# Statistical methods to improve estimates obtained from probability and nonprobability samples

UNIVERSIDAD
DE GRANADA

Doctoral Thesis

Ramón Ferri García
Thesis supervised by Prof. María del Mar Rueda García

Doctorate Program in Mathematical and Applied Statistics
University of Granada

Granada, May, 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ramón Ferri García
ISBN: 978-84-1306-930-2
URI: http://hdl.handle.net/10481/69645

# Contents

# Agradecimientos

*Al andar se hace el camino,*
*y al volver la vista atrás*
*se ve la senda que nunca*
*se ha de volver a pisar.*

Proverbios y cantares (XXIX). Antonio
Machado (1875-1939)

Hacer un listado de gente a la que uno está agradecido es un ejercicio intrínsicamente injusto, puesto que requiere una exhaustividad plena (es decir, que se pueda seleccionar a toda la población objetivo) que en la práctica no es posible alcanzar, puesto que la duración de unos estudios de doctorado es tan extensa (en mi caso, cinco años académicos) como repleta de acontecimientos significativos, mientras que una sección de Agradecimientos tiene una extensión demasiado limitada. Sin embargo, sería aún más injusto renunciar a esta oportunidad de acordarme de tanta gente sin la cual posiblemente no podría escribir estas líneas, así que vamos a intentarlo.

En primer lugar, quiero agradecer a mi directora María del Mar Rueda su apoyo y esfuerzo para que esta tesis saliera adelante en las circunstancias más adversas, así como por su dedicación, empatía y generosidad demostrada en todo momento. Además de su amplísimo dominio del campo de la metodología de las encuestas, ha demostrado ser la mejor mentora que se podría tener en el mundo de la investigación.

Quiero mencionar también en estas líneas al resto de miembros del grupo de investigación: Antonio Arcos, Luis Castro, Beatriz Cobo y David Molina. En concreto, me gustaría agradecer la buena voluntad que han tenido siempre por ayudarme cuando lo he necesitado y por buscar siempre lo mejor para el grupo.

Quiero dar también a todo el equipo de Statistics Canada que me permitió hacer la estancia internacional: Jean-François Beaumont, Keven Bosa, Joanne Charlebois y Kenneth Chu. Las tristes circunstancias pandémicas nos impidieron poder conocernos en persona, pero mi etapa como alumno virtual con ellos fue inmensamente prolífica en lo referido a la cantidad de estadística aprendida y a mi propia autoestima como investigador. I hope we could meet soon in Canada.

También quiero agradecer al departamento de Estadística e Investigación Operativa de la Universidad de Granada las oportunidades y el apoyo brindado durante estos años para introducirme en el mundo de la investigación y docencia universitaria, así como a mi compañero Christian Acal por acompañarme en todas las fatigas de esta ardua marcha y brindarme su apoyo en todo momento, además de las largas conversaciones sobre la otra gran pasión que compartimos: el fútbol.

Me gustaría igualmente mencionar a los jefes que tuve durante el tiempo que trabajé en el departamento de Ingeniería Civil de la Universidad de Granada, Francisco Javier Calvo y Juan de Oña. Les agradezco que me introdujeran el "gusanito"de la investigación y todo lo que aprendí en mi etapa allí.

Por último, y como pilar fundamental, quiero agradecer a todos mis seres queridos el aliento que me han dado estos años para continuar, muy especialmente en los momentos de zozobra. Gracias por hacerme sentir orgulloso de este logro, que es también vuestro.

# Summary

Since their theoretical development in the first half of the XXth century, surveys have been the standard procedure to obtain information from a population of interest. The statistical properties of the estimators of population parameters, such as totals, means or proportions, allow researchers to make inferences about a target population using only a reduced sample of it, as well as obtain a measure of the variability of the estimations.

The first surveys were administrated by directly interviewing the respondents in person, a mode known as face-to-face surveying. This administration mode has been considered the "gold standard"practice in surveys, but their increasing costs and the advances in communication technologies favored the rise of telephone surveys and self-administered questionnaires, such as those used in mail surveys.

In the last decades, these modes have also experienced an increase in costs and coverage problems, as well as a decline in response rates. Again, the development of new technologies has been the factor that has allowed the appearance of a new set of questionnaire administration techniques known as online surveys. Some examples include SMS surveys, e-mail surveys, smartphone surveys, and especially Web surveys, which are those that are administered and completed in web browsers.

Online surveys comprise many advantages for researchers to conduct their studies. Recruitment of participants can be done much faster than in other survey modes, and at largely reduced costs. In addition, the use of technology allows researchers to design questionnaires with a wider spectrum of possibilities than in face-to-face, telephone or mail surveys.

On the other hand, online surveys present several relevant sources of error. By definition, such surveys can only reach online users or people with some kind of access to information and communication technology networks. This is an important coverage issue that can lead to biased estimates if the composition of the offline population differs significantly from that of the online population, which is often the case as the differences are associated to demographics such as education level or age.

In addition, the impossibility to find any reliable sampling frame of the online population contributes to the use of self-selection procedures in online surveys. This practice constitutes an example of nonprobability sampling where the estimators of population parameters and their variance cannot be calculated because of the inability of inclusion probabilities to meet the requirements of a probability sampling. The main consequence of the application of these procedures is selection bias, which can be very relevant if there is any relationship between propensity to participate (self-select) in the survey and the variables of interest of the study.

In those cases where a sampling frame is available for an online survey, and therefore it is possible to design a sampling scheme, non-response bias is also prone to appear. This is a particularly relevant issue in online panel surveys, and it has been associated with factors such as questionnaire length, incentives or invitation reminders.

Some methods have been developed in survey methodology literature to address these issues. Non-response error is a common problem to all probability sampling surveys, and in consequence many methods have been developed to mitigate it, from which imputation and reweighting techniques can be pointed out. The correction of coverage and self-selection biases depends on the auxiliary information available. If only population totals for a set of covariates are available, calibration procedures can be applied; these have been proven to reduce coverage error, but their use in the correction of self-selection bias in online surveys is unclear.

In some cases, a probability survey of reference, conducted in the same target population, is available. The variable of interest has not been measured on it, but if some auxiliary covariates (also measured in the online survey) are available, some adjustments can be considered. The most remarkable ones are Propensity Score Adjustment (PSA) and Statistical Matching or Mass Imputation. These adjustments focus on the mitigation of self-selection bias.

Finally, if a population census is available for some auxiliary covariates (also measured in the online survey), methods based on superpopulation modeling can be considered, such as model-based, model-adjusted and model-calibrated estimators. These methods have been mostly considered in probability sampling contexts, although some recent works adapt some of them to nonprobability sampling problems.

To contribute with the development of online surveys, we propose some methodological advances, such as the development of estimators of general parameters and the estimator of their variance, the study of the properties of the combination of PSA and calibration, the use of modern prediction techniques and variable selection methods in PSA, and the adaptation of all the superpopulation modeling approaches to the nonprobability sampling context considering modern prediction techniques as well.

We also adapt the weight smoothing strategy, developed for increasing the efficiency of the estimators in multipurpose probability surveys, to the nonprobability sampling context. Adapting the weighting adjustments existent for such samples to multipurpose surveys could be the key to their adoptation in the production of official statistics or their inclusion in large-scale studies.

Finally, we use PSA in the study of health-related variables in healthcare professionals using data from an online survey as the main source of information and the population census as the reference sample. We compare the results to the unadjusted case and evaluate the performance of the aforementioned adjustment.

Note: This thesis is presented as a compendium of seven publications in relation with the contents of the thesis. The full version of the papers is included in Appendices A1 - A7.

# Resumen

Desde su desarrollo teórico en la primera mitad del siglo XX, las encuestas han sido el método estándar de obtención de información de una población de interés. Las propiedades estadísticas de los estimadores de parámetros poblacionales, como los totales, las medias o las proporciones, permiten a los investigadores hacer inferencia sobre una población objetivo utilizando únicamente una muestra reducida de ella, así como obtener una medida de la variabilidad de las estimaciones.

Las primeras encuestas fueron administradas entrevistando directamente a los encuestados en persona, un modo conocido como la encuesta cara a cara. Este modo de administración ha sido considerado como la práctica "gold standard.$^{en}$ encuestas, pero sus crecientes costes y los avances en las tecnologías de la comunicación favorecieron el surgimiento de encuestas telefónicas y cuestionarios autoadministrados, como los empleados en encuestas por correo.

En las últimas décadas, estos modos también han experimentado un incremento en costes y problemas de cobertura, así como un declive de las tasas de respuesta. De nuevo, el desarrollo de nuevas tecnologías ha sido el factor que ha permitido la aparición de un nuevo conjunto de técnicas de administración de cuestionarios conocido como las encuestas online. Algunos ejemplos incluyen las encuestas por SMS, las encuestas por e-mail, las encuestas por smartphone y especialmente las encuestas Web, que son aquellas que se administran y se completan en navegadores web.

Las encuestas online incluyen muchas ventajas para los investigadores de cara a realizar sus estudios. El reclutamiento de participantes puede ser realizado mucho más rápido que en otros modos de encuesta, y con costes ampliamente reducidos. Además, el uso de la tecnología permite a los investigadores diseñar cuestionarios con un espectro más amplio de posibilidades que en las encuestas cara a cara, telefónicas o por correo.

Por otra parte, las encuestas online presentan algunas fuentes de error relevantes. Por definición, estas encuestas sólo pueden llegar hasta usuarios online o personas con algún tipo de acceso a las redes de las tecnologías de la información y comunicación. Este es un importante problema de cobertura que puede traducirse en estimaciones sesgadas si la composición de la pobla-

ción offline difiere significativamente de la de la población online, lo que suele ser el caso dado que las diferencias están asociadas a variables demográficas como el nivel educativo o la edad.

Junto a ello, la imposibilidad de encontrar algún marco muestral fiable de la población online contribuye al uso de técnicas de autoselección en las encuestas online. Esta práctica constituye un ejemplo de muestreo no probabilístico donde la varianza no puede ser calculada por la imposibilidad de las probabilidades de inclusión de cumplir los requerimientos de un muestreo probabilístico. La principal consecuencia de la aplicación de estos métodos es el sesgo de selección, que puede ser muy relevante si existe alguna relación entre la propensión a participar (autoseleccionarse) en la encuesta y las variables de interés del estudio.

En aquellos casos en los que haya un marco muestral disponible para una encuesta online, y por tanto sea posible diseñar un esquema de muestreo, el sesgo de no respuesta también es proclive a aparecer. Este problema es particularmente relevante en las encuestas de paneles online, y ha sido asociado a factores como la longitud del cuestionario, los incentivos o los recordatorios de invitación. Se han desarrollado algunos métodos en la literatura para atajar estos problemas. El sesgo de no respuesta es un problema común a todas las encuestas probabilísticas, y en consecuencia se han desarrollado muchos métodos para mitigarlo, de los cuales se pueden destacar las técnicas de imputación y reponderación.

La corrección de los sesgos de selección y cobertura depende de la información auxiliar disponible. Si sólo están disponibles los totales poblacionales para un conjunto de covariables, se pueden aplicar métodos de calibración; se ha comprobado que éstos reducen el error de cobertura, pero su uso en la corrección del sesgo de autoselección en las encuestas online no está claro.

En algunos casos, una encuesta probabilística de referencia, llevada a cabo en la misma población objetivo, está disponible. La variable de interés no ha sido medida en ella, pero si hay disponibles algunas covariables auxiliares (también medidas en la encuesta online), se pueden considerar algunos ajustes. Los más conocidos son el Propensity Score Adjustment (PSA) y el Statistical Matching o Mass Imputation. Estos ajustes se centran en la mitigación del sesgo de selección.

Finalmente, si está disponible un censo de la población para algunas covariables auxiliares (también medidas en la encuesta online), se pueden considerar métodos basados en los modelos de superpoblación, como los estimadores modelo basado, modelo asistido y modelo calibrado. Estos métodos se han considerado principalmente en contextos de muestreo probabilístico, aunque algunos trabajos recientes adaptan algunos de ellos a problemas de muestreo no probabilístico.

Para contribuir al desarrollo de las encuestas online, proponemos algunos avances metodológicos, como el desarrollo de estimadores de parámetros

generales y el estimador de su varianza, el estudio de las propiedades de la combinación de PSA y calibración, el uso de técnicas modernas de predicción y selección de variables en PSA, y la adaptación de todos los métodos de modelos de superpoblación al contexto del muestreo no probabilístico considerando asimismo técnicas modernas de predicción.

Adaptamos también la estrategia de suavizado de pesos, desarrollada para incrementar la eficiencia de los estimadores en encuestas probabilísticas multipropósito, al contexto del muestreo no probabilístico. Adaptar los ajustes de ponderación existentes para estas muestras a las encuestas multipropósito podría ser la clave para adoptarlas en la producción de estadísticas oficiales o incluirlas en estudios a gran escala.

Finalmente, empleamos PSA en el estudio de variables relacionadas con la salud en profesionales sanitarios utilizando datos de una encuesta online como la principal fuente de información y el censo de la población como la muestra de referencia. Comparamos los resultados al caso sin ajustar y evaluamos el rendimiento del mencionado ajuste.

Nota: Esta tesis se presenta como un compendio de 7 publicaciones relacionadas con los contenidos de la tesis. La versión íntegra de los artículos se incluye en los Apéndices A1 - A7.

# Part I

# Chapter 1

# Introduction

Statistics, as the science of quantifying the characteristics present in real word, rely largely on survey sampling theory. Throughout history, the methods used to obtain figures from a population, such as total tax revenue or total number of crops available, were mainly censuses of the entire target population. According to Bethlehem (2009b), the first known application of a sample of individuals to make generalizations about a given population was done in the popular book *Natural and Political Observations Made upon the Bills of Mortality* by John Graunt in 1662. In one of its chapters, Grant attempted to estimate the total population of London using the total number of burials per year across the city, the estimated mean number of family members and the estimated annual number of deceases per family. By combining these figures in a very similar calculation to that of the ratio estimator used nowadays, he estimated a total population of 403,000 inhabitants in 1661 (see Chapter 7 in Hald (2003) for more details on Graunt's book). The rise of empirical sciences in the XIX$^{th}$ century saw an important increase in the use of samples for experimentation purposes. However, as noted in Bethlehem (2009b), those efforts did not account for any statistical sampling principles despite the earlier development of probability theory, although some surveys attempted to follow a sample design to ensure representativeness of the population but without accounting for sample variability or inclusion probabilities. It was not until the work of Neyman (1934) when a theoretical framework for probability sampling was established, which was extended in the following decades. In this sense, we can outline the work of Horvitz and Thompson (1952), which established that any sample with known and positive inclusion probabilities can be used to obtain unbiased estimates of population parameters. This basic definition of probability sample has withstood the test of time; it is the definition that will apply in this work when we refer to probability samples.

Along with the sampling theory, questionnaire administration modes were largely studied throughout the XX$^{th}$ century. Face-to-face interviews, which had been the usual method for sample recruitment, established themselves as the gold standard for representative survey sampling. However, the increasing costs of deploying such method, caused by an increase of refusal rates (see Díaz de Rada (2012) for a review on the matter) favoured the rise of telephone surveys in the second half of the century, along with methods that automatized their administration, such as Computer Assisted Telephone Interview (CATI). Telephone surveys allowed practitioners to put stratified sampling designs into practice, to administer more questionnaires at a lesser cost than face-to-face surveys, and to easily perform multiple contacts in the case of absent respondents (Díaz de Rada, 2012). However, in the last years some serious drawbacks have arisen. Firstly, the fraction of the population without a landline phone has grown critically because of the rise of mobile phone services. This behavior makes those people harder to reach, as telephone lists for mobile phone users rarely exist (in contrast to landline phone owners), and, if available, they are likely to entail coverage errors which may largely contribute to selection bias (see Pasadas-del-Amo (2012) for a review of the matter for the case of Spain). On the other hand, telephone surveys have experienced a notorious decrease in response rates. A study by Kohut et al. (2012) showed that telephone surveys performed by Pew Research Center dropped from 36 % in 1997 to 9 % in 2012. Another study by Marken (2018) showed that response rates for telephone surveys done in the context of the Gallup Poll Social Series dropped from 28 % in 1997 to 7 % in 2017. This decrease might lead to a higher cost per questionnaire and a decrease in the quality of the data.

At the same time, the rise of new information and communication technologies made it possible for practitioners to start using new forms of questionnaire administration. The increasing penetration of internet connections and smartphones across the globe made it possible to consider these communication devices as a potential channel of questionnaire delivery for a wider range of people, and to create the so-called online surveys. These surveys, following the definition from Vehovar and Manfreda (2008), are all of those that use any type of Information and Communications Technology (ICT) network during the survey process. In this wide category, surveys administered via SMS or in local networks in an organization (for instance, employees in a company) are included, but research has focused mainly on Internet and smartphone surveys, which share many principles with the surveys previously described. Internet surveys, following again the definition from Vehovar and Manfreda (2008), are all of those that can be administered and completed using Internet technology. In this category, web surveys (administered and completed in web browsers) and e-mail surveys can be included. Smartphone surveys are regarded in the literature as those that can be completed using a mobile phone device or tablet; however, there is an important divide (especially regarding data quality) between those completed in mobile phone web browsers but which may have not been optimized for them and smartphone app-based surveys or questionnaires designed specifically for being completed in mobile phone devices (Buskirk and Andrus, 2014; Callegaro, 2013; Couper and Peterson, 2017; Wells et al., 2014).

As mentioned above, the main extent of online surveys nowadays refers to smartphone and specially web surveys. Therefore, this dissertation will focus on them, although the conclusions can be extended to the other types of online surveys given that the advantages and inconveniences are caused by the same factors.

There are many strategies to perform web surveys. From the comprehensive list included in Callegaro et al. (2015) and summarized in Díaz de Rada et al. (2019), we could distinguish between probability and nonprobability online surveys. The possibilities for nonprobability sampling include limitless surveys with self-selected samples (which do not require a sampling frame), such as snowball sampling in social media websites, and samples based on lists of individuals who self-selected (also called opt-in panels) or do not cover the whole target population. It is worth mentioning that interception surveys, which çaptureïnternet users who are browsing a particular website and invite them to participate in the survey, are included as a probability sampling option in the list from Callegaro et al. (2015) and as a nonprobability sampling approach in the review by Schonlau and Couper (2017). According to the latter reference, the selection biases that may apply to this type of sampling are uncertain. Regarding probability sampling options, the possibilities include regular probability surveys targeted to online users, probability web surveys targeted to wider populations than internet users and probability-based web panels. The latter option has already been adopted by several statistical organizations around the world. From the review by Schonlau and Couper (2017), most of them rely on recruiting candidates via a probability survey with the objective of inviting them to be part of an online panel. In some cases, the panels offer free computer and internet access to those candidates who have no access to either of them.

Each of the described strategies has its own advantages. For most of them (except those involving offline recruitment), the most obvious one is the reduction in costs regarding interviewers and administration of questionnaires. Online surveys, except for special needs or cases such as online focus groups or chat rooms, are self-administered by the respondent itself, meaning that interviewers are not required and therefore saving the costs of preparing and displaying a team of interviewers to do the fieldwork. The inmediateness of internet connection itself also allows survey designers to save the costs of sending the questionnaire physically by post mail. Research showed that web surveys have much lower costs in comparison to mail surveys, which are also self-administered (Bech and Kristensen, 2009; Greenlaw and Brown-Welty, 2009; Díaz de Rada, 2012), and telephone surveys (Lee et al., 2019). On the other hand, it has been shown that online surveys can highly benefit from incentives, such as paying each respondent a small quantity of money or giving them the chance to win a prize, which increases the final budget (Bosnjak and Tuten, 2003; Göritz, 2006, 2014).

In addition, online surveys offer a substantial decrease in the time needed to achieve a given sample size, because of the smaller effort required to perform the fieldwork. The review done by Ilieva et al. (2002) in the early days of online surveys found that e-mail questionnaires were responded in half of the time that it took to respond mail questionnaires. A study done in three online panels by Reynolds et al. (2009) found that almost two thirds of the sample completed the questionnaire in the first 72 hours after launching the survey. These findings were corroborated in other studies, as mentioned by Díaz de Rada (2012). Moreover, the time required to complete an online questionnaire can be shorter than the time required to complete a questionnaire administered by an interview; for example, the comparison between CATI, web and smartphone respondents done by Lee et al. (2019) showed that web and smartphone respondents took two thirds of the time than CATI responders took to complete the questionnaire. However, this could also be a consequence of acquiescence or survey satisficing (Barge and Gehlbach, 2012), a phenomenon that drives respondents to answer the questions without paying the required attention to understand them, compromising the quality of the data.

Another substantial advantage of online surveys is the computerization of the questionnaire, which entails a wide spectrum of possibilities in terms of flux control, multimedia content, formulation of the question and its response options, etc. As noted in Díaz de Rada et al. (2019), the computerization also allows the introduction of mechanisms to prevent methodological problems in questionnaires, such as randomizing the order of the questions or response options [which has been shown to have an influence in the answers given by respondents (Tourangeau et al., 2004, 2013)], or programming alerts to warn users of the questions they left unanswered, in order to avoid partial nonresponse.

The computerization of the questionnaire is tied to another important advantage: the fact that online surveys are self-administered. Apart from the advantages of self-administration in terms of costs and time mentioned above, the absence of an interviewer might be positive regarding some effects that can bias the results, such as social desirability. Social desirability refers to the behavior which leads the respondent to answer the options that, according to their values, are more socially desirable or accepted to give the interviewer a good impression of them. This is a very common behavior in compromising questions (such as sexual behavior, use of drugs or criminal records) and some methods have been developed in face-to-face surveys to renforce the anonimity of the response (Cobo-Rodríguez, 2018). Their application in online surveys has casted some doubts (Coutts and Jann, 2011), but some studies have pointed out that social desirability in online surveys (both internet and smartphone surveys) is not as common as in face-to-face surveys (Heerwegh, 2009; Mavletova and Couper, 2013). However, a meta-study by Dodou and de Winter (2014) found that there is no effect associated to the administration of the questionnaire via online regarding social desirability, although the heterogeneity level is high and the number of studies is scarce.

As a final advantage, we could mention the fact that self-selection procedures in online surveys may be helpful in order to find members of non-demographic strata in a population. This can happen if the survey is released using the topic that is covering as an incentive to answer (Lehdonvirta et al., 2020); for example, a survey about soccer might make soccer fans more prone to take it so they can express their views. These surveys can be used to find respondents from hard-to-reach populations, at the cost of suffering the disadvantages of self-selection which will be described in the following lines.

Although each of the possibilities for administration of online questionnaire has its own disadvantages, there are several inconveniences which are common to all of them. For example, online surveys are particularly vulnerable to measurement error. The aforementioned informatization of the questionnaires can lead to programming bugs or incompatibility issues that can compromise the completion of the survey or the understanding of the questions by the respondents, who may eventually produce erroneous answers (Díaz de Rada et al., 2019; Elliott and Valliant, 2017). In addition, response error is also prone to appear in online surveys; although the absence of an interviewer can be seen as an advantage (as described in previous lines), it can also lead to undesirable behaviors such as survey satisficing. This behavior can appear in any type of survey, but it constitutes a particularly serious problem in online surveys because of the lack of motivation produced by the absence of an interviewer (Anduiza and Galais, 2017; Gao et al., 2016).

