Blom, A. G. (2019). Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. *Journal of Official Statistics (JOS)*, **35**(3), 653-681.

[36] Schaurer, I., and Wei, B. (2020). Investigating selection bias of online surveys on coronavirus-related behavioral outcomes. *In survey research methods* **14**(2), 103-108.

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

[38] Vehovar, V., Toepoel, V. and Steinmetz, S. (2016). Non-probability sampling. *The Sage handbook of survey methods*, **22**, 329-345.

[39] Wang, G.C. and Katki, L. (2020) Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *J. R. Stat. Soc.* **183**, 1293–1311.

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

[41] Wright, L. and Fancourt, D. (2021). Do predictors of adherence to pandemic guidelines change over time? A panel study of 22,000 UK adults during the COVID-19 pandemic. *Preventive Medicine*, **153**, 106713. https://doi.org/10.1016/j.ypmed.2021.106713

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