Out of the possible sources of error in online surveys, the most important one is selection bias. Generally speaking, selection bias in online surveys happens if, as described in Elliott and Valliant (2017), the characteristics of the sample obtained from the online survey differ significantly from those of the target population in a way that disallows any attempt to generalize the results of the sample. The term "selection bias.ºften refers to several sources of error following the definition by Smith (2020); in the context of online surveys, selection bias can entail non-response, coverage and self-selection (also named as volunteer) bias. Although these biases can be observed in other survey modes, it can be assumed that they play a more relevant role in online surveys because of the characteristics of online population (Schonlau et al., 2009).

Coverage bias is common to practically all online surveys except those that attempt to introduce offline population in their sampling frames. This is a consequence of the lack of exhaustiveness of the internet penetration, which, despite its substantial growth in the last few years, is still insufficient in some specific population strata. According to 2020 data from the Spanish National Institute of Statistics (Spanish National Institute of Statistics, 2020), 93,2 % of the Spanish population between 16 and 74 years of age access the internet at least once every three months, but this number is not homogeneous across population groups. For instance, while 99,5 % of the people with a PhD access the internet at least once every three months, this percentage decreases with the education level to the point that only 76 % and 51,4 % of the population with elemental studies and no studies respectively uses the internet at least once every three months. There is also a notorious age divide: while 99,8 % of the population between 16 and 24 years of age access the internet at least once every three months, only 69,7 % of the population between 65 and 74 years of age does so. Finally, a class divide can be seen as well: while this percentage is 99 % for people living in households with monthly net earnings above 2.500 EUR, it decreases to 84,6 % for people living in households with monthly net earnings below 900 EUR. Evidences of similar divides have been observed in the United States (Couper et al., 2018). The described percentages indicate that any internet survey with no offline recruitment will systematically leave some specific parts of the population out, which makes them an example of nonprobability sampling (because not all population members have a positive inclusion probability) and can lead to significant differences between the population represented in the samples obtained from internet surveys and the actual target population. This discrepancy can be seen as a consequence of selection bias (Elliott and Valliant, 2017).

In nonprobability surveys, such as self-selected samples using snowballing in social media websites and surveys using opt-in online panels, self-selection biases are common. From the definition of Bethlehem (2010), self-selection surveys can be described as the surveys where the selection relies completely on the individuals; they are the ones that choose to be selected in the sample or not. In nonprobability online surveys as the ones described above, self-selection takes place as the questionnaires are just left on the internet so that anyone who browses the web and sees it can be a survey respondent if that user decides to take it, meaning that there is no survey design associated to the recruitment. The absence of that design means that the inclusion probabilities for the members of the target population are unknown, and therefore any sample drawn using this method is a nonprobability sample. The self-selection mechanism can lead to important amounts of selection bias if there is a relationship between the variables of interest of the survey and the probability of participating in it. Numerous real world examples of self-selection bias in online surveys can be found; for instance, Faas (2004) compared the results of an offline election poll, an online election poll selected via online panel, and an online election poll whose respondents had been self-selected. The results showed very large differences between the composition of the latter poll and the composition of the other two polls. Bethlehem (2009a) mentions two examples of electoral polls in the Netherlands in 2003 and 2006, where it is noticeable that online surveys with self-selection mechanisms tended to strongly favour the estimated vote for some parties. Bethlehem (2015) also studied the opposition to Sunday shopping in three samples from the Netherlands: an offline sample from shopping centers' clients, an online panel survey and an online self-selected sample. Results showed that both the demographics and the observed opposition to Sunday shopping was largely different in the online self-selected sample in comparison to the other two samples and the population demographics. Smironva et al. (2020) studied self-selection biases in hotel online reviews, and found significant differences between online and offline mean rating scores. Self-selection biases can be relevant even when the target population is the online population. Faas and Schoen (2006) compared the results of 2002 German Election polls done with face to face interviews of internet users, online panels and using self-selected individuals, finding that the latter poll showed estimates of population parameters and associations significantly different from those obtained from the other sources of data. Khazaal et al. (2014) used a self-selected sample and a random sample to study the virtual characters of online players, finding that the parameter estimates obtained from the self-selected sample differed significantly from those obtained from the random sample.

The third source of bias that takes place in online surveys is non-response bias. It can happen when a fraction of the people able to participate in the survey does not do it, and their characteristics are significantly different from those that participate. This bias is related to self-selection bias, but there are some differences: while self-selection bias refers to those cases when the survey is open to anyone to participate, non-response bias occurs specifically in those cases when a number of individuals have been selected to participate (following a probability sampling scheme or not) in the survey. Non-response bias is very common in online panels, where surveys are sent to certain members of the panel who decide to participate or not. It can also be a problem even when recruiting members of the panel, as noted in Elliott and Valliant (2017). A meta-study by Manfreda et al. (2008) using 45 publications found that the response rate (a key metric for non-response bias measurement) in online surveys was, in average, 11 % lower than response rate in other survey modes. This trend has been observed in more recent studies comparing online surveys with mail surveys (Millar and Dillman, 2011; Loomis and Paterson, 2018). The review on the value of online surveys done by Evans and Mathur (2018) cites some factors that may promote nonresponse, such as excessive survey length, content and wording of the surveys themselves, incentives, invitation wording and reminders, and researcher and sponsor identity.

It must be noted that, despite online surveys are able to recruit more participants that other survey modes, larger sample sizes are not sufficient to remove biases coming from any sources or make the survey estimates efficient (Meng, 2018). Due to the growing interest of online surveys for empirical research, there has been a growing interest on developing methods to mitigate these sources of bias in recent years. Some of that sources have been widely studied besides their importance in online surveys. For example, non-response bias is common to many probability surveys regardless of their mode of administration, and a vast literature has been developed on techniques to reduce this bias. We could mention:

- Imputation for item non-response, where the individual has taken the survey but not answered all the questions. This technique is based on completing the missing responses with imputed values, obtained from models fitted using the rest of the information on each individual (Rubin, 1996).
- Reweighting for unit non-response, which refers to the case where the selected individual has not taken the survey (Särndal and Lundström, 2005)

The case of self-selection and coverage bias is more complicated. Regarding coverage bias, calibration weighting (Deville and Särndal, 1992) has been proposed as a method to approximate the sample to the target population. This methodology, similar to reweighting for non-response treatment, has been studied for the online survey context in several works. Dever et al. (2008) used data from two health surveys to construct calibration estimators to mitigate undercoverage of online surveys. Although the results showed efficiency at addressing nonresponse and coverage errors, the authors were doubtful on the application of the method in volunteer web surveys. Bethlehem (2010) used a fictitious population which aimed to emulate a situation where a sample is drawn from an online population. The study used post-stratification weights (which are a form of calibration) to remove bias; this approach was observed to be effective in Missing At Random (MAR) situations, where the selection mechanism is indirectly related (i.e. via a mediator variable) to the variable of interest, but not in Missing Not At Random (MNAR) situations, where the selection mechanism is directly related to the variable of interest. Valliant and Dever (2011) used calibrated estimates with the general regression estimator (GREG) for a fictitious population based on real data. The estimates were approximately unbiased when volunteering was not directly related to the analysis variables and the relationship between calibration covariates and the analysis variables could be properly described with a linear model. The use of calibration in nonprobability samples, given the lack of a sampling design, is still a challenge which has been discouraged by some authors (Devaud and Tillé, 2019).

On the other hand, methods to correct self-selection bias had not been widely studied before the rise of online surveys. Since then, some older methods developed for other sources of bias were adapted to address self-selection. Among the alternatives, we could mention the following ones:

- Propensity Score Adjustment (PSA). This method was originally developed by Rosenbaum and Rubin (1983) to address selection bias in experimental designs. Although it was adapted to mitigate non-response bias a few years later (Little, 1986), it was not fully developed to control selection bias in online surveys until the work by Lee (2006), which stipulated that a reference probability sample must be available. The method is based on predicting the propensity of an individual to participate in the online sample, and using the predicted propensity as the inclusion probability to be used in weighted estimators. The efficiency of PSA at successfully removing self-selection bias was proven in later works by Lee and Valliant (2009) and Valliant and Dever (2011). However, this efficiency can only be achieved if some conditions apply; more precisely, the covariates used for the adjustment must be related to both the selection mechanism and the variables of interest, and the propensity weights obtained must be adjusted via calibration weighting. In addition, the application of PSA increases the variance of the estimators.

- Propensity Score Matching. This option is closely related to PSA, but in this case the method is based on assigning each individual of the reference probability sample an individual from the online sample who approximately matches their values, producing a new sample that matches the population characteristics while being able to estimate population parameters relative to the target variable (which is only measured in the online sample). This approach has also been studied for the case of opt-in online panels, with promising results (Rivers, 2006; Vavreck and Rivers, 2008; Terhanian and Bremer, 2012).

- Kernel weighting (KW). This approach was developed in Wang et al. (2020) and it is closely related to PSA (it can even be considered a generalization of the methods intended to transform propensities into weights). The propensity estimates are used to calculate the distance between every member of the nonprobability sample and every member of the probability sample, and the resulting distances are smoothed via a kernel function. Finally, the weight for an individual of the nonprobability sample is obtained as the sum of each smoothed distance between that individual and every member of the probability sample, multiplied by the design weight of the correspondent member of the probability sample. Research shows that this method can largely benefit from advanced data mining and prediction techniques (Kern et al., 2020).

- Statistical Matching (SM). This method, which is also named as "mass imputationïn literature, was developed in Rivers (2007) as a technique to address selection bias in web surveys by means of predictive modelling. Beaumont and Bissonnette (2011), on their adaptation of SM to address non-response bias, developed some important properties of SM estimators, such as their variance.
- Doubly Robust Inference. This method was originally proposed in Kim and Haziza (2014) for the adjustment of nonresponse error in probability surveys, and it was adapted to nonprobability sampling in Kim and Wang (2019) and Chen et al. (2020b). It is based on the application of Statistical Matching, but includes a term that takes into account the prediction error, using the prediction residuals in the nonprobability sample and weighting each one according to their estimated participation propensity.
- Estimators based in superpopulation models. This approach was firstly developed by Royall (1970) for the probability sampling case where the full census is available for some variables which have also been measured in the sample and a superpopulation model is assumed. Valliant et al. (2000) developed some properties of these estimators for the same context. The use of model-based estimators in nonprobability surveys has been recently studied by Buelens et al. (2018), using data on annual mileages driven by vehicles in the Netherlands to build a pseudopopulation from which several samples (with different levels of bias) are drawn. Model-based estimators were calculated using a wide range of predictive models, including Machine Learning algorithms. Results showed that selection bias can be removed if the right covariates are used for prediction.

The aim of this thesis is to provide methodological advances in the estimation from online surveys, both by combining or improving the methods that already exist, and by adapting methods from probability sampling to the nonprobability sampling context. In Appendix 1, we study the feasibility of the combination of PSA and calibration, using actual or estimated population totals, for mitigating bias in nonprobability online surveys. In Appendix 2, we consider Machine Learning classification algorithms for propensity estimation in PSA, and we study if they can be an alternative to logistic regression. In Appendix 3, we develop a theoretical framework for the estimation of general population parameters using nonprobability samples, including the two-phase procedure for estimation and the estimator of the variance of the parameters estimated in that procedure. In Appendix 4, we study the impact of the application of variable selection techniques to obtain subset of optimal covariates, in terms of bias reduction, to be used as input variables in PSA. In Appendix 5, we adapt the methods from superpopulation modeling to the nonprobability online sampling context, applying modern prediction techniques to compare their performance. In Appendix 6, we adapt the methodology of weight smoothing for multipurpose nonprobability surveys, using two advanced prediction algorithms to obtain the smoothed weights and comparing the results. In Appendix 7, we apply PSA in a real world problem where we aim to estimate the prevalence of several health-related issues in a population of healthcare professionals, studying several algorithms for propensity estimation and the results that they provide.

# Chapter 2

# Objectives

## 2.1. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

The combination of Propensity Score Adjustment (PSA) and calibration weighting to address selection bias has been explored in literature (Lee and Valliant, 2009; Valliant and Dever, 2011). On the other hand, calibration weighting using estimated population totals instead of actual population totals in the calibration equations was also studied by Bethlehem (2010). However, research on the different approaches to treat propensities obtained with PSA in order to use them in weighted estimators is scarce. A comparison between stratified and non-stratified propensities was done in Valliant and Dever (2011), but some other transformations can be used, such as the stratification proposed in Lee and Valliant (2009), which uses design weights of the reference probability sample, and inverse propensity weighting proposed in Schonlau and Couper (2017), which takes into account the fact that individuals from the online sample have to be removed from the target population of the reference sample (assuming that there is no overlap between samples).

We study the combination of PSA and further calibration reweighting, using the weights obtained from propensity scores as initial design weights in the calibration process. To do so, we use a simulated population with a multiclass target variable that reflects each of the three possible scenarios for the selection mechanism: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). We transform propensity scores into weights using two approaches: the inverse propensity weighting proposed in Schonlau and Couper (2017), and the stratified propensity weighting using design weights from the reference sample proposed in Lee and Valliant (2009). We also compare the results from the combination of PSA and calibration using two approaches for the latter: actual population totals and estimated population totals (from the reference sample) in the calibration equations. The efficiency is studied through the measures of bias and standard deviation of the estimates.

## 2.2. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

PSA is based on predicting propensities of being included in the nonprobability sample using a combination of both the nonprobability (convenience) sample and the probability (reference) sample. As developed in theory, PSA uses logistic models for this prediction, which enables the obtention of probabilities based in the logistic formula with some covariates. Although logistic regression is a robust model which can offer good results in many situations, its application in the online survey context poses some challenges. The use of larger samples with a larger number of covariates, with different levels of association with the selection mechanism and the target variable, can compromise the behavior of logistic regression. On the other hand, this approach requires to establish in advance the interactions between variables, which can be unfeasible in many situations.

The use of more advanced methods for predictive modelling, such as Machine Learning (ML) classification algorithms, can be an alternative to logistic regression in this context. This approach has been already studied for propensity scoring to address non-response issues, showing good results (Phipps and Toth, 2012; Buskirk and Kolenikov, 2015).

We propose the use of ML algorithms for PSA with the objective of mitigating selection bias in online surveys. We aim to study the application of a subset of the most popular classification algorithms in the propensity modelling step, after optimizing their hyperparameters with regard to minimizing the log-Loss metric (which is related to the accuracy in the prediction of the probability).

## 2.3. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

Research on the estimation of population parameters using adjusted nonprobability samples has been focused on linear parameters, such as population means, totals and proportions. On the other hand, the development of a formula to estimate the variance of PSA estimators is still under study. This work aims to fill the gap in literature, developing a theoretical framework for the estimation of general population parameters. The framework is based on the work by Chen et al. (2020a) at developing the properties of the doubly robust estimator, extending them to the estimation of general parameters, along with an expression for the estimation of the estimator's variance. The estimator and its properties is tested in a simulation study using three different real world datasets.

## 2.4. Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

The efficacy of adjustments for nonprobability online samples depend to a large extent on the available auxiliary information. That information must be relevant regarding target variables and the selection mechanism of the sample. For this reason, it is critical to properly select the information to be used in the adjustments, in order to avoid including non-relevant or redundant information that could lead to overfitting situations or unstability of the estimators. In this context, recent data mining techniques which were developed to extract relevant features in large datasets could be use to automatize the process of selecting the proper covariates to be included in the adjustments, without the need to have any prior information about the relationships between variables.

The objective of the study is to evaluate the performance of variable selection methods prior to PSA in different contexts, where Machine Learning classification algorithms are used and PSA weights are used on their own or combined with Raking calibration (in which the initial weights are the PSA weights).

## 2.5. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

Estimators based in superpopulation modeling, such as model-based, model-assisted and model-calibrated estimators, have been explored to reduce bias in nonprobability surveys, with promising results (Valliant et al., 2000). These estimators are often adjusted using linear regressions, which may entail several disadvantages for large populations in comparison to modern prediction methods such as ML algorithms. The use of those algorithms for model-assisted estimators was proposed in Breidt et al. (2017), and they were applied in the study by Buelens et al. (2018) on model-based estimation in nonprobability surveys.

This study aims to obtain a full picture of the efficiency of model-based, model-assisted and model-calibrated estimators for nonprobability online surveys, and the ability of ML algorithms to be an alternative to linear regression models. We compare the three approaches using three simulated populations, obtained from real world datasets, and a set of prediction models including advanced linear regression models, tree-based models, boosting models, k-nearest neighbours classification and neural networks with and without regularization.

## 2.6. Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

The importance of online surveys in official statistics is growing steadily in the last years, as they can offer substantial advantages in some crucial areas which were a cause of concern in traditional survey modes. Apart from their use in probability surveys as an alternative to personal interviews, which can be troublesome in the case of hard-to-reach interviewees, nonprobability online surveys are starting to be considered as a considerably fast way of obtaining timely estimates of multiple features from a given population. The drawbacks that they pose might be corrected with the usual techniques for addressing selection bias.

However, official surveys are often multipurpose (i. e. have multiple variables of interest) and, in some cases, the exact variables of interest are unknown to the researcher at the moment of adjusting the sample. This is an important issue because it makes unfeasible to apply adjustments based on predicting the target variable: in the best-case scenario, we would need one model for each variable of interest, which can be impractical and also increase the probabilities of model misspecification for any of the variables. For these reasons, adjustments based on weighted estimates, which only require a single model to obtain a single vector of weights, are seen as a more adequate choice.

Weighting adjustments use auxiliary covariates that should ideally be related to the selection mechanism and the variable of interest. In a multipurpose survey, covariates can be strongly related to the propensity to participate but weakly related to the variables of interest, or can also be related to some target variables but unrelated to others. Such covariates do not contribute to bias reduction and might increase the variance of the estimates. In this context, weight smoothing techniques developed for probability samples can be used. Weight smoothing aims to model the relationship between the vector of weights and the variables of interest, with the objective of substituting the weights by the fitted values of the developed models.

The objectives of this work are twofold. The first objective is to evaluate the feasibility of weight smoothing techniques in the nonprobability online survey context, where the weights are not given by design but by adjusment methods that tend to increase the variability of the estimates and therefore increase the contribution of the noise in the vector of weights. The second objective is to study the performance of modern prediction techniques for the modelization of the weight smoothing model; having algorithms that perform variable selection prior to the prediction may be helpful in terms of discarding those variables of interest that are not related to the weights, leading to better model specifications.

## 2.7. Self-perceived health, life satisfaction and related factors among healthcare professionals and the general population: analysis of an online survey, with propensity score adjustment

Health status of healthcare professionals (HCP) is an issue of growing importance, due to the high levels of stress, anxiety and burnout that they have to convey, especially after the irruption of the COVID-19 pandemic. However, health status and life satisfaction among HCP has not been widely studied because of limited data and resources to carry out quality studies on the HCP population. In this context, nonprobability surveys can help to fill this gap, as long as methods to mitigate biases inherent to such surveys are applied.

The objective of this study is to estimate the prevalence of health issues and the factors related to life satisfaction and self-perceived health in HCP from the region of Andalusia (Spain), and to study how PSA can address selection bias in a real world situation. PSA is applied using several approaches to predict propensities in order to compare their adequacy to such situations: logistic regression, and a set of ML algorithms with varying hyperparameter configurations. We also aim to study empirically the statistical behavior of the weights, and the relationships between their stability and the properties of the estimators that they produce.

# Chapter 3

# Methodology

## 3.1. Framework of probability and nonprobability samples

Let $U$ be a population of interest of size $N = 1, 2, 3....$ We are interested in the estimation of a population parameter for a given variable of interest $y$, $\theta_y$. For such a task, we draw a sample $s$ of size $n$ from $U$, with a sample design $(s_d, p_d)$ where $p_d$ represents the probability of drawing a given sample $s_d$ which belongs to the space of possible samples, $\delta = \{s|s \in U\}$. As a result, the first order inclusion probability of the individual $i$, $\pi_i$, can be defined as

$$\pi_i = \sum_{s \ni i} p_d(s), i = 1, ..., n. \tag{3.1}$$

and the second order inclusion probability of the individuals $i$ and $j$, $\pi_{ij}$, can be defined as

$$\pi_{ij} = \sum_{s \ni \{i,j\}} p_d(s), i, j = 1, ..., n. \tag{3.2}$$

In a probability sampling design, $p_d > 0, \forall s_d \in \delta$. From the results in Horvitz and Thompson (1952), it can be shown that linear parameters, such as the population total, $T_y$, can be unbiasedly estimated through the formula

$$\hat{T}_y = \frac{\sum_{i \in s_d} y_i}{\pi_i} = \sum_{i \in s_d} d_i y_i \tag{3.3}$$

where $d_i = 1/\pi_i$ is the design weight of individual $i$. The population mean, $\overline{Y}$ (which is equivalent to the population proportion of a condition if $y$ is an indicator variable where 1 indicates the presence of the condition), can also be estimated through the formula

$$\hat{\overline{Y}}^{HT} = \frac{1}{N}\frac{\sum_{i\in s_d} y_i}{\pi_i}. \tag{3.4}$$

An expression for the variance of both estimators, as well as for the estimator of the variance, is also given in the paper:

$$Var(\hat{T}_y) = \sum_{k\in U}\sum_{l\in U}\frac{y_k y_l}{\pi_k \pi_l}\pi_{kl} - \pi_k\pi_l \tag{3.5}$$

$$V\hat{a}r(\hat{T}_y) = \sum_{k\in s_d}\sum_{l\in s_d}\frac{y_k y_l}{\pi_k \pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \tag{3.6}$$

$$Var(\hat{\overline{Y}}^{HT}) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\frac{y_k y_l}{\pi_k \pi_l}\left(\pi_{kl} - \pi_k\pi_l\right) \tag{3.7}$$

$$V\hat{a}r(\hat{\overline{Y}}^{HT}) = \frac{1}{N^2}\sum_{k\in s_d}\sum_{l\in s_d}\frac{y_k y_l}{\pi_k \pi_l}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \tag{3.8}$$

The Horvitz-Thompson estimator of the mean is not linearly invariant, as shown in Tillé (2020). This is an undesirable property, as it introduces an extra error term independent of $y$. The Hajek estimator, developed in Hájek (1981), is an alternative for the estimation of a population mean based on considering the sum of inverse inclusion probabilities as the population total:

$$\hat{\overline{Y}}^{H} = \frac{\sum_{i=1}^{n} y_i/\pi_i}{\sum_{i=1}^{n} 1/\pi_i}. \tag{3.9}$$

The Hajek estimator is linearly invariant, but it can be biased in some applications since it depends on the quotient of two random variables.

In the case of nonprobability samples, the design $(s_d, p_d)$ does not exist as the probabilities of drawing each sample, $p_d$, are not known or the condition $p_d > 0, \forall s_d \in \delta$ is not met. If a nonprobability sample (for example, a sample of volunteers) is available, $s_v$ with a sample size $n_v$, the usual estimator of the mean is given by the expression

$$\hat{\overline{Y}}^{NP} = \frac{\sum_{i\in U} y_i R_i}{\sum_{i\in U} R_i} = \frac{\sum_{i\in s_v} y_i}{n_v}, \tag{3.10}$$

where $R$ is an indicator variable which measures whether an individual belongs to $s_v$ or not:

$$R_i = \left\{\begin{array}{ll} 1 & i \in s_v \\ 0 & i \notin s_v \end{array}\right., i \in U \tag{3.11}$$

It can be easily shown that $\sum_{i\in U} R_i = n_v$. The usual estimator of the total is given by a very similar formula, but multiplying by the size of the population such that

$$\hat{T}_y^{NP} = N\frac{\sum_{i\in U} y_i R_i}{\sum_{i\in U} R_i} = N\frac{\sum_{i\in s_v} y_i}{n_v} = \sum_{i\in s_v} d_{vi} y_i, \tag{3.12}$$

where the values of the vector $d_v = n_v/N$ are equivalent to the design weights previously defined, and can be seen as the weights that are assigned to each unit of the nonprobability sample when they are unknown.

These estimators are subject to the selection mechanism that influences the indicator variable $R$. Given $\{R_i, i \in U\}$, the error of $\hat{\overline{Y}}^{NP}$ can be written as in Kim and Wang (2019):

$$\hat{\overline{Y}}^{NP} - \overline{Y} = \frac{N}{n_v}Cov(R,Y), \tag{3.13}$$

where

$$Cov(R,Y) = \frac{1}{N}\sum_{i\in U}(R_i - \overline{R})(y_i - \overline{Y}) \tag{3.14}$$

Given that $\overline{R} = \frac{\sum_{i\in U} R_i}{N} = \frac{n_v}{N}$, the previous formula for $Cov(R,Y)$ can be rewritten as

$$Cov(R,Y) = \frac{1}{N}\sum_{i\in U}(R_i - \frac{n_v}{N})(y_i - \overline{Y}) \tag{3.15}$$

The mean quadratic error, which is a measure of the error in the estimation of $\overline{Y}$, can be expressed as

$$E_R[(\hat{\overline{Y}}^{NP} - \overline{Y})^2] = \left(\frac{N}{n_v}\right)^2 E_R[Cov(R,Y)^2]. \tag{3.16}$$

where $E_R[.]$ denotes the expectation with respect to the selection mechanism for $R$. An alternative expression of $E_R[(\hat{\overline{Y}}^{NP} - \overline{Y})^2]$ for general cases was developed in Meng (2018):

$$E_R[(\hat{\overline{Y}}^{NP} - \overline{Y})^2] = E_R[Cov(R,Y)^2] \cdot \left(\frac{N}{n_v} - 1\right) \cdot \sigma^2, \tag{3.17}$$

where $\sigma^2 = \frac{\sum_{i\in U}(y_i - \overline{Y})^2}{N}$. This form allows to divide the error in the estimation in three parts: the inherent variability associated to the variable of interest ($\sigma^2$), the amount of available data, represented as the inverse of the sampling fraction $\left(\frac{N}{n_v} - 1\right)$, and the relationship between the selection mechanism and the variable of interest. If that relationship is strong, the error in the estimation tends to grow more rapidly than in the case where the sample size is reduced, as shown in Meng (2018). Therefore, a reasonable and efficient strategy is to focus on reducing the effect of the relationship between the selection mechanism and the target variable, rather than focusing on increasing the sample size, which may not be as effective unless the sampling fraction is exceptionally large. The available techniques to mitigate the bias induced by the selection mechanism will be described in the following subsections.

## 3.2. Estimation from nonprobability samples with auxiliary population totals available

Let $\mathbf{T_x} = (T_{x_1}, ..., T_{x_p})$ be a vector of population totals of $p$ auxiliary variables $\mathbf{x} = (x_1, ..., x_p)$, which have been measured for all the individuals in the sample $s$. If $s$ has been drawn following a non-probabilistic scheme, it is possible that, as the bias appears, some individuals that take certain values for $\mathbf{x}$ are more prone to be in the sample than others, mainly because of undercoverage of certain subsets of $U$.

Calibration weighting, developed by Deville and Särndal (1992) for probability samples, uses the population totals to create new weights that solve the aforementioned unbalance between the population totals and the observed sample. Let $d = \frac{1}{\pi}$ be the vector of design weights for $s$; the calibration procedure aims to minimize the distance between the new weights, $w$, and the design weights, while at the same time respecting the calibration equations, according to which the sample estimates of the auxiliary variables using weights $w$ should be equal to $\mathbf{T}_X$:

$$\begin{aligned} min \quad & \textstyle\sum_{k\in s} G(w_k, d_k) \\ subject\ to \quad & \textstyle\sum_{k\in s} w_k X = \mathbf{T_x} \end{aligned} \tag{3.18}$$

where $G(.,.)$ can be any differentiable distance function. Deville and Särndal (1992) propose a class of distance measures, with the linear distance being the simplest case:

$$G(w_k, d_k) = \frac{(w_k - d_k)^2}{2d_k} \tag{3.19}$$

When using linear distance, the solution to the optimization problem in Eq. (3.18) via Lagrange multipliers provides an estimator for the population total, $\hat{T}_y^{cal}$, which is equivalent to the generalized regression estimator developed in Cassel et al. (1976) (further description of the estimator will be done in Section 3.9):

$$\hat{T}_y^{cal} = \hat{T}_y + (\mathbf{T_x} - \sum_{k\in s} d_k \mathbf{x}_k)'\hat{\mathbf{B}}_s = \hat{T}_y + (\mathbf{T_x} - \sum_{k\in s} d_k \mathbf{x}_k)'(\sum_{k\in s} d_k q_k \mathbf{x_k} \mathbf{x_k}')^{-1}(\sum_{k\in s}\sum_{k\in s} d_k q_k \mathbf{x_k} \mathbf{y_k}) \tag{3.20}$$

where $q_k$ is a known positive weight unrelated to $d$ for the individual $k \in s$. The usual choice in applications is $q_k = 1$, hence it has no relevance in the results, although some cases might motivate other choices (see Example 1 in Deville and Särndal (1992)).

The case where $s$ is a nonprobability sample $s_v$ is more complicated. In fact, the lack of sampling design for nonprobability samples leads to the inability of using $d$ for calibration reweighting. In such a situation, the weights $w$ will be equal to those obtained after post-stratification (Smith, 1991). This post-stratification strategy was studied in Bethlehem (2010) in order to reduce selection bias in nonprobability online surveys, with successful results.

## 3.3. Estimation from nonprobability samples with a reference probability sample using Propensity Score Adjustment

### 3.3.1. Propensity estimation

Let $s_r$ be a probability (reference) sample of size $n_r$, drawn from $U$, with a sample design $(s_{rd}, p_{rd})$ where $p_{rd}$ represents the probability of drawing a given sample $s_{rd}$ which belongs to the space of possible samples, $\delta = \{s_r | s_r \in U\}$, and $p_{rd} > 0, \forall s_{rd} \in \delta$. As a result, the first order inclusion probability of the individual $i$, $\pi_{ri}$, can be defined as

$$\pi_{ri} = \sum_{s_r \ni i} p_{rd}(s_r), i = 1, ..., n_r. \tag{3.21}$$

Let $s_v$ be a nonprobability (convenience) sample of size $n_v$, drawn from $U$ or a subset of $U$ which represents the potentially covered population, $U_{pc} \subseteq U$. There is no sample design and therefore there are no design weights nor known inclusion probabilities. Let $\mathbf{x} = (x_1, ..., x_p)$ be a vector of covariates that has been measured in both $s_r$ and $s_v$, and $y$ a variable of interest that has been measured only in $s_v$. For the estimation of parameters of $y$, several techniques can be applied taking into consideration the reference sample $s_r$.

Propensity Score Adjustment (PSA) assumes that a relationship underlies between the inclusion probability for $s_v$, $\pi_{vi}$, and some auxiliar covariates $\mathbf{x}$.

$$\pi_{vi} = Pr(R_i = 1 | \mathbf{x}_i), i \in U \tag{3.22}$$

This assumption implies that the selection mechanism for $s_v$ is ignorable, which means that the effect of this relationship in the final estimates can be reduced or eliminated with proper adjustments. In this case, we could say that the non-sampled individuals follow a Missing At Random (MAR) mechanism, given that the selection is not directly related to the variable of interest, although it can be indirectly related if some kind of relationship lies between $\mathbf{x}$ and $y$. If that relationship does not exist, the mechanism is Missing Completely At Random (MCAR) and, in that case, the use of $s_v$ will provide unbiased estimates for parameters of $y$.

In a nonprobability survey, the inclusion probabilities $\pi_{vi}$ are not known, but they can be estimated using the data available from $s_v$ and $s_r$. We define the indicator variable $R^*$, defined for any individual in $s_v \cup s_r$, which measures whether an individual belongs to $s_v$ or $s_r$:

$$R_i^* = \begin{cases} 1 & i \in s_v \\ 0 & i \in s_r \end{cases}, i \in s_v \cup s_r \tag{3.23}$$

This variable is a proxy of the actual indicator variable $R$, which will be less accurate as the overlap between $s_v$ and $s_r$ (number of individuals belonging to both samples at the same time) increases. If the sampling fractions $\frac{n_v}{N}$ and $\frac{n_r}{N}$ are small enough, the overlap probability will remain very low, making $R^*$ a good approximation of $R$ for $s_v \cup s_r$. We can therefore define the approximate inclusion probabilities, $\pi_{vi}^*$, as follows:

$$\pi_{vi}^* = Pr(R_i^* = 1|\mathbf{x}_i), i \in s_v \cup s_r \tag{3.24}$$

PSA is based on estimating the expected value of $\pi_{vi}^*$ for any $i \in s_v \cup s_r$ through a model $M$ if covariates $\mathbf{x}$ are available for both samples:

$$\hat{\pi}_{vi}^* = E_M\left[R_i^* = 1|\mathbf{x}_i\right], i \in s_v \cup s_r \tag{3.25}$$

The usual choice is to consider a logistic regression model to estimate the probabilities, hence their expression can be described as

$$\hat{\pi}_{vi}^* = \frac{1}{1 + exp(-\beta\mathbf{x}_i)}, i \in s_v \cup s_r \tag{3.26}$$

where $\beta$ is the vector of coefficients estimated in the logistic regression modeling according to some optimization criteria.

### 3.3.2. Propensity weighting

Once the propensities to participate have been estimated for each individual in $s_v$, they can be transformed into weights to be used in the estimators of the population total and the population mean. The simplest approach is the inverse probability weighting (Valliant, 2020):

$$w_i^{PSA1} = \frac{1}{\hat{\pi}_{vi}^*}, i \in s_v \tag{3.27}$$

A similar alternative is the one proposed in Schonlau and Couper (2017), which takes into account that members of $s_v$ do not belong to the target population in the sampling design of $s_r$ as their inclusion would enable both samples to overlap, which is an undesirable behavior in this context.

$$w_i^{PSA2} = \frac{1 - \hat{\pi}^*_{vi}}{\hat{\pi}^*_{vi}}, i \in s_v \tag{3.28}$$

Another approach that might be useful, specially if the vector of propensities present extreme values (close to 0 or 1) that can affect the variability of the estimates, is the stratification of propensities. Lee and Valliant (2009) proposed a system based on obtaining the vector of propensities $\hat{\pi}^*_i, i \in s_v \cup s_r$ for the combination of both samples, sorting the propensities and partitioning them on $C$ classes. Ideally, $C = 5$ following Cochran (1968) results which show that five subclasses are sufficient to remove 90 % of the bias produced by subclassification in univariate analysis. Rosenbaum and Rubin (1984) showed that the percentage of removal of the bias from subclassification also applies to propensity scores. Once the partition is made, we define a correction factor for a propensity stratum $c$, $f_c$, based on the design weights of each sample:

$$f_c = \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_j^v / \sum_{j \in s_v} d_j^v} \tag{3.29}$$

where $d^r$ and $d^v$ represent the design weights of the reference and the convenience sample respectively, and $s_r^c$ and $s_v^c$ are the subset of individuals from the reference and the convenience sample respectively that belong to the $c$-th stratum. Note that, in the case where design weights are not available, an alternative for $f_c$ can be defined:

$$f_c = \frac{n_r^c / n_r}{n_v^c / n_v} \tag{3.30}$$

where $n_r^c$ and $n_v^c$ are the number of individuals from the reference and the convenience sample respectively that belong to the $c$-th stratum. The final weights of $s_v$ are defined as the product of the correction factor and the original design weights:

$$w_i^{PSA3} = f_c d_i^v, i \in s_v, c \ni i \tag{3.31}$$

If the design weights are the inverse of the sampling fraction, $d^r = \frac{N}{n_r}$ and $d^v = \frac{N}{n_v}$, which is a common choice when design weights are unknown, $f_c$ is equivalent to the approach described in Eq. (3.30). In that case, when the vector $w^{PSA3}$ is applied in the Hajek estimator of the mean, the final estimator is equivalent to the Horvitz-Thompson estimator of the mean. Given that

$$w_i^{PSA3} = f_c d_i^v = \frac{n_r^c n_v}{n_v^c n_r} \cdot \frac{N}{n_v} = N \frac{n_r^c}{n_v^c n_r}, i \in s_v, c \ni i, \tag{3.32}$$

it can be shown that the sum of the weights vector $w^{PSA3}$ equals the population size if it is known and $d_i^v = \frac{N}{n_v}, \forall i \in s_v$:

$$\sum_{i=1}^{n_v} w_i^{PSA3} = \sum_{i=1}^{n_v} N \frac{n_r^c}{n_v^c n_r} = \frac{N}{n_r} \sum_{i=1}^{n_v} \frac{n_r^c}{n_v^c} = \frac{N}{n_r} \sum_{c=1}^{C} \frac{n_v^c n_r^c}{n_v^c} = \frac{N}{n_r} \sum_{c=1}^{C} n_r^c = \frac{N}{n_r} n_r = N \tag{3.33}$$

When substituting in $\hat{\overline{Y}}^H$ we get

$$\hat{\overline{Y}}^{H\_PSA3} = \frac{\sum_{i=1}^{n_v} w_i^{PSA3} y_i}{\sum_{i=1}^{n_v} w_i^{PSA3}} = \frac{\sum_{i=1}^{n_v} w_i^{PSA3} y_i}{N} = \hat{\overline{Y}}^{HT\_PSA3} \tag{3.34}$$

An alternative approach that is also based in propensity stratification is the one proposed in Valliant and Dever (2011). The segmentation step follows the same principle (sorting propensities and dividing individuals in $C$ subgroups according to their values), but in this case, the weights are defined as the inverse of the mean propensity in the stratum to which each individual belongs to:

$$w_i^{PSA4} = \frac{n_v^c}{\sum_{j \in s_v^c} \hat{\pi}_{vj}^*}, i \in s_v, c \ni i \tag{3.35}$$

### 3.3.3. Machine Learning algorithms in Propensity Score Adjustment

The propensity estimation step in PSA accepts the use of Machine Learning (ML) classification algorithms as an alternative to logistic regression, given that the goal in this context is to predict probabilities for the presence of the attribute (in this case, being in the nonprobability sample) measured by a binary variable ($R^*$). Most of the classification algorithms are able to provide such probabilities, based on non-parametric procedures. For instance, tree-based propensity estimation can be summarized in the formula

$$\hat{\pi}_{vi}^* = \begin{cases} \frac{n(s_v^{J_1})}{n((s_v \cup s_r)^{J_1})} & \{i \in s_v / \mathbf{x}_i \in J_1\} \\ \dots & \dots \\ \frac{n(s_v^{J_k})}{n((s_v \cup s_r)^{J_k})} & \{i \in s_v / \mathbf{x}_i \in J_k\} \end{cases} \tag{3.36}$$

where $J_1, ..., J_k$ represent the $k$ terminal nodes of a decision tree fitted using data from $s_v \cup s_r$ with $R^*$ being the target variable, each one representing a multivariate range of values according to which individuals are classified, and $n(s_v^{J_i})$ and $n((s_v \cup s_r)^{J_i})$ are the number of individuals from the convenience sample and the combined sample respectively that meet the criteria to be classified into the terminal node $J_i, i = 1, ..., k$. In bagging algorithms such as Random Forests (Breiman, 2001), a set of $m$ decision trees (known as *weak classifiers*) can be trained and then averaged to compute the propensities as

$$\hat{\pi}^*_{vi} = \frac{\sum_{j=1}^{m} \phi_j(\mathbf{x}_i)}{m}, \phi_j(\mathbf{x}_i) = \begin{cases} 1 & \{i \in s_v / \mathbf{x}_i \in J^j_{pr}\} \\ 0 & \{i \in s_v / \mathbf{x}_i \in J^j_{ab}\} \end{cases}, \tag{3.37}$$

where $J^j_{pr}$ and $J^j_{ab}$ represent the set of terminal nodes of the $j$th decision tree, $j = 1, ..., m$, where the individuals from the nonprobability sample are majority and minority respectively:

$$J^j_{pr} = \{J^j_l, l = 1, ..., k : \frac{n(s_v^{J^j_l})}{n((s_v \cup s_r)^{J^j_l})} \geq 0{,}5\} \tag{3.38}$$

$$J^j_{ab} = \{J^j_l, l = 1, ..., k : \frac{n(s_v^{J^j_l})}{n((s_v \cup s_r)^{J^j_l})} < 0{,}5\} \tag{3.39}$$

The $m$ trees can be also fitted through an iterative process aiming to improve the accuracy of the classification, also known as boosting. In this sense, Gradient Boosting Machine (GBM) algorithm (Friedman, 2001) is able to provide propensities according to the formula

$$\hat{\pi}^*_{vi} = \frac{1}{1 + exp(-w^T J(\mathbf{x}_i))}, i \in s_v \tag{3.40}$$

where $J(\mathbf{x}_i)$ is a matrix of terminal nodes of $m$ decision trees, fitted through the aforementioned iterative process, whose multivariate range adjusts to $\mathbf{x}_i$, and $w$ is the weight assigned to each decision tree.

Propensities can be estimated using other ML approaches. k-Nearest Neighbours (kNN) algorithm (Cover and Hart, 1967) estimates the probability of $R^*_i = 1, i \in s_v$ as the proportion of individuals belonging to $s_v$ among the $k$ neighbours of individual $i$:

$$\hat{\pi}^*_{vi} = \frac{\sum_{j \in s_r \cup s_v / d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k)})} R^*_j}{k}, i \in s_v \tag{3.41}$$

where $d$ is the distance function that measures the similarity between two individuals given their covariates, and $\mathbf{x}_{(k)}$ represents the covariates of the $k$-th closest individual.

The Bayes theorem can also be used to estimate the propensities through the Naïve Bayes classifier as follows:

$$\hat{\pi}^*_{vi} = P(R^*_i = 1|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|R^*_i = 1)P(R^*_i = 1)}{P(\mathbf{x}_i)}, i \in s_v \tag{3.42}$$

In spite of its simplicity (it assumes that the covariates are mutually independent), this approach can show a good behavior in classification tasks; however, it can lead to unstable estimates if the cardinality of the covariates is high. This formula is equivalent to that of the Linear Discriminant Analysis (LDA) approach, but in that case several parametric assumptions are made on the conditional probability $P(\mathbf{x}_i|R_i^* = 1)$ which ultimately allows the estimates to have an analytic expression. In Naïve Bayes, the probabilities rely entirely on the available data from $s_v \cup s_r$ without making any assumptions on their distribution.

## 3.4. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

We propose an approach based on applying PSA to obtain weights for a nonprobability sample based in estimated propensities, and using them as initial weights in calibration, using the weights provided by the latter procedure to estimate population parameters. In order to evaluate this approach, we perform a simulation study based on the work of Bethlehem (2010) using a fictitious population of size $N = 50,000$ with four covariates (age, education, nationality and gender), an indicator variable which measures whether an individual has internet access or not, and a variable of interest measuring the voting intention in a fictitous election with three parties (Party 1, 2 and 3). Across $1,000$ simulation runs, a reference sample of size $n_r = 500$ was drawn with simple random sampling without replacement (SRSWOR) from the full population, while seven convenience samples of sizes $n_v = 500, 750, 1000, 2500, 5000, 7500, 10000$ were drawn with SRSWOR only from the subset of the population with internet access.

Vote to Party 1 is related to gender, which is only slightly related to age, making the estimation of vote to Party 1 a Missing Completely At Random (MCAR) situation as the target variable is not related to the selection mechanism. Vote to Party 2 is related to age, which is directly related to internet access, making the estimation of vote to Party 2 a Missing At Random (MAR) situation as the target variable is indirectly related to the selection mechanism. Finally, vote to Party 3 is directly related to internet access, which constitutes a Missing Not At Random (MNAR) case for the estimation of vote to Party 3 as the target variable is directly related to the selection mechanism.

The covariates were divided in four scenarios. In each one of them, a couple of variables were used for PSA, while one variable was used for calibration. In Scenarios 1 and 2, age and education were used in PSA, while in Scenarios 3 and 4 age and nationality were used. In Scenarios 1 and 4, gender was used as the calibration variable, while nationality and education were used for that task in Scenario 2 and 3 respectively. The estimation of the voting intention for each of the three parties was done using the following approaches:

1. Non-adjusted estimates from the convenience sample.

2. Calibration using actual population totals for the covariates selected in the scenario.

3. Calibration using estimates for the population totals for the covariates selected in the scenario, obtained by applying the estimator of the total in the reference sample.

4. PSA, transforming estimated propensities into weights using the formula from Schonlau and Couper (2017) (weights for the Hajek estimator), with no further adjustments.

5. PSA, transforming estimated propensities into weights using the formula from Lee and Valliant (2009) and assuming $d^r = \frac{N}{n_r}$ and $d^v = \frac{N}{n_v}$ (weights for the Horvitz-Thompson estimator), with no further adjustments.

6. PSA, transforming estimated propensities into weights using the formula from Schonlau and Couper (2017) (weights for the Hajek estimator), and using those weights as initial weights for calibration with actual population totals of the covariates.

7. PSA, transforming estimated propensities into weights using the formula from Schonlau and Couper (2017) (weights for the Hajek estimator), and using those weights as initial weights for calibration with estimates for the population totals of the covariates.

8. PSA, transforming estimated propensities into weights using the formula from Lee and Valliant (2009) and assuming $d^r = \frac{N}{n_r}$ and $d^v = \frac{N}{n_v}$ (weights for the Horvitz-Thompson estimator), and using those weights as initial weights for calibration with actual population totals of the covariates.

9. PSA, transforming estimated propensities into weights using the formula from Lee and Valliant (2009) and assuming $d^r = \frac{N}{n_r}$ and $d^v = \frac{N}{n_v}$ (weights for the Horvitz-Thompson estimator), and using those weights as initial weights for calibration with estimates for the population totals of the covariates.

## 3.5. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

We study the feasibility of considering ML algorithms an alternative to logistic regression regarding propensity estimation in PSA. To do so, two simulation studies were considered: one using fictitious data and another one using a real dataset as pseudopopulation.

The first simulation used the same data described in Section 3.4 with two exceptions: the removal of the relationship between gender and age, so estimation for Party 1 would be fully MCAR, and the nonprobability sample $s_v$ would be drawn in each simulation run using three different mechanisms. These are described in the following list:

1. SRSWOR from the subset of the population with internet access.
2. Sampling with unequal selection probabilities from the subset of the population with internet access, where the probabilities were described by the formula:

$$\pi_{iv} = \frac{1}{1 + exp(-1 + 0{,}05Age_i)}, i \in U_I \tag{3.43}$$

   where $U_I$ represents the individuals with internet access in the full population $U$.

3. Sampling with unequal selection probabilities from the subset of the population with internet access, where the probabilities were described by the formula:

$$\pi_{iv} = \frac{1}{1 + exp(1 - sin(Age_i/20))}, i \in U_I \tag{3.44}$$

   where $U_I$ represents the individuals with internet access in the full population $U$.

The variable of interest was the voting intention, which estimation was done using PSA (with no further adjustments) and the approaches described in Section 3.3.3 for propensity estimation, with three algorithms for decision tree fitting: C4.5, C5.0 (Quinlan, 1993) and CART (Breiman et al., 1984). In each case, a grid of hyperparameters was defined so PSA would be performed once for each combination of hyperparameters. The grid was the following one:

- Decision trees: 0.1, 0.25 and 0.5 as confidence values for pruning, 0.5 %, 1 % and 5 % of the training dataset size ($n_r + n_v$) as the minimum number of observations per node.

- kNN: $k = 3, 5, 7, 9, 11, 13$.

- Naïve Bayes: Laplace smoothing (number that is added to the numerator result in Eq. 3.42 so the probability is not zero) equal to 0, 1, 2, 5 and 10.

- Random Forest: 500 trees and 1, 2 and 4 variables randomly selected out of the input covariates to fit each tree.

- Gradient Boosting Machine: interaction depth of 4, 6 and 8 levels and learning rates of 0.1, 0.01 and 0.001.

The second simulation used data from the 2012 edition of the Spanish Living Conditions Survey. A sample of $28,210$ individuals (after applying a filtering procedure to the original sample of size $33,573$) and 61 variables was used as a pseudopopulation from which $s_r$ and $s_v$ were obtained across 500 simulation runs, with $n_r = 500$ and $n_v = 500, 750, 1000, 2000, 5000$. $s_r$ was drawn with SRSWOR from the full pseudopopulation, while $s_v$ was drawn with SRSWOR from the subset of the pseudopopulation with a computer at home. Two parameters were estimated: the proportion of the population living in a household with more than two members, and the proportion of the population whose self-reported health was poor. To do so, PSA was applied using the same algorithms as in the first simulation study, but the hyperparameters were selected via an optimization process this time, using cross-validation to find the combinations of hyperparameters (across the same grid) that minimized the log-Loss of the propensities. In addition, four different combinations of covariates were considered: demographic covariates, demographic plus health covariates, demographic plus poverty-related covariates, and all eligible covariates.

## 3.6. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

We define a population parameter of interest, $\theta_N \in \mathbb{R}^p, p \geq 1$, as the solution of the census estimating equations

$$U(\theta_N) = \frac{1}{N} \sum_U u_i(y_i, \theta_N) \tag{3.45}$$

where $u_i(y_i, \theta_N)$ is a function of $\theta_N$, which can define a general parameter of the variable of interest $y$ that can be linear (mean, total) or nonlinear (distribution function, quantiles). The estimate $\hat{\theta}$ is defined as the solution of Eq. (3.45),

$$\hat{U}(\theta_N) = \sum_U \frac{R_i u_i(y_i, \theta_N)}{\pi_{vi}} = \sum_{s_v} \frac{u_i(y_i, \theta_N)}{\pi_{vi}} = 0, \tag{3.46}$$

which is unbiased for the selection mechanism of $s_v$, this is, the true propensity scores model. If we assume that the selection mechanism is a case of Poisson random sampling, the solution of the equation above provides a consistent (Godambe and Thompson, 1974) and normally distributed (Binder, 1983) estimator for $\theta_N$. This assumption, which considers that the second order inclusion probabilities are the product of first order inclusion probabilities such that $\pi_{vij} = \pi_{vi}\pi_{vj}$, also allows to obtain an expression of the variance of $\hat{\theta}$:

$$Var(\hat{\theta}) = J(\hat{\theta})^{-1} Var(\hat{U}(\theta_N)) J(\hat{\theta}) \tag{3.47}$$

where

$$J(\hat{\theta}) = \frac{1}{N} \sum_U \frac{\partial u_i}{\partial \theta} \tag{3.48}$$

and

$$Var(\hat{U}(\theta_N)) = \sum_U \frac{(1 - \pi_{vi}) u_i^2}{\pi_{vi}^2} \tag{3.49}$$

Given that actual inclusion probabilities $\pi_{vi}$ are unknown in a nonprobability sampling framework, PSA must be applied to obtain estimates of the propensities to be used in the estimators described above. If we assume that the propensities can be modeled such that

$$\pi_{vi} = m(\lambda_0, \mathbf{x}_i), i \in U, \tag{3.50}$$

where $m(.,.)$ is a function twice differentiable with respect to an unknown parameter $\lambda_0$, we can obtain the Maximum Likelihood Estimator (MLE) of $\pi_{vi}$ as $m(\hat{\lambda}, \mathbf{x}_i)$, where $\hat{\lambda}$ is the value that maximizes the log-likelihood function:

$$l(\lambda) = \sum_{s_v} log \frac{m(\lambda, \mathbf{x}_i)}{1 - m(\lambda, \mathbf{x}_i)} + \sum_U log\,(1 - m(\lambda, \mathbf{x}_i)) \tag{3.51}$$

In practice, we consider the value that maximizes the pseudo-likelihood function, which only uses sampled units from the population:

$$\tilde{l}(\lambda) = \sum_{s_v} log \frac{m(\lambda, \mathbf{x}_i)}{1 - m(\lambda, \mathbf{x}_i)} + \sum_{s_r} \frac{1}{\pi_{ri}} log\,(1 - m(\lambda, \mathbf{x}_i)) \tag{3.52}$$

The estimator provided by the nonprobability sample, $\hat{\theta}_v$, can be obtained by calculating $\hat{\lambda}_{pl}$ via solving the score equations

$$\partial\tilde{l}(\mathbf{x}_i, \lambda)/\partial\lambda = 0 \tag{3.53}$$

and obtaining the solution of the estimating function

$$\hat{U}_V(\theta) = \sum_U \frac{R_i u_i(y_i, \theta)}{m(\hat{\lambda}_{pl}, \mathbf{x}_i)} = 0 \tag{3.54}$$

We study the properties of $\hat{\theta}$ and $Var(\hat{\theta})$ under a PSA approach, developing a theoretical framework for estimation of general parameters and deploying a simulation study to obtain empirical results on the efficiency of the estimators in real world applications. The 2012 edition of the Spanish Living Conditions Survey was used for the simulation study, where two sampling schemes were studied for the reference sample: a stratified cluster design, where the strata were the NUTS2 regions and the clusters were the households (with probabilities proportional to the household size), and a unequal probability sampling design, with probabilities proportional to the income. The convenience sample was drawn with an unequal probability sampling, where the inclusion probabilities depended on the gender, the age, the population density of the living area, and having a computer at home.

## 3.7. Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

The choice of the covariates is largely relevant regarding the efficiency of PSA at removing selection bias. Literature on PSA for reweighting of treatment and control group in non-randomised studies (Hirano and Imbens, 2001; Brookhart et al., 2006; Austin, 2008; Schneeweiss et al., 2009; Austin, 2011; Myers et al., 2011; Patrick et al., 2011; Austin and Stuart, 2015) show that selection of variables via expert knowledge or statistical procedures such as stepwise can be fruitful, with better results when the selected covariates are related to the variable of interest or both to the variable of interest and the randomization mechanism.

Although literature on PSA in the context of nonprobability samples adjustment recommends the use of all available covariates (Lee, 2006), the use of variable selection techniques such as the LASSO regression have been considered and studied (Breidt et al., 2017; Chen et al., 2019). The objective of this work is to evaluate the impact of using a set of covariates $\mathbf{x}'$ of length $p' \leq p$ with $p$ being the number of covariates originally considered. Selecting a subset of $p'$ relevant variables leads to less complex and more stable models for propensity estimation that can result in lower variances of the estimators.

We have studied the performance of some variable selectors and filters in two simulation studies (one with a fictitious simulated population and another one with a real world survey dataset used as pseudopopulation). The first simulation study used a fictitious simulated population of size $N = 500,000$, from which $s_r$ and $s_v$ were drawn (with $n_r = n_v = 1000$) with SRSWOR and unequal probability sampling respectively, with eight covariates, eight variables of interest and a inclusion probability for $s_v$ for each individual (which was not used in estimation to emulate a nonprobability survey context). Four of the covariates were not related to any other variable, while the remaining four were related both to the inclusion probability and four of the target variables. Four covariates were used in Raking calibration; two out of the four not in the first group and two out of the four in the second group. Out of the eight target variables, two were not related to any other variable, two were related to the inclusion probability, two were related to the four covariates described above and the remaining two were related both to the inclusion probability and the four covariates described above. This configuration enabled the study of variable selection methods under different missing data mechanisms.

The second simulation study used the dataset of the January 2019 Barometer Survey conducted by the Spanish Centre for Sociological Research (CIS). The sample of $n = 2,156$ individuals (filtered to remove extreme cases and missing data from the original dataset of $n = 2,989$ individuals) and 17 selected variables was bootstrapped up to a size of $N = 500,000$ individuals to be used as a pseudopopulation. In the 17 selected variables, there were 6 target variables, 10 covariates (out of which 3 were used in Raking calibration) and a variable measuring the use of internet in the three months prior to the survey as the delimiter of the internet population. Again, $s_r$ and $s_v$ were drawn (with $n_r = n_v = 1000$) with SRSWOR from the full population and the internet population respectively.

The variable selection approaches used were the following:

- Correlation-based Feature Selection (CFS) (Hall, 1999).
- Chi-Square filter based on Cramer's V.
- Gain ratio (Quinlan, 1986).
- One-R (Holte, 1993).
- Random Forest importance filter (Breiman, 2001).
- Boruta (Kursa and Rudnicki, 2010).
- LASSO regression (Tibshirani, 1996).

Variables selected by these algorithms would be used for PSA with logistic regression, kNN, GBM and neural networks for propensity estimation. In addition, two options were considered for the final weights: direct PSA weights (using the formula $w_i^{PSA1} = 1/\hat{\pi}_{iv}^*$) or Raking calibration weights using $w^{PSA1}$ as initial weights. For those approaches that provide a variable importance index (Chi-Square filter, gain ratio, One-R and Random Forest), the set of covariates $\mathbf{x}'$ was selected using the largest difference criteria for cut-off. In the rest of cases, the algorithms provided a list of selected variables which could sometimes be empty; if such thing happened, PSA would not be applied and therefore the weights remained unitary.

## 3.8. Estimation using Statistical Matching

A model-based alternative for estimation of a parameter of $y$, given a reference probability sample $s_v$, is Statistical Matching or Mass Imputation. We assume that the population $U$ of size $N$ is a realization of a superpopulation model $m$ such that:

$$y_i = m(\mathbf{x}_i) + e_i, i = 1, ..., N, e \sim N(0, \sigma^2) \tag{3.55}$$

where $m(\mathbf{x}_i) = E_m[y_i|\mathbf{x}_i]$. This method implies that the relationship between $y$ and a set of covariates $\mathbf{x}$ can be described with a model $SM$, and therefore the imputation of values for $y$ in $s_r$ such that

$$\hat{y}_j = E_{SM}[y_j|\mathbf{x}_j, R_j], j \in s_r \tag{3.56}$$

enables the obtention of unbiased estimates for population parameters of $y$ through the usual estimators (described in Section 3.1) as long as the model $SM$ is properly specified:

$$\hat{T}_y^{SM} = \frac{\sum_{i\in s_r} \hat{y}_i}{\pi_i} \tag{3.57}$$

$$\hat{\overline{Y}}^{SM_{HT}} = \frac{1}{N}\frac{\sum_{i\in s_r} \hat{y}_i}{\pi_i}. \tag{3.58}$$

$$\hat{\overline{Y}}^{SM_{H}} = \frac{\sum_{i\in s_r} \hat{y}_i/\pi_i}{\sum_{i\in s_r} 1/\pi_i}. \tag{3.59}$$

Note that the model $SM$ must be fitted using data from $s_v$, where the variable of interest $y$ has been measured. Under the ignorability assumption $P(R_i = 1|\mathbf{x}_i, y_i) = P(R_i = 1|\mathbf{x}_i)$, the selection bias should not affect the reliability of $SM$ if the right covariates are used for modeling.

In the first reference to this approach (Rivers, 2007), the imputation was done by giving to individual $j \in s_r$ the value of the closest match (according to $\mathbf{x}$) to $j$ in the nonprobability sample. This approach is equivalent to the case where the model $SM$ is a k-Nearest Neighbors algorithm with $k = 1$. The theoretical properties of this particular case were developed eleven years later in Yang and Kim (2018) (although the earlier work of Beaumont and Bissonnette (2011) developed some properties and the variance estimator for mass imputation in the nonresponse context, known as composite imputation), who showed that the matching estimator is consistent under certain assumptions, and gave an expression for the estimator of the variance as well, but pointed out that the asymptotic bias of the estimator is not negligible and the use of kNN is subject to the curse of dimensionality. The article also considers the application of other prediction models such as Generalized Additive Models (GAM). Kim et al. (2018) developed the theoretical properties of Mass Imputation with semi-parametric linear models as predictive models for $SM$, proving their unbiasedness if the coefficients of the models are equal to the true coefficients (i.e. the model is properly specified). Chen et al. (2020a) developed the theoretical framework and properties for the application of nonparametric models in mass imputation of nonprobability samples, including GAM and kernel smoothing.

## 3.9. Estimation from nonprobability samples with population values for auxiliary variables

If a census of the full target population $U$ is available for some covariates, this is, we have observed $\mathbf{x}_i, i \in U$, we can use an approach very similar to Statistical Matching which has been largely developed in literature throughout the years. This approach is summarized in the model-based adjustments for estimation of population parameters, which can be applied using different formulas.

In model-based estimation, we assume the superpopulation model from Eq. (3.55), but this time we are able to extend the prediction of unobserved values of $y$ to the whole population through a model $SP$:

$$\hat{y}_j = E_{SP}[y_j|\mathbf{x}_j], j \in U \tag{3.60}$$

Let $s_d$ be a probability sample drawn from $U$ with a sample design $(p_d, s_d)$. The simplest approach in the context of superpopulation modelling is the model-based estimator, which is defined for estimation of the population total as follows:

$$\hat{T}_y^{MB} = \sum_{i \in s_d} y_i + \sum_{j \in U - s_d} \hat{y}_j \tag{3.61}$$

Royall (1970) developed the theoretical properties of this estimator in the context of probability sampling for the case where the model $SP$ is a linear regression one, showing that $\hat{T}_y^{MB}$ is unbiased as long as the estimated coefficients of the linear regression model, $\hat{\beta}$, are unbiased. Cassel et al. (1976) developed the model-assisted estimator in probability sampling, which can be defined for estimation of the population total as follows:

$$\hat{T}_y^{MA} = \sum_{i \in U} \hat{y}_i + \sum_{j \in s_d} w_j (y_j - \hat{y}_j) \tag{3.62}$$

where $w$ is the vector of weights (typically design weights but could be any type of adjustment weights) of $s_d$. In the particular case where $SP$ is linear, $\hat{T}_y^{MA}$ is also named difference estimator and is equivalent to the general regression estimator (Deville and Särndal, 1992). The model-assisted estimator, as shown in Breidt et al. (2017), is asymptotically design-unbiased for any model $SP$ and has a known variance whose expression is

$$Var(\hat{T}_y^{MA}) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) w_k (y_k - \hat{y}_k) w_l (y_l - \hat{y}_l) \tag{3.63}$$

Another case of superpopulation modelling estimator is the model-calibrated estimator, developed in Wu and Sitter (2001) which can be defined as follows:

$$\hat{T}_y^{MC} = \sum_{i \in s_d} w_i^{MC} y_i \tag{3.64}$$

where $w^{MC}$ is a vector of calibrated weights where the calibration variable is the variable representing the fitted values of $SP$, $\hat{y}$, such that $w^{MC}$ minimizes the distance with the design weights of $s_d$ while respecting the calibration equations:

$$\frac{1}{N} \sum_{i \in s_d} w_i^{MC} = 1 \qquad \sum_{i \in s_d} w_i^{MC} \hat{y}_i = \sum_{j \in U} \hat{y}_j \tag{3.65}$$

The restriction $\frac{1}{N} \sum_{i \in s_d} w_i^{MC} = 1$ can be dropped out, leading to an alternative estimator $\hat{T}_y^{MC^*}$. Both estimators are equivalent to the general regression estimator if $SP$ is a linear regression model.

These estimators have been developed in the context of probability sampling. If we assume that the avaliable sample is a nonprobability sample $s_v$, we can reformulate the estimators. That was done in Buelens et al. (2018) for the model-based estimator, which in the presence of $s_v$ can be rewritten as

$$\hat{T}_y^{MB} = \sum_{i \in s_v} y_i + \sum_{j \in U - s_v} \hat{y}_j \tag{3.66}$$

An adaptation for the model-assisted and model-calibrated estimators was developed in Rueda et al. (2020) for the nonprobability sampling context such that

$$\hat{T}_y^{MA} = \sum_{i \in U} \hat{y}_i + \sum_{j \in s_v} w_j (y_j - \hat{y}_j) \tag{3.67}$$

where $w$ is a vector of weights defined for the nonprobability sample $s_v$, and

$$\hat{T}_y^{MC} = \sum_{i \in s_v} w_i^{MC} y_i \tag{3.68}$$

where $w_i^{MC}$ minimize the distance with the initial weights of the nonprobability sample while satisfying the calibration equations

$$\frac{1}{N} \sum_{i \in s_v} w_i^{MC} = 1 \qquad \sum_{i \in s_v} w_i^{MC} \hat{y}_i = \sum_{j \in U} \hat{y}_j \tag{3.69}$$

Note that the lack of a sampling design for $s_v$ makes the theoretical properties on the estimators not applicable in this case. On the other hand, and similarly to the case of calibration in nonprobability sampling, the lack of design weights affects the model-assisted and model-calibrated estimators. The solutions to the issue might be the same as in the calibration case.

## 3.10. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

We conduct a study to evaluate the efficiency of model-based estimators in the nonprobability online surveys context, using alternatives to linear regression in the modelization of $SP$. The study is based in three experiments with real world datasets conforming pseudopopulations of simulation experiments, with 500 runs per simulation and three different sample sizes (1000, 2000 and 5000) for $s_v$.

The first simulation is based on data from the 2012 edition of the Spanish Living Conditions Survey following the same procedure as in Section 3.5. In this case, $s_v$ was drawn with two different sampling schemes: SRSWOR from the population with a computer at home, and unequal probability sampling with the inclusion probability defined as a quadratic function of the age (the younger, the more likely to be included in $s_v$). The second simulation is based on the BigLucy dataset (Gutiérrez, 2009) on financial data from $N = 85,396$ industrial companies which were used as the pseudopopulation. Again, $s_v$ was drawn following two different sampling schemes: SRSWOR from the medium-sized and big-sized companies without SPAM options, and unequal probability sampling with inclusion probabilities proportional to the income tax of each company (the larger, the more likely to be included in

$s_v$). The third simulation is based on the Bank Marketing Data Set from Moro et al. (2014) which comprises data of $N = 41,188$ phone calls made in the context of a marketing campaign. The two sampling designs tried for selecting $s_v$ were SRSWOR schemes; in the first case, the sample was drawn from those who were contacted more than three times, and in the second case, from those who were contacted more than twice.

In each simulation, the population means of the selected variables of interest were estimated via model-based, model-assisted and model-calibrated estimators, using data from $s_v$ to train the models and assuming unitary design weights for each individual of $s_v$. The approaches used for the modelization of $SP$ were:

- Generalized linear models (GLM).
- LASSO regression, with and without Bayesian priors.
- Ridge regression.
- Bagged trees.
- Gradient Boosting Machine (GBM).
- Neural networks, with and without Bayesian regularization.

## 3.11. Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

In surveys with multiple variables of interest, also named multipurpose surveys, the use of model-based approaches such as estimators based on the superpopulation model assumption can be impractical. Those approaches would require one adjustment per variable of interest, which would result in the need for multiple model specifications and predictions, and therefore be impractical or even unfeasible if the number of variables of interest is large.

For this reason, reweighting approaches are a more appropiate option for multipurpose surveys. These approaches allow the adjustment of such samples to mitigate selection bias, but only require one modelization step and provide a single vector of weights that can be used for every variable of interest. However, as the efficiency of the weights at removing selection bias depends on the covariates used for their estimation, their use can be more adequate in some variables than in others. As discussed in Section 3.7, adjustments may be better if the covariates used to estimate the weights are related to the variable of interest. If we consider a single vector of weights for every variable, it might work well for some variables but not so well for other ones, as the covariates used for their estimation might be related only to certain variables of interest, or strongly related to some variables of interest but weakly related to other ones.

Weight smoothing is a technique developed in Beaumont (2008), in the context of probability sampling, to deal with this kind of situation. It is based on the assumption that the adjustment weights $w$ of a given probability sample $s$ are related to the variables of interest $\mathbf{y}$ through some measurable function with a random noise term:

$$w_i = f(\mathbf{y}_i, \gamma) + e_i, i \in s \tag{3.70}$$

where $\gamma$ is a vector of unknown parameters and $e$ is a random variable with $E[e] = 0$ and $Var(e) < +\infty$. Weight smoothing is based on fitting a model $WS$ which represents that relationship, and substituting the original weights by the smoothed weights, $\tilde{w}$, which are the predictions provided by $WS$ for each individual in $s$:

$$\tilde{w}_i = E_{WS}[w_i|\mathbf{y}_i], i \in s \tag{3.71}$$

Beaumont (2008) showed that this approach provides unbiased and efficient estimates with the Horvitz-Thompson estimator. In the nonprobability survey context, the design weights that would occupy the place of $w$ are not available, but instead the vector of weights can be a vector of calibration weights or estimated propensity weights obtained with PSA. The model formulated in Eq. 3.70 can then be redefined for a nonprobability sample $s_v$ such that

$$w_i = f(\mathbf{y}_i, \gamma) + e_i, i \in s_v \tag{3.72}$$

and the smoothed weights can be defined for $s_v$ as well, with $WS$ fitted using data from the nonprobability sample (as it might be the only data source where $\mathbf{y}$ has been observed) such that

$$\tilde{w}_i = E_{WS}[w_i|\mathbf{y}_i], i \in s_v \tag{3.73}$$

Two simulation studies were performed for the attainment of the objectives. The first study used a fictitious simulated population of size $N = 500,000$ with 10 covariates, 10 variables of interest and a variable measuring the inclusion probability, with an U-shaped distribution, for each individual. The covariates were not related to each other, but three of them were related to three variables of interest, and another three of them were related to the inclusion probabilities. Two scenarios, representing the extreme cases, were considered: in the first one, there was no relationship between the variables of interest and the inclusion probabilities, while in the second one every variable of interest was related to the inclusion probabilities. The reference sample

$s_r$ was drawn with SRSWOR from the full population, and the convenience sample $s_v$ was drawn with unequal probability sampling using the vector of inclusion probabilities.

The second simulation study used data from the 2012 edition of the Spanish Living Conditions Survey following the same procedure as in Section 3.5 but bootstrapping the population up to $N = 1,000,000$ at the end and removing refusal answers for a final pseudopopulation size of $N = 990,838$, and the same two scenarios for the selection of $s_v$ as in Section 3.10. In this case, 10 variables were chosen as variables of interest: 3 related to house issues, 3 related to deprivation, 2 related to health and 2 random variables generated via simulation procedures which were not related to any other variable. Two groups of covariates were defined: the first group had 9 demographic covariates, and the second group had 8 variables measuring economic and material deprivation.

Each simulation had 500 runs, where $s_r$ and $s_v$ were drawn with equal sample sizes ($n_r = n_v = 1000$). In each one, two procedures were applied to obtain adjustment weights: PSA and Tree-Based Inverse Propensity Weighted estimation (TrIPW) (Chu and Beaumont, 2019), which uses a modified version of the CART algorithm (Breiman et al., 1984) and the design weights from $s_r$ (which were unitary in this simulation, as $s_r$ was always drawn with SRSWOR) to estimate propensities. The propensities were transformed into weights using the formula from Valliant (2020), $w_i^{PSA1} = 1/\hat{\pi}^*_{vi}, i \in s_v$, and finally those weights were substituted by the smoothed weights $\tilde{w}$ by fitting models which used $\mathbf{y}$ as the input predictors and $w_i^{PSA1}$ as the target variable. Two modelling approaches were used for weight smoothing: XGBoost and LASSO. Finally, population means of $\mathbf{y}$ were estimated through the Hajek estimator using $\tilde{w}$ as the estimator weights.

## 3.12. Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment

We deployed the methods for adjustment of nonprobability online surveys in the study of self-perceived health and life satisfaction among healthcare professionals (HCP) in the Spanish region of Andalusia. This study conducted an online survey among the students (university graduates working in the Andalusian Health System) of an online course in holistic care for patients with chronic pain organized by the Andalusian School of Public Health in 2014. The final sample size was $n = 1,797$ and the objective of the study was to measure several variables related to health status (hours of sleep, alcohol consumption, discapacity, presence of chronic diseases or health problems, satisfaction with life and self-perceived health). In addition, ordinal regression models were developed for the study of factors associated to the two latter variables: satisfaction with life (10-point Likert scale) and self-perceived health (5-point Likert scale).

The target population of the survey was the population of Andalusian HCP with an university degree ($N = 73,465$), meaning that the survey was non-probabilistic as the probability of each member of the population to take the survey was unknown, as their probability to take the online course was unknown, and the selection did not follow any scheme: it was administered (and responded) by all the participants of the course. This eventually meant that each member indirectly self-selected to participate in the online survey, which can be a source of selection bias given that the population of HCP with internet access or interest in the course might have different characteristics than the general population of HCP.

In order to correct the selection bias, the parameters of interest were estimated via Propensity Score Adjustment, using the full census of the population (which was available for age, sex, healthcare area and degree subject area, which were the covariates used for propensity estimation) as the reference sample. The use of full censuses as reference samples for PSA has been acknowledged in literature as a possibility for data integration (Elliott and Valliant, 2017). Estimated propensities were transformed into weights using inverse propensity weighting $w_i^{PSA1} = \frac{1}{\hat{\pi}^*_{vi}}, i \in s_v$ (Valliant, 2020) and the propensity stratification approach proposed in (Lee and Valliant, 2009). The results of both approaches (Hajek in the first case, Horvitz-Thompson in the second given that weights were unitary in both cases) were compared in terms of stability and similarity of the weights, and variability of the estimates on the population proportion of some characteristics (<7 hours of sleep, dissatisfaction with life, presence of discapacity, chronicity or health problems, poor self-perceived health and consumption of alcohol at least once a week). The weights were also used in ordinal regression models for the study of factors associated to self-perceived health and life satisfaction.

# Chapter 4

# Results

## 4.1. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

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

- Propensity Score Adjustment provides an important reduction in bias if the selection mechanism is Missing at Random (MAR), and a noticeable but modest reduction if the selection mechanism is Missing Not At Random (MNAR).
- The use of calibration after PSA provides different results according to the calibration variable used. Using education provides modest gains in terms of bias removal, while using nationality adds bias to the estimates. Using gender, which is poorly related to any other variable, does not have any effect on the estimates.
- Bias of the estimates on voting intention for each party does not vary across sample sizes or population totals used for calibration (actual or estimated).
- Standard deviation of the estimates, which drop considerably as the sample size increased, is consistently larger for PSA and PSA + calibration than calibration only for larger sample sizes when the selection mechanism is MAR or NMAR.
- If the selection mechanism is MCAR, using gender as the calibration variable provides an increase in standard deviation of the estimates only when the calibration totals used are the estimated ones (using reference sample data) when the convenience sample size is large. In the rest of the cases, the standard deviation is similar across adjustments.

- The transformation of estimated propensities into weights for Hajek or Horvitz-Thompson estimators does not make any difference in the bias reduction or the standard deviation of the estimates in the first two situations (where age and education are used as PSA covariates). In the other two situations (where age and nationality are used as PSA covariates), weights for Horvitz-Thompson estimators are associated to greater bias reduction than weights for Hajek estimators when the selection mechanism is MAR, while the opposite situation occurs when the selection mechanism is MNAR.

- In the situations where age and nationality are used as PSA covariates, weights for Horvitz-Thompson estimators are associated to slightly smaller variances than weights for Hajek estimators under any selection mechanism when the sample size of the convenience sample is small.

Regarding the application study, PSA using all the available covariates (age, gender and CAST score) without calibration provided the closest estimation to the mean SDS score estimated by the reference sample, while the smaller estimated standard error (according to the results from jackknife) corresponded to the estimation using PSA (without further calibration) with gender and CAST score as covariates.

## 4.2. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

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

- Approaches based on decision trees (C4.5, C5.0 and CART) are not advantageous in terms of bias reduction in any situation or volunteering scheme, especially when compared to PSA with logistic regression, except for the case where the selection mechanism is MCAR and the convenience sample size is small. Mean Square Error (MSA) converged to the unadjusted case. There were small differences across decision-tree algorithms, which narrowed even more as the convenience sample size increased, and parameter tuning only had a noticeable but small effect when the selection mechanism was MCAR and the inclusion probability for the convenience sample was proportional to age.

- The use of k-Nearest Neighbours algorithm (k-NN) provided larger bias reductions than PSA with logistic regression in all situations and volunteering schemes, as well as smaller MSE in the majority of the situations. The choice of $k$ is relevant, as it is shown that larger values of $k$ provide better results overall when the selection mechanism is MAR or MNAR, while lower values of $k$ provide lower bias and MSE when the selection mechanism is MCAR.

- The use of Naïve Bayes classifier in PSA for the case where the convenience sample is drawn with SRSWOR from the computer population is associated to a greater bias reduction in comparison to the case of PSA with logistic regression when the selection mechanism is MAR, but this advantage does not apply to the MCAR and MNAR situations except when the convenience sample size is small and Laplace smoothing is substantial. In the case of unequal probability sampling for the convenience sample, Naïve Bayes provides better results than logistic regression, decision trees or k-NN, both in terms of bias reduction and MSE, when the selection mechanism is MNAR.
- Random Forest in PSA provides substantially better results in bias reduction and MSE than the rest of the alternatives for propensity estimation when the selection mechanism is MNAR, regardless of the sampling scheme for the convenience sample. However, its performance depends on the number of covariates sampled to fit the trees and the convenience sample size. For smaller samples, fewer covariates provide better results, while for larger samples more covariates provide better results. In MCAR or MAR situations, Random Forest provide worse results than the rest of alternatives.
- GBM for propensity estimation gives very different results depending on the hyperparameter configuration. The algorithm removes bias to a greater extent when the learning rate is kept in low values, especially when the selection mechanism is MAR. In terms of MSE, the best results are obtained when the learning rate is low in MCAR situations, and when the learning rate is high in MAR and MNAR situations.

The results of the simulation study using real data from the 2012 edition of the Spanish Living Conditions Survey are summarized in the following points:

- Estimation of the population fraction who perceive their health to be poor.
    - When only demographic covariates are available, PSA with Naïve Bayes provides the largest bias reduction among all the algorithms studied, regardless of the convenience sample size. However, Naïve Bayes induces an instability that makes it provide the largest variance among all the methods and, as a consequence, a larger MSE. In this case, the method that provides the estimation with the smallest MSE is logistic regression.

    - The results regarding capacity across methods to improve estimations are the same when health or poverty covariates are available and included in the models. However, it is noticeable that the bias reduction is larger than in the case where only demographic covariates are available; in this sense, bias reduction is larger when poverty covariates are used, which is the group of covariates related to the variable of interest.
    - When all covariates are used, PSA with logistic regression leads to almost unbiased estimates. However, the variance of the estimates is large, and therefore the MSE for PSA with logistic regression presents poor values. The best choice, in terms of MSE, depends on the convenience sample size: if it is small, decision trees present the smallest term of error, and if it is large, GBM and Random Forest occupy that place.
    - Across all the algorithms, sample sizes and groups of covariates used, the lowest MSE is observed in the case where PSA with logistic regression and demographic and health covariates are used. Its MSE is close to that of PSA with GBM using all of the covariates.

- Estimation of the population fraction who live in households with more than two members.

    - When only demographic covariates are available, PSA with Random Forest provides the largest bias removal and the smallest MSE out of all the algorithms studied when the convenience sample size is large, but the opposite behavior is observed when the convenience sample size is small. In the latter case, the performance of the rest of algorithms is very similar between them, although PSA with logistic regression seem to provide better results both in terms of bias removal and MSE.
    - When health or poverty covariates are available in addition to demographics, the same patterns than in the case with demographics only can be observed (albeit a limited bias reduction is observed in comparison to the aforementioned case), except for the fact that, when poverty covariates are used, PSA with logistic regression provides the estimates with the smallest MSE out of all the algorithms studied.
    - When all covariates are included, J48 and GBM are the algorithms that provide the smallest MSE when small and large convenience sample sizes, respectively, are available. Naïve Bayes presents the largest bias reduction for all sample sizes, but its variability leads to poorer MSE values.

- Across all the algorithms, sample sizes and groups of covariates used, the lowest MSE is observed in the case where PSA with Random Forest and demographic covariates are used. Its MSE is close to that of PSA with GBM using all of the covariates, and PSA with Naïve Bayes using all of the covariates as well.

## 4.3. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

For a nonprobability sample which it is assumed to constitute a realization of a Poisson sampling, it is proven that:

- A consistent and asymptotically normal estimator of a general population parameter, $\hat{\theta}_v$, can be obtained via a two step optimization procedure under the assumptions of double-differentiability for the estimating function, and unbiasedness and normality for the estimating equation.

- An expression for the asymptotic variance-covariance matrix of the estimators of a general population parameter and the population value of $\lambda_{MLE}$ can be obtained under the assumptions of double-differentiability for both the estimating function and the estimating equation, and unbiasedness and normality for the estimating equation.

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

- The relative mean bias in the estimation of the Gini index, poverty risk (HCI), interquartile range (IQR) and interdecile range (IDR) is substantially lower after the application of PSA with logistic regression, especially in the estimation of HCI. The reduction in relative bias was larger when the reference sample is drawn with inclusion probabilities proportional to the income, in comparison to the case where it is drawn with stratified cluster sampling (with probabilities proportional to the household size).

- The standard deviation of the estimates for the Gini index, IQR and IDR are largely similar in all cases, both the adjusted and the unadjusted ones, although small samples provide smaller standard deviations in the case where the reference sample is drawn with inclusion probabilities proportional to the income. However, PSA-adjusted estimates of HCI have a substantially larger variability.

- The Root Mean Square Error (RMSE) of the Gini index, HCI, IQR and IDR estimates is remarkably lower in the cases where PSA is applied.

Once again, the reduction in RMSE is larger when the reference sample is drawn with inclusion probabilities proportional to the income.

- Regarding the estimation of HCI under a cubic relationship and a cosine-shaped relationship between the age and the inclusion probability in the convenience sample, PSA provided estimates with lower relative bias and RMSE in all of the situations considered.

## 4.4. Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

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

- Variable selection provides estimates with a slightly reduced relative bias for MCAR variables when the algorithm used to estimate propensities is neural networks. However, the least biased results are observed when using PSA with kNN and no variable selection or StepWise selection. Reductions in MSE provided by variable selection are more noticeable. Generally speaking, variable selection improves the efficiency of the estimates for obtained from PSA + Raking calibration. In the case of one of the MCAR variable, the MSE was 10 % lower when k-NN and Raking are used to estimate it, with the Stepwise algorithm (in comparison to using all variables). In the absence of Raking calibration, the Chi-square, Gain Ratio, LASSO and OneR methods provide reductions of 7 % in MSE.

- In situations where the selection mechanism is MAR, variable selection are able to reduce the relative bias for the majority of situations studied (especially when PSA is performed using logistic regression), but the most relevant intervention in that sense is Raking calibration, which markedly reduces the bias in the estimates. The use of several variable selection methods leads to a significant improve of estimates' efficiency, especially in the case where Raking calibration is included among the adjustments, with MSE reductions between 10 % and 50 %.

- In terms of relative bias reduction, the estimation of parameters of variables subject to MNAR selection benefit from the application of Raking calibration and variable selection techniques as well, but to a more limited extent than the observed in the MAR case. The reductions in MSE, in comparison to the case where no variable selection techniques are used, are around to 20 % for OneR (when PSA with logistic regression is applied) and Gain Ratio (for all situations as long as the outcome is fixed as the target variable of the selection algorithm) techniques.

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

- The best choice in variable selection depends on the propensity estimation model considered. In the majority of cases, PSA using k-NN provided the best results when using all the available covariates. When other algorithms were used in PSA, selection algorithms improve the results of the estimates, although there is not a single variable selection algorithm that fits all the cases but an ideal algorithm for each case.

- Raking calibration has a modest positive effect on the variables measuring the economic situation in Spain, the preference for a unitary national state without autonomous communities and whether the respondent self identified as only Spanish, but the impact on the estimation of other variables is insignificant.

- Regarding efficiency of the estimators, using variable selection algorithms is associated to reductions in the MSE in all situations. These reductions have a magnitude of 10 % of the MSE using all the available covariates, although it can be up to 20 % in some cases (OneR algorithm in the estimation of perceived economic situation in Spain, CFS in the estimation of ideological self-positioning, and CFS, Chi-square, Gain ratio and Random Forest in the estimation of the preference for an unitary national state).

The results from the application study show that the use of variable selection algorithms in the estimation from the nonprobability sample of students are associated to point estimates closer to the estimate from the reference sample. If we assume that the latter sample is closer to a probability sampling, which is probably the case as it was drawn following a stratified cluster sampling design, we can assume that these results are a proof of bias reduction associated to variable selection algorithms, along with Raking calibration (which also provides estimates closer to the reference sample one). The estimated variance of the estimators, obtained with the jackknife method, the increase in variance induced by PSA is atenuated when variable selection algorithms are used prior to PSA.

## 4.5. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

The results of the simulation study show that model-assisted, model-based and model-calibrated estimators provide important reductions in relative bias and RMSE for all of the datasets and volunteering schemes studied. No differences are observed between any of the three estimators. The behavior of the estimators was also consistent across all the convenience sample sizes.

- For the first dataset (Spanish Living Conditions Survey data), the reduction is maximized by bayesian-regularized neural networks (BRNN) and linear regression methods with and without regularization (GLM, Ridge and LASSO), regardless of the inclusion probability scheme for the convenience sample.

- In the second dataset (BigLucy), the largest reduction in bias and MSE under the first sampling design of the convenience sample is observed when using GLM, Ridge regression (with and without bayesian priors), and Bayesian LASSO regression, as well as BRNN when the sample size is large. However, in the second sampling design (with inclusion probabilities proportional to taxes), the largest bias and MSE reduction corresponds to k-Nearest Neighbors algorithm.

- In the third dataset (Bank Marketing Data Set), the results are very similar to those of BigLucy: the largest reduction in bias and MSE under the first sampling design of the convenience sample is observed when using GLM, Ridge regression, and LASSO regression, as well as Gradient Boosting Machine when the sample size is $n_v = 1,000$ and Bagged Trees algorithm when the sample size is $n_v = 2,000$. However, in the second sampling design (with SRSWOR from people contacted more than twice), the largest bias and MSE reduction corresponds to k-Nearest Neighbors algorithm, and BRNN when the sample size is $n_v = 1,000$

Results of the linear mixed-effects regression confirm that there is no evidence in the simulations' results that Ridge regression, GLM, LASSO maximum-likelihood regression (both bayesian and non-bayesian), k-Nearest Neighbors or Bayesian-regularized Neural Networks provide different effects on the efficiency of the estimates.

## 4.6. Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

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

- The application of weight smoothing does not produce any important change in relative bias, regardless of whether all the target variables are unrelated or directly related to the inclusion probability. In the former case, the relative bias is negligible (given that the selection mechanism is MCAR), while in the latter case a noticeable amount of bias can be observed, although it can be succesfully removed up to half of the original relative bias when using TrIPW. In this case, PSA is also able to reduce the bias of the estimates to a lesser extent.
- Weight smoothing is able to substantially improve the efficiency of the adjusted estimates in the case where the selection mechanism is MCAR, especially when the adjustment is done via TrIPW. In this case, weight smoothing with LASSO regression makes the MSE comparable to the unadjusted case (where no variance is induced by the adjustment).
- In the case where all the variables of interest are directly related to the inclusion probability, the application of weight smoothing did not provide any noticeable change in the efficiency of the estimates.

The results of the simulation study using a real dataset as the pseudopopulation depend on the sampling scheme of the nonprobability sample. When the sample is drawn with SRSWOR from the subset of the pseudopopulation with a computer at home, the results are the following ones:

- When using the demographics covariates for propensity adjustment (in PSA or TrIPW), the bias is reduced (and very similarly for both adjustments) only in the variables measuring self-reported health, number of months working part time, and whether the household has noise problems or not. When using the deprivation covariates, the bias reduction is markedly more generalized amongst all the variables of interest, with TrIPW presenting the best results in this sense. Weight smoothing did not produce any noticeable change in the bias of the estimates.
- The efficiency of the estimates depends on the covariates used in the adjustments. In general, there is always an adjustment that leads to an improve in the efficiency of the estimates, in comparison to the

unadjusted case. The impact of weight smoothing also vary across adjustments; in those cases where the efficiency is below 1 (the proposed method leads to a more efficient estimate than the unadjusted case), the application of weight smoothing is counterproductive. However, when the efficiency is above 1, the smoothing procedure leads to a reduction of the quotient (and therefore, an improval of the efficiency), especially when LASSO regression is applied.

When the sample is drawn with unequal probability sampling, with inclusion probabilities proportional to the age, the results are the following ones:

- The relative bias of the estimates does not get reduced after the application of the adjustments, except for the variable measuring whether the household has a heating system or not, where the application of TrIPW with deprivation covariates leads to a reduction of the bias up to the point of making it almost zero. These results are also observed in the MSE of the estimates.
- This time, weight smoothing is able to improve the efficiency in many situations, including some of those where the MSE of the estimates is already below 1. LASSO regression provides again the largest improvements in efficiency.

## 4.7. Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment

Results of the weighting process lead to the discarding of Horvitz-Thompson weights, given their large variability which induces an increase in the variance of the estimators. The correlation between Horvitz-Thompson and Hajek weights, when using the same algorithms for propensity estimation in PSA, is high in the majority of cases, but Horvitz-Thompson weights present larger skewness values and a considerable number of outliers in some cases.

In addition, Hajek weights obtained from PSA with decision trees and neural networks with five units are discarded as well. The reason is that those algorithms were unable to fit any model, and therefore provided unitary propensities. Those propensities made Hajek weights to be equal to the initial weights of the nonprobability sample, meaning that the results would be equivalent to the unadjusted case.

The Multidimensional Scalling (MDS) analysis of the weights shows two differentiated groups of weights largely similar between them: those provided by PSA with neural networks (regardless of the number of units in the hidden layer), and those provided by PSA with logistic regression, Gradient Boosting Machine or Naïve Bayes. In addition, the analysis also shows three clear outliers: weights provided by PSA with Random Forests (both Horvitz-Thompson and Hajek) and Horvitz-Thompson weights obtained from PSA with a neural net with five units in the hidden layer.

The results of the prevalence analysis are very similar across adjustments. Although some of the point estimates differ slightly, the estimated confidence intervals (at a confidence level of 95 %) overlap in all variables and PSA approaches. In addition, it can be observed that the application of PSA leads to an increase of the estimators' variance in comparison to the unadjusted case. The minimum increase in variance is observed for PSA with logistic regression. For this reason, we use the results provided by PSA with logistic regression as the reference results, which show the following numbers:

- 10.3 % of male HCPs and 12.6 % of female HCPs are dissatisfied with their life and 8.4 % of male and 7.8 % of female professionals perceived their own health as poor. These estimates are significantly lower than the estimates for the general Andalusian population.

- 62.3 % of the men and 42.8 % of the women drank alcohol at least once a week. These estimates are signficantly higher than the estimates for the general Andalusian population.

- 31.1 % of the men and 26.7 % of the women slept for less than seven hours a day. These estimates are signficantly higher than the estimates for the general Andalusian population.

- 31.8 % of the men and 22.3 % of the women reported having at least one chronic disease. The estimate for women HCP is significantly higher than the estimate for the general Andalusian females population, but that is not the case for men.

- 26.3 % of the men and 20.6 % of the women had one health problem, and 10.4 % and 6 %, respectively, had two or more health problems.

- 7 % of men and 6 % of women had a disability. These estimates are signficantly higher than the estimates for the general Andalusian population.

The regression analysis was performed using weights from PSA with logistic regression and PSA with a neural net with one unit, with the objective of representing both of the groups of very similar weights that have been found in MDS. The results were compared to the same regression analysis with no weighting adjustment, showing that weighting makes the estimates of the regression coefficients to be closer to zero, which also influences the strength of the evidences against the null effect of a given factor on the target variable.

The results of the regression analysis are summarized in the following points:

- The prior existence of one health problem increases the likelihood of poor self-perceived health by 3 and 2 times, respectively, for men and women. In the case of two or more health problems, this probability rise to 8 and 10 times, respectively. In addition, there is evidence that the presence of obesity, according to the BMI index, is significantly associated with a lower probability of good health among women (OR = 2.1).

- Nursing qualifications are significantly associated with poorer self-perceived health, compared with respondents with a degree in medicine, regardless of sex (OR = 1.8), or even among women those whose degree subject was reported as neither medicine nor nursing (OR = 2). However, no significant differences in OR are observed between those who worked in primary care or other level of healthcare.

- Smoking every day is associated with a greater likelihood of poorer self-perceived health in women; no physical activity or only occasional activity is also associated with poorer self-perception of health, especially in men, as is sleeping less than seven hours per night.

- The strongest negative association with life satisfaction is measured for prior health problems, and this relationship become significantly stronger for both male and female respondents as the number of pre-existing health problems increases. For men, furthermore, working in primary rather than other levels of healthcare is also associated with less life satisfaction.

- Physical inactivity is associated with lower levels of life satisfaction. Male and female HCPs who performed no physical activity at all are 5 and 2.5 times, respectively, more likely to have less satisfaction with life than their more physically active counterparts.

- Women who smoked (whether every day or less frequently) are more likely to report lower levels of life satisfaction than those who have never smoked.

- HCPs who sleep less than seven hours per night are around 1.5 and 1.8 times (for men and women, respectively) more likely to report low levels of life satisfaction than those who sleep for longer, assuming all other variables remain constant.

# Chapter 5

# Conclusions

## 5.1. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

The adequacy of PSA (with different weighting approaches) and the combination between PSA and calibration is studied through a simulation study that represents all types of missing data mechanisms, and an application study on a real online survey of university students. We also study the behavior of calibration using actual population totals and estimated population totals (from the reference sample) . The results show that the bias can be almost entirely reduced with the right combination of adjustments if the selection mechanism is MAR, while it can be reduced only to a modest extent if the selection mechanism is MNAR. The results support the hypothesis that the combination of PSA and calibration can be more adequate for mitigating bias than using only one of the methods on its own, as long as the covariates for calibration and propensity estimation are properly specified. The observed differences between weighting procedures are not significant in terms of bias or estimators' variance reduction. It can also be concluded that the application of PSA may lead to a larger variance of the estimates in comparison to the unadjusted or calibrated-only case, but the increase depends on the convenience sample size, with larger samples leading to a larger increase of the variance. Finally, it can be concluded that the use of actual population totals or estimated population totals does not make a difference in bias reduction or variance of the estimates (except for the case where selection mechanism is MCAR) as long as the reference sample is representative of the target population.

## 5.2. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

We study the application of Machine Learning classification algorithms in PSA in two simulation studies with fictitious and real world data, considering different hyperparameter configurations for each algorithm tried in the first case. The results show that Machine Learning algorithms can be viewed as an advantageous alternative to logistic regression for propensity estimation, although the latter method is also shown to be a robust, reliable one. There is not a single optimum approach for every case, and the best Machine Learning algorithm in terms of efficiency might be different for every scenario and dataset. When selection follows a MCAR scheme, CART algorithm and Gradient Boosting Machine (GBM) are the best alternatives, although the majority of the algorithms tested also improve upon the results obtained by PSA with logistic regression, especially as the convenience sample size increases. With MAR or MNAR selection, logistic regression generally provides good adjustments, especially when the dimensionality is low and the covariates are not very discriminant. However, if more covariates are available, logistic regression tends to destabilise and the MSE increases, and GBM, k-NN, decision trees and Random Forests all represent good alternatives. The presence of balancing and overfitting issues suggests that data preprocessing should be a key step in propensity estimation.

## 5.3. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

We establish a theoretical framework for estimation of general parameters in nonprobability surveys using PSA to estimate the unknown inclusion probability. We propose a two-phase estimation procedure, in which the first step is focused on finding the value that maximizes the pseudo-likelihood of the model considered to predict the inclusion probability conditional to a set of covariates, and the second one is focused on finding the solution to the estimating equation. We prove that the provided estimators are consistent and asymptotically normal under several conditions. We also provide the expression of the estimator for the variance of the estimators developed in the two-step procedure.

Results observed in the simulation study provide strong evidence on the efficiency of methods based in estimating equations with estimated propensities. However, it must be noted that the efficiency depends on the selection mechanisms of nonprobability samples and the availability of covariates for propensity estimation. In our simulations, results show that PSA is more efficient when the propensity of being in the nonprobability sample is less related to the variable of interest.

## 5.4. Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

We perform two simulation studies using synthetic data and a real survey where PSA is used to reduce selection bias in nonprobability samples, applying variable selection methods to obtain an optimal set of covariates for propensity estimation. We also apply Machine Learning classification algorithms for propensity estimation and Raking calibration after PSA, in order to study the adequacy and efficiency of each technique or combination of techniques.

Our analysis shows that variable selection makes a significant contribution to reducing relative bias, although the best method for finding the best subset of covariates depends on the dataset considered and the adjustment choices made. Variable selection is associated with a reduction of model complexity which leads to more efficient estimators. Selecting variables according to their impact on the outcome variable provided the best results overall.

In addition, we observe that the application of Raking calibration after PSA is the most efficient technique in almost all cases. On the other hand, the use of classification algorithms instead of logistic regression for estimating propensities was only advantageous for certain algorithms, which are different depending on the scenario.

## 5.5. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

We describe some options for estimation in nonprobability samples using ML techniques in three superpopulation modeling based approaches: model-based, model-assisted and model-calibrated estimators. We clarify the requirements for the application of these estimators and illustrate how they perform empirically in three simulation studies using real world datasets. The main conclusion of the simulation study is that the selection of the predictive model used in the process is vastly more important than the approach used in the estimation or the convenience sample size. No significant differences was found in efficiency between any of the three estimators proposed nor any of the three sample sizes studied. Regarding the modeling choice, we observe that advanced regression techniques, such as LASSO regression or Ridge regression, are similar in their good performance, although other techniques such as bayesian regularized neural networks or k-Nearest Neighbors can provide efficient estimates in some specific situations. Linear models, which are the standard choice considered in superpopulation modeling literature and its theoretical framework, provide good results overall.

## 5.6. Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

We apply weight smoothing methods in two simulation studies, using artificial and real world data, in the context of nonprobability sampling. PSA and TrIPW methods are used to estimate propensities, which are transformed into weights using the inverse probability weighting approach. The results of the simulation study with an artificial population show that weight smoothing contributes to reduce the variance of the estimates in those situations where the efficiency of non-smoothed estimates is poor. When the estimates are already efficient, weight smoothing does not add much, with some exceptions where the set of covariates is good but not optimal. The bias of the estimates remains unchanged after the application of weight smoothing.

In the real data simulation, the conclusions that arise from the results are different. The adjusted estimators (with PSA or TrIPW) are largely inefficient in some cases, with weight smoothing being unable to increase their efficiency. This could be caused by misspecified propensity models which might contribute to increasing the bias, making the MSE to increase because of the bias (which is the term that weight smoothing is unable to address) and not the variance.

LASSO regression presents better results than XGBoost in terms of MSE reduction in the majority of cases. Regarding differences in propensity estimation methods, TrIPW provides better results overall than PSA.

## 5.7. Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment

We study a real world problem where the variables of interest of the analysis (prevalence of certain diseases, life satisfaction and self-perception of health) had been measured in an online nonprobability survey. PSA is applied to remove selection bias using different predictive models for propensity estimation and two different weighting procedures. The results on prevalence estimates show that there are some differences across the estimations provided by different adjustments and estimators, although several groups of algorithms for PSA with similar behaviours have been spotted according to the weights that they provide. Estimates provided by the Horvitz-Thompson estimator have larger estimated variances, and Random Forest algorithm provides more extreme weights and therefore skewed vectors of weights, which also contribute to an increase in the variance of the estimates. Some adjustments, especially PSA with logistic regression, present smaller variances, making them more desirable in terms of reducing estimation error.

The regression analysis shows no significant differences across selection bias adjustments, although a shift towards the null hypothesis can be observed for some of the regression coefficients after PSA reweighting. Prior health problems, sleeping for less than seven hours per night, physical inactivity and smoking (by women) are all associated with the perception of poorer health, while obesity (among women), working as a nurse or in primary healthcare (among male HCPs) are associated with less satisfaction with life.

Chapter 6

# Conclusiones

## 6.1. Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys

La adecuación del PSA (con diferentes métodos de ponderación) y la combinación entre PSA y calibración se estudia a través de un estudio de simulación donde se representan todos los mecanismos de datos faltantes, y un estudio de aplicación en una encuesta online real de estudiantes universitarios. Se estudia también el comportamiento de la calibración utilizando totales poblacionales reales y totales poblacionales estimados (a partir de la muestra de referencia). Los resultados muestran que el sesgo puede reducirse casi en su totalidad con la combinación correcta de ajustes si el mecanismo de selección es MAR, mientras que se puede reducir sólo hasta una modesta cantidad si el mecanismo de selección es MNAR. Los resultados apoyan la hipótesis de que la combinación de PSA y calibración puede ser más adecuada para reducir el sesgo que utilizar únicamente uno de los métodos, en tanto en cuanto las variables auxiliares para la calibración y la estimación de la propensión estén correctamente especificadas. Las diferencias observadas entre métodos de ponderación no son significativas en términos de reducción de sesgo o varianza de los estimadores. Se puede concluir también que la aplicación del PSA puede traducirse en una mayor varianza de las estimaciones en comparación con el caso sin ajustar o con sólo calibración, pero el incremento depende del tamaño de la muestra de conveniencia, con mayores tamaños muestrales conllevando un mayor aumento de la varianza. Finalmente, se puede concluir que el uso de totales poblacionales reales o estimados no supone ninguna diferencia en cuanto a reducción del sesgo o de la varianza de las estimaciones (salvo en caso de que el mecanismo de selección sea MCAR) en tanto en cuanto la muestra de referencia sea representativa de la población objetivo.

## 6.2. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

Se estudia la aplicación de algoritmos de clasificación de Machine Learning en el PSA en dos estudios de simulación con datos tanto ficticios como del mundo real, considerando diferentes configuraciones de hiperparámetros para cada algoritmo estudiado en el primer caso. Los resultados muestran que los algoritmos de Machine Learning pueden verse como una alternativa ventajosa a la regresión logística para la estimación de la propensión, aunque este último método se muestra también como uno robusto y fiable. No hay un solo método óptimo para todos los casos, y el mejor algoritmo de Machine Learning en términos de eficiencia puede ser diferente en cada escenario y conjunto de datos. Cuando la selección sigue un esquema MCAR, los algoritmos CART y Gradient Boosting Machine (GBM) son las mejores alternativas, aunque la mayoría de algoritmos probados también mejoran los resultados obtenidos por el PSA con regresión logística, especialmente a medida que el tamaño de la muestra de conveniencia aumenta. Cuando la selección es MAR o MNAR, la regresión logística generalmente produce buenos resultados, especialmente cuando la dimensionalidad es baja y las covariables no son muy discriminantes. Sin embargo, si hay más covariables disponibles, la regresión logística tiende a desestabilizarse y el Error Cuadrático Medio (ECM) aumenta, y el GBM, los k-NN, los árboles de decisión y los Random Forests representan buenas alternativas. La presencia de problemas de balanceo y sobreajuste sugieren que el preprocesamiento de datos debería ser un paso clave en la estimación de propensiones.

## 6.3. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

Establecemos un marco teórico para la estimación de parámetros generales en encuestas no probabilísticas utilizando el PSA para estimar la probabilidad de inclusión desconocida. Proponemos un proceso de estimación en dos fases, en el cual el primer paso se centra en encontrar el valor que maximiza la pseudo-verosimilitud del modelo considerado para predecir la probabilidad de inclusión condicionada a un conjunto de covariables, y el segundo paso se centra en encontrar la solución a la ecuación de estimación. Demostramos que los estimadores proporcionados son consistentes y asintóticamente normales bajo ciertas condiciones. También proporcionamos la expresión del estimador para la varianza de los estimadores desarrollados en el proceso en dos fases.

Los resultados observados en el estudio de simulación proporcionan importantes evidencias de la eficiencia de los métodos basados en ecuaciones de estimación con propensiones estimadas. Sin embargo, se debe notar que la eficiencia depende de los mecanismos de selección de las muestras no probabilísticas y la disponibilidad de covariables para la estimación de la propensión. En nuestras simulaciones, los resultados muestran que el PSA es más eficiente cuando la propensión a estar en la muestra no probabilística está menos relacionada con la variable de interés.

## 6.4. Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys

Realizamos dos estudios de simulación utilizando datos sintéticos y una encuesta real en la que se utiliza el PSA para reducir el sesgo de selección en muestras no probabilísticas, aplicando métodos de selección de variables para obtener un conjunto óptimo de covariables para la estimación de la propensión. También aplicamos algoritmos de clasificación de Machine Learning para la estimación de la propensión y calibración Raking después del PSA, con el objetivo de estudiar la adecuación y la eficiencia de cada técnica o combinación de técnicas.

Nuestro análisis muestra que la selección de variables supone una contribución significativa en la reducción del sesgo relativo, aunque el mejor método para encontrar el mejor subconjunto de covariables depende en el conjunto de datos considerado y en los ajustes realizados. La selección de variables está asociada a una reducción en la complejidad del modelo que se traduce en estimadores más eficientes. Seleccionar variables según su impacto en la variable objetivo proporcionó los mejores resultados en líneas generales.

Además, observamos que la aplicación de la calibración Raking tras el PSA es la técnica más eficiente en casi todos los casos. Por otra parte, el uso de algoritmos de clasificación en lugar de regresión logística para estimar propensiones sólo es ventajosa para ciertos algoritmos, que son diferentes en función del escenario.

## 6.5. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

Se describen algunas opciones para la estimación en muestras no probabilísticas empleando técnicas de Machine Learning en tres métodos basados en modelos de superpoblación: los estimadores modelo basado, modelo asistido y modelo calibrado. Se clarifica los requerimientos para aplicar estos estimadores y se ilustra cómo actúan empíricamente en tres estudios de simulación utilizando conjuntos de datos del mundo real. La conclusión principal del estudio de simulación es que la selección del modelo predictivo utilizado en el proceso es mucho más importante que el método utilizado para la estimación o el tamaño de la muestra de conveniencia. No se encontraron diferencias significativas entre ninguno de los tres estimadores propuestos ni ninguno de los tres tamaños de muestra estudiados. En cuanto a la elección del modelo, observamos que las técnicas avanzadas de regresión, como la regresión LASSO o la regresión Ridge, son similares en cuanto a su buen funcionamiento, aunque otras técnicas como las redes neuronales regularizadas de forma bayesiana o los k-vecinos más cercanos pueden proporcionar estimaciones eficientes en algunas situaciones específicas. Los modelos lineales, que son la elección estándar considerada en la literatura de modelos de superpoblación y su marco teórico, proporcionan en general buenos resultados.

## 6.6. Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

Se aplican métodos de suavizado de pesos en dos estudios de simulación, utilizando tanto datos artificiales como del mundo real, en el contexto del muestreo no probabilístico. Se emplean los métodos PSA y TrIPW para estimar las propensiones, que se transforman en pesos utilizando el método de ponderación de la probabilidad inversa. Los resultados del estudio de simulación con una población artificial muestran que el suavizado de pesos contribuye a reducir la varianza de las estimaciones en aquellas situaciones en las que la eficiencia de las estimaciones no suavizadas es pobre. Cuando las estimaciones ya son eficientes, el suavizado de pesos no añade mucho, con algunas excepciones donde el conjunto de covariables es bueno pero no óptimo. El sesgo en las estimaciones permanece sin cambios tras la aplicación del suavizado de pesos.

En la simulación que emplea datos reales, las conclusiones que surgen de los resultados son diferentes. Los estimadores ajustados (con PSA o TrIPW) son ampliamente ineficientes en algunos casos, con el suavizado de pesos siendo incapaz de aumentar su eficiencia. Esto podría estar causado por la especificación errónea de modelos de propensión que podrían contribuir a incrementar el sesgo, haciendo que incremente el Error Cuadrático Medio (ECM) debido al sesgo (que es el término en el que el suavizado de pesos no es capaz de intervenir) y no la varianza.

La regresión LASSO presenta mejores resultados que el XGBoost en términos de reducción del ECM en la mayoría de los casos. En cuanto a las diferencias entre métodos de estimación de la propensión, el TrIPW proporciona mejores resultados que el PSA.

## 6.7. Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment

Se estudia un problema del mundo real en el que la única información disponible sobre ciertas variables relacionadas con la salud es una muestra no probabilística. Se aplica el PSA para eliminar el sesgo de selección utilizando diferentes modelos predictivos para la estimación de las propensiones y dos métodos de ponderación diferentes. Los resultados de las estimaciones de prevalencia muestran que hay algunas diferencias entre las estimaciones proporcionadas por diferentes ajustes y estimadores, aunque se han identificado algunos grupos de algoritmos para PSA con comportamientos similares, de acuerdo a los pesos que proporcionan. Las estimaciones proporcionadas por el estimador de Horvitz-Thompson tienen mayores varianzas estimadas, y el algoritmo Random Forest proporciona más pesos extremos y por tanto vectores de pesos asimétricos, lo que también contribuye a un incremento en la varianza de las estimaciones. Algunos ajustes, especialmente PSA con regresión logística, presentan varianzas más pequeñas, lo que les hace más deseables en términos de reducir el error de estimación.

El análisis de regresión no muestra diferencias significativas entre ajustes para el sesgo de selección, aunque se puede observar un desplazamiento hacia la hipótesis nula en algunos de los coeficientes de regresión tras la reponderación con PSA. Problemas de salud previos, dormir menos de siete horas por noche, inactividad física y fumar (en mujeres) están asociados con una autopercepción de salud más pobre, mientras que la obesidad (en mujeres) trabajar de enfermera o en atención primaria (en hombres) están asociados con una menor satisfacción con la vida.

# Chapter 7

# Future Research

The research presented in this dissertation entails several limitations that should be accounted for in future research. A major issue is the lack of a theoretical framework for estimation in nonprobability surveys for many of the methods studied. Although such framework has been provided for estimation of any parameter with Propensity Score Adjustment (PSA), the development of the theoretical properties for estimators based in superpopulation modeling in the nonprobability sampling context remains as a challenge to be considered in future research lines. The importance of this development is stated by the promising results provided by empirical research.

Regarding PSA, future research should focus on the inclusion of the design weights, often present in the reference sample, in the propensity estimation procedure. Although an approach can be consulted in Valliant (2020), there is still a lack of research on the path that provides the best results in terms of bias reduction via propensity estimation. The Tree-based Inverse Propensity Weighted estimator method, which has been studied as an alternative to PSA in order to obtain propensity weights, consider the design weights to estimate the population fraction that would fit in the terminal node of a decision tree. The promising results provided by TrIPW in our simulation studies might be a motivation to consider this issue as a future research line.

Another important research line that has already shown its potential is the combination of estimation procedures to reduce bias more significantly. Doubly robust estimators are an example of that combination: they use the results from the Statistical Matching estimator but complement them with a measure of the prediction error provided by the information present in the nonprobability sample, which is weighted using PSA to make it representative for the population. The combination of different approaches might be useful to provide robustness and protection against misspecifications that could be present in the statistical models used in the majority of adjustment methods.

The results presented in this dissertation are also limited by the reduced set of predictive models and missing data situations studied. Although the inclusion of real world datasets makes it possible to take into account complex relationships and phenomena that can take place in data sets, the range of possible situations that may appear in an application could be different from those studied. On the other hand, the algorithms used for propensity estimation or mass imputation, although modern, only account for a minimum part of the state-of-the-art prediction algorithms which could be used for the matter. In addition, the preprocessing step, which has not been exhaustively studied in the context of bias reduction in nonprobability surveys (except for variable selection), could help to provide better results in terms of bias reduction.

Finally, a major issue that should be addressed in future research is the adjustment of selection bias in those cases where the selection mechanism is directly associated to the variable of interest, this is, the missing responses for the variable of interest are Missing Not At Random (MNAR). These are the most problematic yet most common cases in real data situations, and the methods to mitigate the bias developed in literature can only remove it to a small extent in such situations, although some promising results could be observed (for example, PSA applying Random Forest to estimate the propensities provided considerably better results in MNAR situations).

# 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.

Austin, P. C. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, vol. 27(12), pp. 2037–2049, 2008.

Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, vol. 46(3), pp. 399–424, 2011.

Austin, P. C. and Stuart, E. A. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, vol. 34(28), pp. 3661–3679, 2015.

Barge, S. and Gehlbach, H. Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, vol. 53(2), pp. 182–200, 2012.

Beaumont, J.-F. A new approach to weighting and inference in sample surveys. *Biometrika*, vol. 95(3), pp. 539–553, 2008.

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

Bech, M. and Kristensen, M. B. Differential response rates in postal and web-based surveys in older respondents. *Survey Research Methods*, vol. 3(1), pp. 1–6, 2009.

Bethlehem, J. *Applied survey methods: A statistical perspective*, vol. 558. John Wiley & Sons, 2009a.

Bethlehem, J. *The rise of survey sampling.* Statistics Netherlands, 2009b.

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

Bethlehem, J. Essay: sunday shopping-the case of three surveys. *Survey Research Methods*, vol. 9(3), pp. 221–230, 2015.

Binder, D. A. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pp. 279–292, 1983.

Bosnjak, M. and Tuten, T. L. Prepaid and promised incentives in web surveys: An experiment. *Social science computer review*, vol. 21(2), pp. 208–217, 2003.

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. 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.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. Variable selection for propensity score models. *American journal of epidemiology*, vol. 163(12), pp. 1149–1156, 2006.

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

Buskirk, T. D. and Andrus, C. H. Making mobile browser surveys smarter: results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods*, vol. 26(4), pp. 322–342, 2014.

Buskirk, T. D. and Kolenikov, S. 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*, pp. 1–17, 2015.

Callegaro, M. From mixed-mode to multiple devices: web surveys, smartphone surveys and apps: has the respondent gone ahead of us in answering surveys? *International Journal of Market Research*, vol. 55(2), pp. 317–320, 2013.

Callegaro, M., Manfreda, K. L. and Vehovar, V. *Web survey methodology*. SAGE, 2015.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, vol. 63(3), pp. 615–620, 1976.

Chen, J. K. T., Valliant, R. L. and Elliott, M. 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)*, vol. 68(3), pp. 657–681, 2019.

Chen, S., Yang, S. and Kim, J. K. Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, 2020a.

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, 2020b.

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.

Cobo-Rodríguez, B. *Using auxiliary information in indirect questioning techniques*. PhD thesis, Universidad de Granada, 2018.

Cochran, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pp. 295–313, 1968.

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.

Couper, M. P. and Peterson, G. J. Why do web surveys take longer on smartphones? *Social Science Computer Review*, vol. 35(3), pp. 357–377, 2017.

Coutts, E. and Jann, B. Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, vol. 40(1), pp. 169–193, 2011.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, vol. 13(1), pp. 21–27, 1967.

Devaud, D. and Tillé, Y. Rejoinder on: Deville and särndal's calibration: revisiting a 25-year-old successful optimization problem. *TEST*, vol. 28(4), pp. 1087–1091, 2019.

DEVER, J. A., RAFFERTY, A. and VALLIANT, R. Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, vol. 2(2), pp. 47–60, 2008.

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.

DODOU, D. and DE WINTER, J. C. Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, vol. 36, pp. 487–495, 2014.

ELLIOTT, M. R. and VALLIANT, R. Inference for nonprobability samples. *Statistical Science*, pp. 249–264, 2017.

EVANS, J. R. and MATHUR, A. The value of online surveys: A look back and a look ahead. *Internet Research*, 2018.

FAAS, T. Online or not online?. a comparison of offline and online surveys conducted in the context of the 2002 german federal election. *Bulletin de méthodologie sociologique. Bulletin of sociological methodology*, (82), pp. 42–57, 2004.

FAAS, T. and SCHOEN, H. Putting a questionnaire on the web is not enough-a comparison of online and offline surveys conducted in the context of the german federal election 2002. *Journal of official statistics*, vol. 22(2), p. 177, 2006.

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.

GODAMBE, V. and THOMPSON, M. Estimating equations in the presence of a nuisance parameter. *The Annals of Statistics*, pp. 568–571, 1974.

GÖRITZ, A. S. Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, vol. 1(1), pp. 58–70, 2006.

GÖRITZ, A. S. *Determinants of the starting rate and the completion rate in online panel studies*, chapter 7, pp. 154–170. Wiley Online Library, 2014.

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

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

Hájek, J. Sampling from a finite population. Technical report, 1981.

Hald, A. *A history of probability and statistics and their applications before 1750*. John Wiley and Sons, 2003.

Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Citeseer, 1999.

Heerwegh, D. Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, vol. 21(1), pp. 111–121, 2009.

Hirano, K. and Imbens, G. W. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, vol. 2(3), pp. 259–278, 2001.

Holte, R. C. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, vol. 11(1), pp. 63–90, 1993.

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.

Ilieva, J., Baron, S. and Healey, N. M. Online surveys in marketing research. *International Journal of Market Research*, vol. 44(3), pp. 1–14, 2002.

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.

Khazaal, Y., Van Singer, M., Chatton, A., Achab, S., Zullino, D., Rothen, S., Khan, R., Billieux, J. and Thorens, G. Does self-selection affect samples' representativeness in online surveys? an investigation in online video game research. *Journal of medical Internet research*, vol. 16(7), p. e164, 2014.

Kim, J. K. and Haziza, D. Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, vol. 24(1), pp. 375–394, 2014.

Kim, J. K., Park, S., Chen, Y. and Wu, C. Combining non-probability and probability survey samples through mass imputation. *arXiv preprint arXiv:1812.10694*, 2018.

Kim, J. K. and Wang, Z. Sampling techniques for big data analysis. *International Statistical Review*, vol. 87, pp. S177–S191, 2019.

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

Kursa, M. and Rudnicki, W. Feature selection with the boruta package. *Journal of Statistical Software*, vol. 36(11), pp. 1–13, 2010.

Lee, H., Kim, S., Couper, M. P. and Woo, Y. Experimental comparison of pc web, smartphone web, and telephone surveys in the new technology era. *Social Science Computer Review*, vol. 37(2), pp. 234–247, 2019.

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.

Lehdonvirta, V., Oksanen, A., Räsänen, P. and Blank, G. Social media, web, and panel surveys: using non-probability samples in social and policy research. *Policy & Internet*, 2020.

Little, R. J. Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, pp. 139–157, 1986.

Loomis, D. K. and Paterson, S. A comparison of data collection methods: Mail versus online surveys. *Journal of Leisure Research*, vol. 49(2), pp. 133–149, 2018.

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.

Mavletova, A. and Couper, M. P. Sensitive topics in pc web and mobile web surveys: is there a difference? *Survey Research Methods*, vol. 7(3), pp. 191–205, 2013.

Meng, X.-L. Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, vol. 12(2), pp. 685–726, 2018.

Millar, M. M. and Dillman, D. A. Improving response to web and mixed-mode surveys. *Public opinion quarterly*, vol. 75(2), pp. 249–269, 2011.

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.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M. and Glynn, R. J. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, vol. 174(11), pp. 1213–1222, 2011.

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.

Pasadas-del-Amo, S. Cell phone-only population and election forecasting in spain: The 2012 regional election in andalusia. *REIS*, vol. 162, pp. 55–72, 2012.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J. and Stürmer, T. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety*, vol. 20(6), pp. 551–559, 2011.

Phipps, P. and Toth, D. Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, pp. 772–794, 2012.

Quinlan, J. R. Induction of decision trees. *Machine learning*, vol. 1(1), pp. 81–106, 1986.

Quinlan, J. R. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Reynolds, S., Sharp, A. and Anderson, K. Online surveys: Response timeliness and issues of design. In *ANZMAC*. 2009.

Rivers, D. Sample matching: Representative sampling from internet panels. *Polimetrix White Paper Series*, 2006.

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

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, vol. 70(1), pp. 41–55, 1983.

Rosenbaum, P. R. and Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, vol. 79(387), pp. 516–524, 1984.

Royall, R. M. On finite population sampling theory under certain linear regression models. *Biometrika*, vol. 57(2), pp. 377–387, 1970.

Rubin, D. B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, vol. 91(434), pp. 473–489, 1996.

Rueda, M., Ferri-García, R. and Castro, L. The r package nonprobest for estimation in non-probability surveys. *The R Journal*, vol. 12(1), pp. 406–418, 2020.

Särndal, C.-E. and Lundström, S. *Estimation in surveys with nonresponse*. John Wiley & Sons, 2005.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H. and Brookhart, M. A. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, vol. 20(4), p. 512, 2009.

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.

Smironva, E., Kiatkawsin, K., Lee, S. K., Kim, J. and Lee, C.-H. Self-selection and non-response biases in customers' hotel ratings–a comparison of online and offline ratings. *Current Issues in Tourism*, vol. 23(10), pp. 1191–1204, 2020.

Smith, L. H. Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports*, pp. 1–11, 2020.

Smith, T. M. Post-stratification. *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 40(3), pp. 315–323, 1991.

SPANISH NATIONAL INSTITUTE OF STATISTICS. Survey on equipment and use of information and communication technologies in households. results. 2020. https://ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176741&menu=resultados&idp=1254735576692. Accessed 10 March 2021.

TERHANIAN, G. and BREMER, J. A smarter way to select respondents for surveys? *International Journal of Market Research*, vol. 54(6), pp. 751–780, 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.

TILLÉ, Y. *Sampling and estimation from finite populations*. John Wiley & Sons, 2020.

TOURANGEAU, R., COUPER, M. P. and CONRAD, F. Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public opinion quarterly*, vol. 68(3), pp. 368–393, 2004.

TOURANGEAU, R., COUPER, M. P. and CONRAD, F. G. üp means good": the effect of screen position on evaluative ratings in web surveys. *Public Opinion Quarterly*, vol. 77(S1), pp. 69–88, 2013.

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.

VAVRECK, L. and RIVERS, D. The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, vol. 18(4), pp. 355–366, 2008.

VEHOVAR, V. and MANFREDA, K. L. *Overview: online surveys*, chapter 10, pp. 177–194. Sage: London, UK, 2008.

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.

Wells, T., Bailey, J. T. and Link, M. W. Comparison of smartphone and online computer survey administration. *Social Science Computer Review*, vol. 32(2), pp. 238–255, 2014.

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.

Yang, S. and Kim, J. K. Integration of survey data and big observational data for finite population inference using mass imputation. *arXiv preprint arXiv:1807.02817*, 2018.

# Part II

# Appendices

# Appendix A1

## Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys

Ferri-García, Ramón; Rueda, María del Mar (2018)
Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys
*SORT: statistics and operations research transactions*, vol. 42, number 2, pp. 156 - 182
DOI: 10.2436/20.8080.02.73

| STATISTICS & PROBABILITY | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2018 | 1.125 | 58/123 | Q3 |

### Abstract

One of the main sources of inaccuracy in modern survey techniques, such as online and smartphone surveys, is the absence of an adequate sampling frame that could provide a probabilistic sampling. This kind of data collection leads to the presence of high amounts of bias in final estimates of the survey, specially if the estimated variables (also known as target variables) have some influence on the decision of the respondent to participate in the survey. Various correction techniques, such as calibration and Propensity Score Adjustment or PSA, can be applied to remove the bias. This study attempts to analyze the efficiency of correction techniques in multiple situations, applying a combination of PSA and calibration on both types of variables (correlated and not correlated with the miss of data) and testing the use of a reference survey to get the population totals for calibration variables. The study was performed using a simulation of a fictitious population of potential voters and a real volunteer survey aimed to a population for which a complete census was available. Results showed that PSA combined with calibration results in a bias removal considerably larger when compared with calibration with no prior adjustment. Results also showed that using population totals from the estimates of a reference survey instead of the available population data does not make a difference in estimates accuracy, although it can contribute to slightly increment the variance of the estimator.

## 1 Introduction

Traditional surveys are experiencing, along with society, a number of changes which affect their validity and applicability. Several reasons can be cited (e. g., see Couper (2017) and Schonlau et al. (2009)) on the decline of participation and completion rates in surveys conducted using traditional modes of contact, such as telephone or face-to-face surveys. A review performed by Díaz de Rada (2012) stated that response rates in traditional surveys have been dropping for two decades. The increasing difficulty of contacting households members in face-to-face surveys results in increased costs per interview and therefore non-sampling errors are problematic to deal with in this context; regarding telephone surveys, the rise of mobile phones makes it more difficult for government agencies to keep an adequate sampling frame, in terms of coverage, of landline phones (Pasadas-del-Amo, 2018).

At the same time, the arrival of the Internet and mobile phone lines has led to the usage of new survey administration methods, with online surveys and smartphone surveys being the most popular and promising ones to deal with the mentioned issues in order to contact respondents. Online surveys can be defined, given how they are conducted nowadays as described by Mei and Brown (2017), as surveys filled from computers that respondents can access anytime. Questionnaires might have a conventional structure adapted to the online context (e. g. SurveyMonkey) and might also be provided using online social networks. Smartphone surveys differ in the filling mode: any survey completed using a mobile device or a tablet can be considered a smartphone survey. Sometimes, the questionnaire might be hosted in an URL, thus it could be considered a browser survey and therefore an online survey. This states a clear divide in the smartphone surveys between those app-based questionnaires or related and those filled using a browser available in the device itself, as the latter do not properly seize the advantages of a mobile device.

The change from the traditional survey to the internet survey has brought important changes and new challenges have arisen (Díaz de Rada and Domínguez (2015); Díaz de Rada and Domínguez (2016)). These new methods offer substantial advantages against traditional survey techniques, specially in terms of monetary and time costs as they usually do not require any effort by any interviewer and the information collection becomes instantaneous. In addition, online surveys are considered to be more advantageous for information collection; despite the advantages of smartphones such as the audiovisual options and the possibility to retrieve data on certain variables without the need of any extra question in the survey, web surveys take less time to be completed by interviewers, as proved by Couper and Peterson (2017).

Along with the described advantages, some serious concerns often arise when using these new survey methods. As noted in Elliott and Valliant (2017), Internet surveys (even when an structured voluntary panel is used) suffer mostly from selection bias, specially from the bias induced by the Internet availability and penetration in the general population. This issue will be broadly discussed later. Internet surveys are also affected by nonresponse bias; a meta-analysis conducted by Manfreda et al. (2008) estimated that online surveys are associated with a decrease in response rates between 6% and 15% in comparison to other survey modes. In addition, the use of incentives as a method to improve cooperation have been proved as less efficient in online surveys (Díaz de Rada, 2012). Other important sources of non sampling errors in online and smartphone surveys are measurement errors; although the social desirability effect is less prone to appear in online surveys (Heerwegh, 2009), they still suffer from other effects such as technical issues (p. e. poor Internet connection may lead to a lack of completion of a survey) or lack of veracity in the responses given, which in the online case has a variety of causes.

Nonresponse bias, as well as measurement errors, have been widely studied in the literature as they have been common issues in traditional survey methods since their initial development. However, selection bias presents some particular characteristics in the new survey methods which require other strategies in order to tackle it. In all cases, online and smartphone surveys are often applied under inadequate sampling conditions; they are generally filled by self-selected respondents which conform a non-probabilistic sampling. Even if an acceptable random sampling is eventually performed, it may be particularly troublesome to establish a reliable sampling frame required to meet the probabilistic sampling assumptions (Couper (2000); Couper and Peterson (2017)). On the other hand, the coverage of such surveys is also limited by the population access to the Internet. Although no interview mode is exempt from suffering coverage bias, it happens to be much more important in Internet surveys (Couper (2007), according to Schonlau et al. (2009)), as Internet access is often associated to sociodemographic variables which could be eventually related to the outcome variables of a certain study. To mention some examples, data from Pew Research Center (2017) reveals that in 2016 while 99%

of U. S. adults between 18 and 29 years old could be considered Internet users, only a 64% of those above 65 years of age fell into the same group. In the case of Spain, the generation gap is wider according to the National Institute of Statistics (2017a); while the penetration rate is above 90% for all age groups below 54 years of age, in citizens between 65 and 74 years old Internet penetration rate is 43.7%.

It is obvious that such a problem can be responsible for a large increase in the bias of the final results. Therefore, developing methods to deal with the lack of representativity has become a priority. To date, the more relevant methods are considered to be calibration techniques and Propensity Score Adjustment (PSA). Calibration weighting using auxiliary information (Deville and Särndal, 1992) has been established as the main technique to deal with problematic sampling frames, but its efficacy can decrease when the self-selection procedure is tied (directly or not) to the target variables (Bethlehem, 2010). Calibration for coverage issues has also been studied using the superpopulation model approach through general regression (GREG) weights (Dever et al., 2008); despite it successfully address both nonresponse and noncoverage in online surveys, it requires an structured sampling design, something that does not apply to volunteer surveys. When calibration is ineffective, PSA can be a proper substitute if it is feasible to use a probabilistic sample on the same target population, on which a subset of variables measured on the non-probabilistic sample have been measured on the probabilistic sample as well. Research findings have shown that PSA successfully removes bias in some situations, but at the cost of increasing the variance of the estimates (Lee (2006); Lee and Valliant (2009)). The efficacy of bias removal by PSA is strongly dependent on using covariates related to the actual propensity to participate and the target variables (Schonlau and Couper, 2017), and its sole application without any further adjustment can lead to biased estimates (Valliant and Dever, 2011). The aim of this study was to examine the behavior of the estimators when both techniques, PSA and calibration, are applied, in comparison to the situations where only calibration is performed or where no weighting technique is applied at all. Given that, for most situations, auxiliary information can be troublesome to find, calibration is tested using known population totals and using population estimates coming from the reference (probabilistic) sample that it is supposed to be available. Under the initial hypothesis of the study, the combined weighting of PSA in a first step and calibration in a second one would outperform the estimates obtained with calibration weighting only in terms of bias reduction, although the estimators will have a higher variance as the reference sample size gets smaller in comparison to the convenience (non-probabilistic) sample size.

# 2 Methodology

## 2.1 Calibration weighting

Surveys often have a coverage error associated to them, in the sense of being made using a sampling frame that does not cover the entire population to which survey results are to be extrapolated. This coverage error, which can be the result of several irregularities, can be controlled by the use of reweighting or calibration techniques. Calibration was defined by Särndal (2007) as the combination of three items:

> a) a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s), b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units, c) an objective to obtain nearly design unbiased estimates as long as non-response and other non-sampling errors are absent.

Calibration theory can be explained as follows (Deville and Särndal, 1992): let $y$ be the interest variable in the survey estimation and $s$ the sample collected in the survey, with each element $k$ in the sample having an associated probability of selection, $\pi_k = 1/d_k$. Without any auxiliary information, the population total of $y$, $Y$, is estimated in a non-biased way with the Horvitz-Thompson estimator:

$$\hat{Y}_{HT} = \sum_{k \in a} d_k y_k \tag{1}$$

Let $\mathbf{x}$ be an auxiliary vector associated to $y$, 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 obtaining of a new weights vector $w_k$ for $k \in s$ which modifies as little as possible the original sample weights, $d_k$, which have the desirable property of producing unbiased estimations, respecting at the same time the calibration equations:

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

Given a distance $G(w_k, d_k)$, the calibration process consists on finding the solution to the minimization problem

$$\min_{w_k} E\{\sum_{k \in s} G(w_k, d_k)\} \tag{3}$$

while respecting the calibration equation (2). Several distances were defined in Deville and Särndal (1992), being the linear distance one of the most commonly used (Rueda et al. (2010), Martínez et al. (2010)). This distance is calculated by:

$$\sum_{k \in S} \frac{(w_k - d_k)^2}{q_k d_k} \tag{4}$$

$q_k$ are positive weights that are usually assumed as uniform (i. e. $1/q_k = 1$), although unequal weights $1/q_k$ are sometimes used. The problem now concerns finding the minimum of (4) subject to ( 2), leading to the calibrated weight:

$$w_k = d_k(1 + q_k \mathbf{x}_k' \lambda) \tag{5}$$

where the vector of multipliers, $\lambda$, is calculated as:

$$\lambda = T_s^{-1}(\mathbf{X} - \sum_s \mathbf{x}_k d_k) \tag{6}$$

$T_s$, whose inverse is assumed to exist, is the equivalent of:

$$T_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k' \tag{7}$$

The resulting estimator of $Y$ is the general regression estimator (Cassel et al., 1976)

$$Y = \sum_s w_k y_k = \sum_s y_k d_k + (\mathbf{X} - \sum_s \mathbf{x}_k d_k)' \hat{B}_s \tag{8}$$

where $\hat{B}_s$ is

$$\hat{B}_s = T_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k \tag{9}$$

In general, the resulting estimator for $Y$ is biased, but it is assumed to be asymptotically unbiased as the new weights $w_k$ would approach to the sampling weights $d_k$.

## 2.2 Propensity Score Adjustment (PSA)

The Propensity Score Adjustment method was originally developed by Rosenbaum and Rubin (1983) which sought to reduce the bias due to treatment and control assignment in non-randomized studies. The main idea of the adjustment is to balance the differences between groups in non-randomized designs with the computation of a score whose distribution is the same for all groups. The proposed score for a given unit is equivalent to its probability of being in the treatment group, which can be modeled using a regression model. Although the implications of this approach in survey nonresponse were considered shortly after (Rubin (1986), according to Little and Rubin (2002)), it was not proposed for online surveys until Harris Interactive took it into account in their Internet research (Taylor (2000); Taylor (2001)). To a lesser extent, these first attempts added one element to the requirements for performing PSA: a reference survey. The concept of reference survey was extended in further studies (see Lee (2006)).

When treating an online survey, it is expected that the sampling was conducted in a non-probabilistic manner or even not conducted at all, with the survey being filled by volunteer respondents. It is feasible to consider that the decision to take part on the survey depends on a probability which, depending on the respondent characteristics, might be higher or lower. In this case, a reference survey can be very helpful to determine the mentioned probability. A reference survey is conducted on the same target population than the online 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 probabilistic survey is performed on it.

Once data is collected from both surveys, the propensity for an individual to take part on the volunteer (non-probabilistic) survey is obtained by binning the data together and training a logistic regression model on the dichotomous variable, $y$, which measures whether the respondent took part in the volunteer survey or in the reference survey. The model uses covariates, $\mathbf{x}$, that have been measured in both surveys, thus the formula to compute the propensity of taking part in the volunteer survey, $\pi$, can be displayed as

$$\pi(x) = \frac{1}{e^{-(\gamma^T \mathbf{x}_k)} + 1} \tag{10}$$

for some vector $\gamma$, as a function of the model covariates.

We denote by $s_R$ the reference sample and by $s_V$ the volunteer sample. Following the approach described in Lee and Valliant (2009) which will be used in this study, propensity scores are divided in $g$ classes, with $g$ = 5 as the conventional choice following Cochran (1968), where all units may have the same propensity score or at least be in a very narrow range. For each class, an adjustment factor is calculated as stated in (11):

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

where $s_{R_g}$ is the set of individuals in the reference sample that are in the $g$th class of propensity scores, and $d_{Rk}$ is the original design weight of the $k$ individual in the reference sample, $s_{V_g}$ is the set of individuals in the volunteer sample that are in the $g$th class of propensity scores, and $d_{Vk}$ is the original design weight of the k individual in the volunteer sample. Finally, the adjusted weights $d^*$ 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 $g$th propensity class in the volunteer sample) is computed as indicated in (12). These weights are equivalent to the weights used for the Horvitz-Thompson (H-T) estimator.

$$d_k^* = f_g d_{Vk} = \frac{\sum_{k \in s_{R_g}} d_{Rk} / \sum_{k \in s_R} d_{Rk}}{\sum_{k \in s_{V_g}} d_{Vk} / \sum_{k \in s_V} d_{Vk}} d_{Vk} \tag{12}$$

Alternatively, the approach proposed in Schonlau and Couper (2017) can be used to obtain weights for a Hajek-type estimator using propensity scores. This approach has the particularity of adjusting to the population of the probabilistic sample, rather than the combined population of the two samples. Weights are defined as the inverse propensity scores, as indicated in (13)

$$w_i = \frac{1 - \hat{\pi}(\mathbf{x}_i)}{\hat{\pi}(\mathbf{x}_i)} \tag{13}$$

where $\hat{\pi}(\mathbf{x}_k i$ is the estimated response propensity for the individual $i$ of the volunteer sample as predicted by logistic regression with covariates $\mathbf{x}$.

# 3 Simulation study

## 3.1 Data description

To explore the effectivity of PSA with further calibration compared to calibration alone, a fictitious population was simulated in order to analyze and establish conclusions for the behaviour of these techniques when applied in real situations. The simulation was based on the study presented in Bethlehem (2010), introducing several changes to extend the spectrum of possible cases in which adjustment methods can be used. In the proposed simulation study, a survey would be conducted to examine population's voting intention. The population had a fixed size of N=50000, and six variables were included in the study: age, nationality (native/non-native), gender, education (primary/secondary/tertiary), access to the Internet (yes/no), and party to which they intended to vote, with four possible options: Party 1, Party 2, Party 3 and Abstention. The distribution of the variables and the relationships between them were fixed as follows:

- Age followed a Beta distribution with $\alpha = 2$ and $\beta = 3$ to make it similar to the Spanish population pyramid (National Institute of Statistics, 2017b), and it ranged from 18 to 100 years old.

- Probability of being non-native depended on the age, which was divided in three classes ($< 35$, 35-65, and $>$65 years old) and individuals on each had a probability of 0.15, 0.1 and 0.025 respectively of being non-native. This probability is similar to the nationality distribution by ages in Spain (National Institute of Statistics, 2016).

- Probability of being a woman was fixed in 0.5 for everyone, except for individuals above 75 years old, whose probability of being a woman was 0.65, as women in Spain tend to have a greater representation in older ages (National Institute of Statistics, 2017b).

- Probabilities of having a specific education level were fixed to resemble as much as possible to the Spanish adult population (National Institute of Statistics, 2017c). These probabilities can be consulted in Table 1.

Table 1: Probability of each education level as the highest achieved by the fictitious individual, by age groups

| Education level / Age group | $<$ 35 years old | 35-65 years old | $>$ 65 years old |
|---|---|---|---|
| Primary education | 0.35 | 0.45 | 0.8 |
| Secondary education | 0.2 | 0.25 | 0.1 |
| Tertiary education | 0.45 | 0.3 | 0.1 |

- Access to the Internet was made dependent of two variables: age and nationality. This time the probabilities assignment was not based in real data, in order to capture more patterns in the experiment. Probability of access by age groups and nationalities can be consulted in Table 2.

Table 2: Probability of access to the Internet by a given individual, by age groups and nationality

| Nationality / Age group | $<$ 35 years old | 35-65 years old | $>$ 65 years old |
|---|---|---|---|
| Native | 0.9 | 0.7 | 0.5 |
| Non-native | 0.2 | 0.1 | 0.0 |

- Probability of voting to each party depended on the party itself. The following relationships were established to make sure all kinds of missing data mechanisms would be represented in the analysis:
  - Voting to Party 1 depended on the gender of the individual; women had a probability of 0.2 to vote for this party while men had a 0.0 probability. Gender is not related to Internet access (which is the responsible for non-response) thus the missing data mechanism could be considered as MCAR (Missing Completely At Random).
  - Voting to Party 2 depended on the age of the individual; voting probability was 0.0 for people younger than 35 years old, 0.4 for people between 35 and 65 years old, and 0.6 for people older than 65 years old. Given that age, which is an auxiliar variable, is related to Internet access, the missing data mechanism was MAR (Missing At Random).
  - Voting to Party 3 depended on the access to the Internet and the age; people with no access to the Internet had a 0.1 probability, no matter what old they were, while people with access had a 0.6, 0.4 and 0.2 probability for each respective age group. In this case, the target variable is directly related to the non-response mechanism, configuring a NMAR (Not Missing At Random) situation.

### 3.2 Results

To estimate the bias for every possible situation, several configurations of sample sizes for the volunteer sample were considered, letting it vary between 500 and 10,000 individuals. On the other hand, the reference sample size was fixed in 500 individuals for all the experiments. For each volunteer sample size, 1,000 simulations were computed for the results on estimated percent of vote for each of the parties, using the following methods:

- Non-adjusted (unweighted) estimates from the volunteer sample.

- Calibrating the volunteer sample with population totals or estimated population totals (from the reference sample).

- Reweighting with PSA and applying those weights directly to the sample with no further adjustments.

- Reweighting with PSA and calibrating those weights with population totals or estimated population totals (from the reference sample).

Propensity scores were calculated using both approaches presented in Section 2.2 (with $g = 5$ for stratification in the Horvitz-Thompson estimator weights computation). Variables used for PSA and calibration were assigned in four different situations with the following combinations:

- Situation 1: age and education as PSA covariates, gender as calibration variable.

- Situation 2: age and education as PSA covariates, nationality as calibration variable.

- Situation 3: age and nationality as PSA covariates, education as calibration variable.

- Situation 4: age and nationality as PSA covariates, gender as calibration variable.

For each method and situation, the bias, as a result of the difference between real vote % and estimated vote %, was calculated, as well as the standard deviation of the voting estimation for the 1000 simulations. Figures 1 and 2 summarize results for Situation 1.

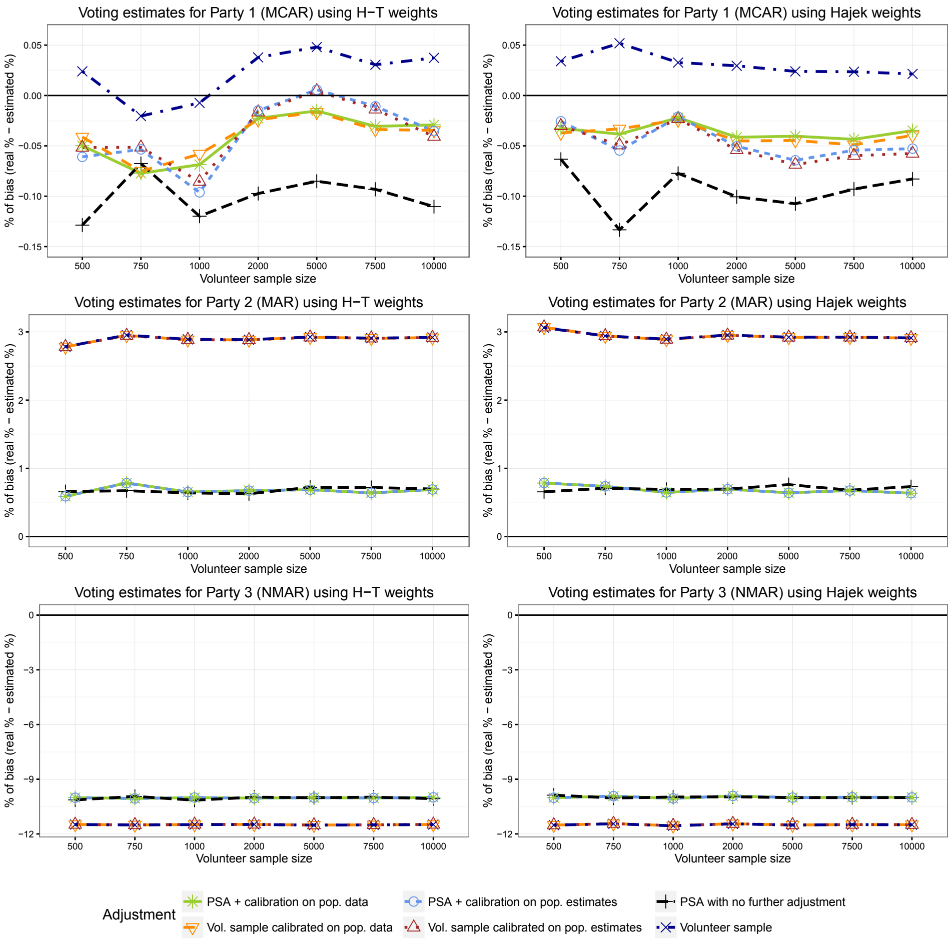


Figure 1: Bias of each method in voting intention estimations by party in Situation 1

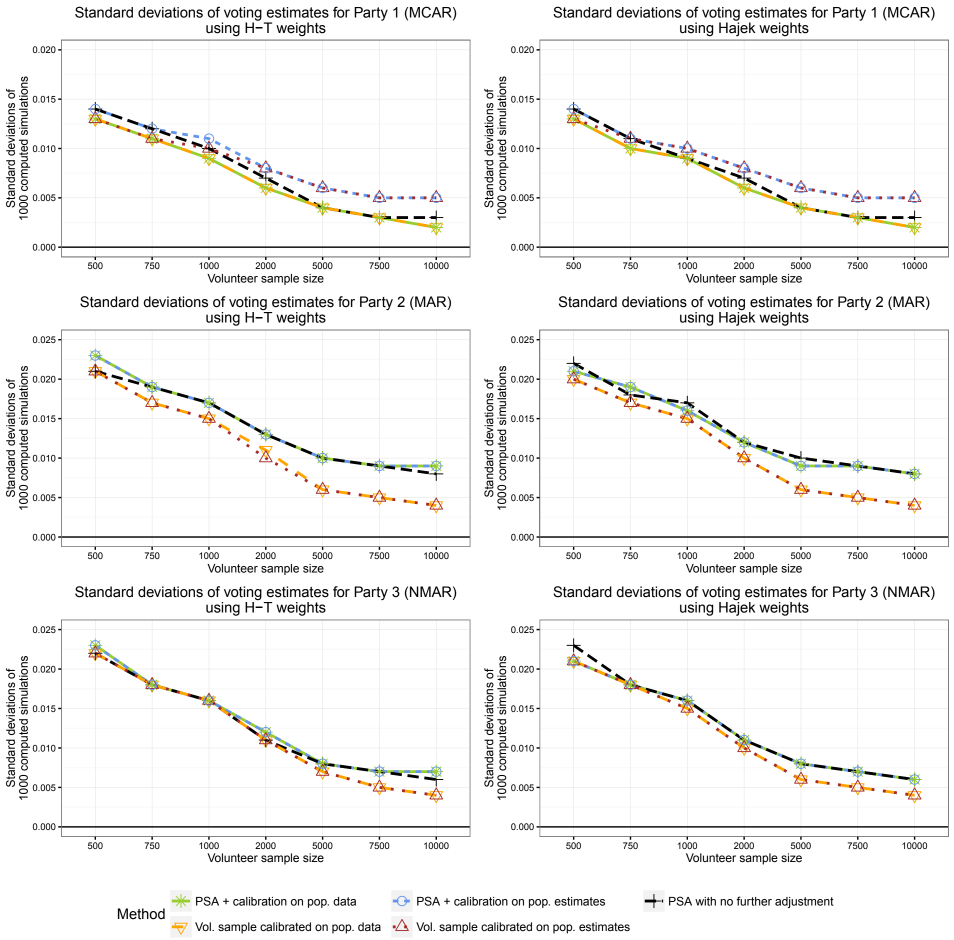


Figure 2: Standard deviation of voting intention estimations by party provided by each method in Situation 1

Results showed that the difference in bias when the missing data mechanism was completely random is negligible; however, when data was MAR or NMAR, using PSA (regardless of doing calibration afterwards or not) resulted in a reduction in the amount of bias, although this reduction was much higher when data is MAR. It is worth mentioning that these statements could be extended to all the studied sample size situations.

In terms of standard deviations, which give a measure of the variance of the estimator for each method, it can be observed that methods involving PSA resulted in an increase in variance in comparison to methods involving calibration only. However, it is important to point out that the use of estimates of population totals did not increase variance of the survey estimates in MAR and NMAR cases. For the MCAR case, methods involving estimates of population totals resulted overall in greater variance of the estimators.

It is worth to mention that using Horvitz-Thompson weights or Hajek weights after the computation of the PSA scores made almost no difference in final results in terms of bias reduction or estimators' variance. The very slight differences that could be observed between results may be attributed to the randomness of the experiment rather to an actual effect of the type of weighting.

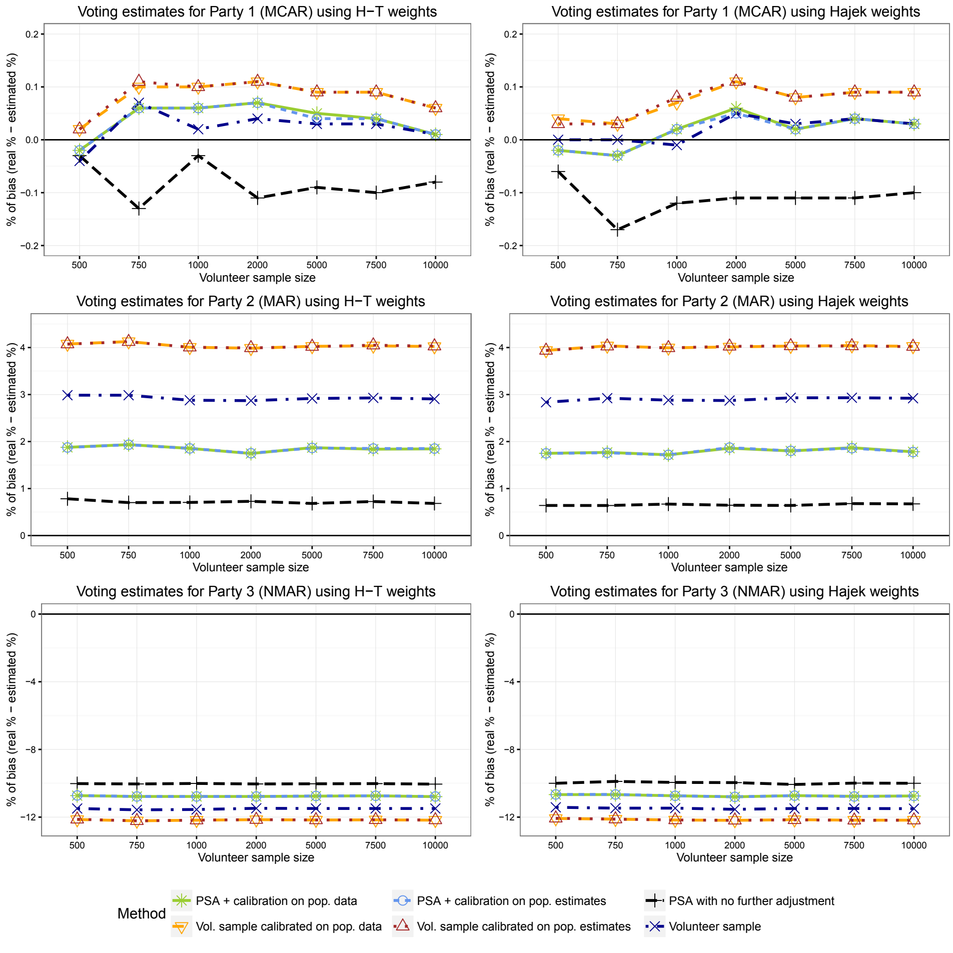


Figure 3: Bias of each method in voting intention estimations by party in Situation 2

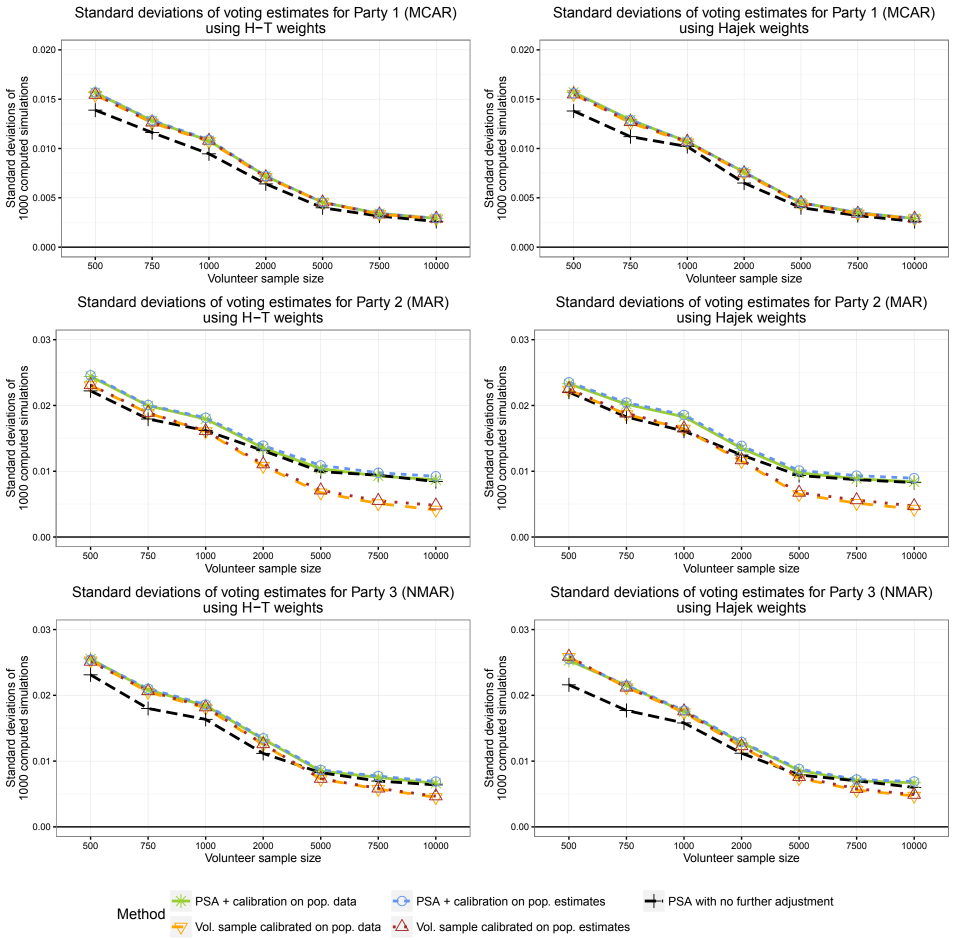


Figure 4: Standard deviation of voting intention estimations by party provided by each method in Situation 2

Figures 3 and 4 summarize results for Situation 2. Bias reduction kept its consistence between weighting methods (Horvitz-Thompson and Hajek), but some differences were found in reference to Situation 1. The only difference between them was the calibration variable used (nationality instead of gender), but it turned out to be a critical choice. As it can be seen in Figure 3, the application of calibration in Situation 2 resulted in an increase of bias on the estimates, while PSA with no further adjustment produced the same bias reduction than the registered in Situation 1. Estimates involving calibration also had a higher variance, as it can be observed in Figure 4.

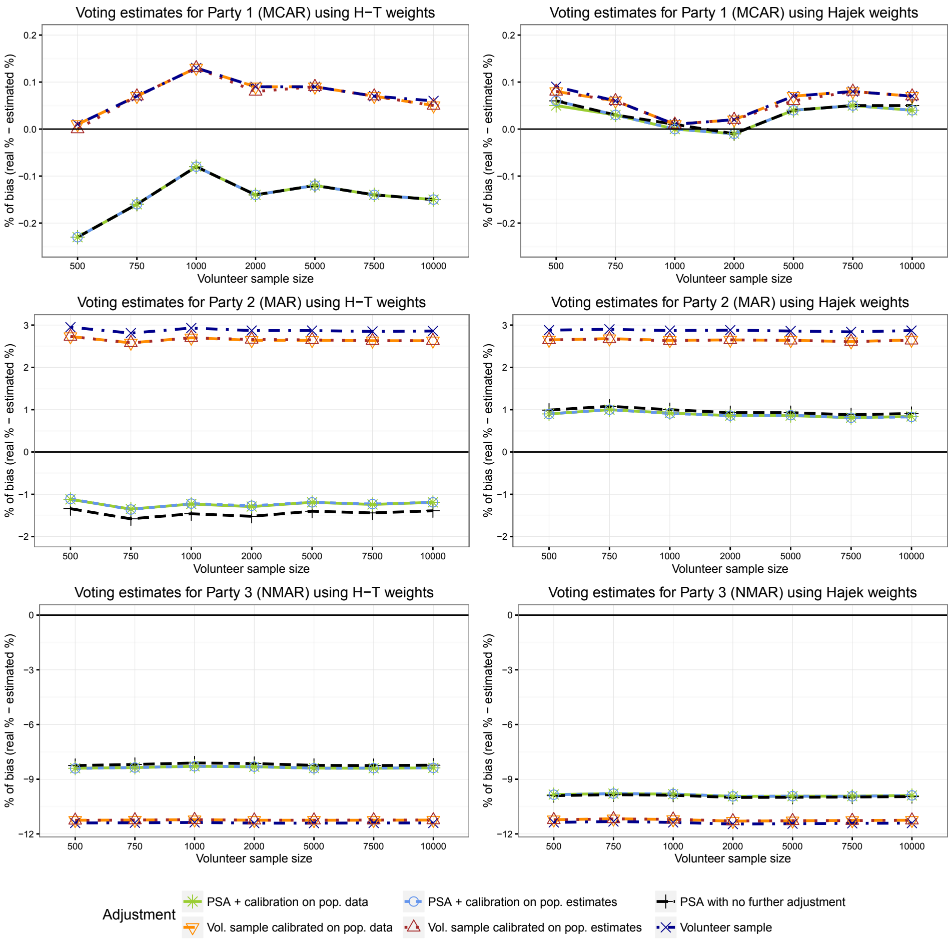


Figure 5: Bias of each method in voting intention estimations by party in Situation 3

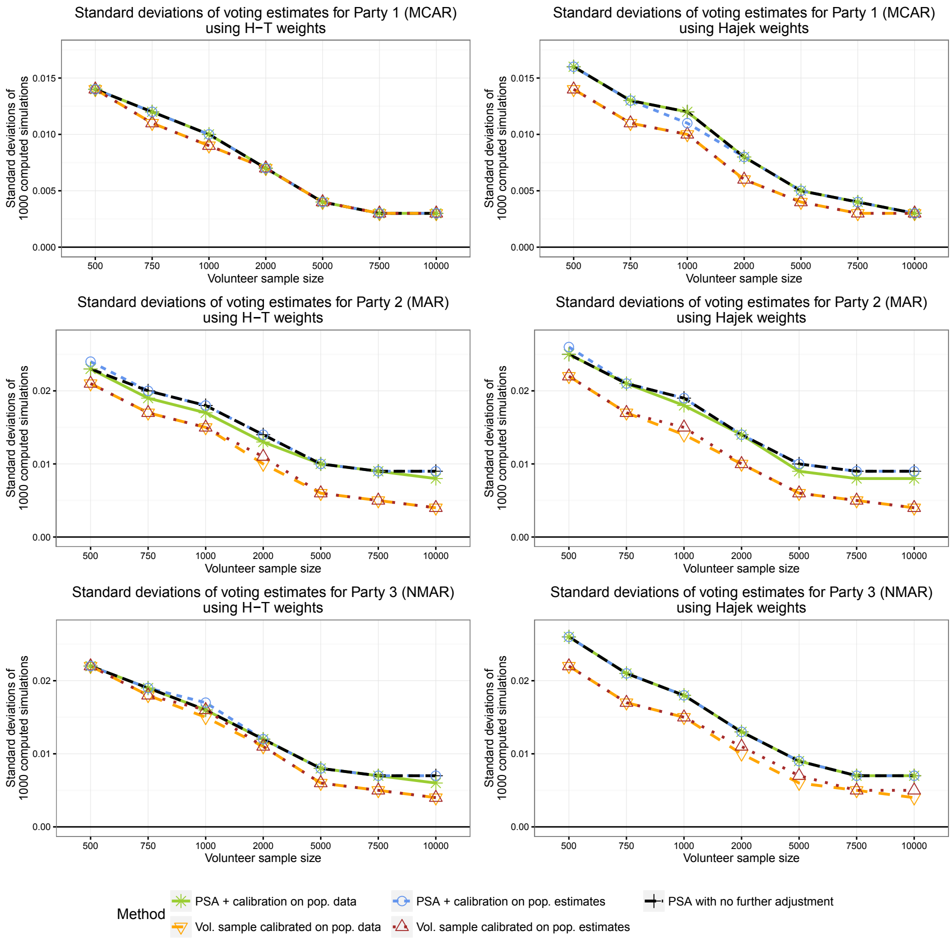


Figure 6: Standard deviation of voting intention estimations by party provided by each method in Situation 3

Figures 5 and 6 summarize results for Situation 3. In this case, there is a difference in bias reduction motivated by the weighting method used. It is noticeable that Hajek-type estimates are less biased than Horvitz-Thompson-type estimates in the MCAR and MAR cases. It is also worth to mention that PSA with calibration removed more bias than PSA with no adjustment in the MAR case using Horvitz-Thompson weights. On the contrary, in the NMAR case Horvitz-Thompson-type estimates are less biased than Hajek-type estimates. Finally, in terms of variance, it can be observed in Figure 6 that Hajek-type estimators have a greater variance than Horvitz-Thompson-type estimators, specially when the volunteer sample size is relatively small.

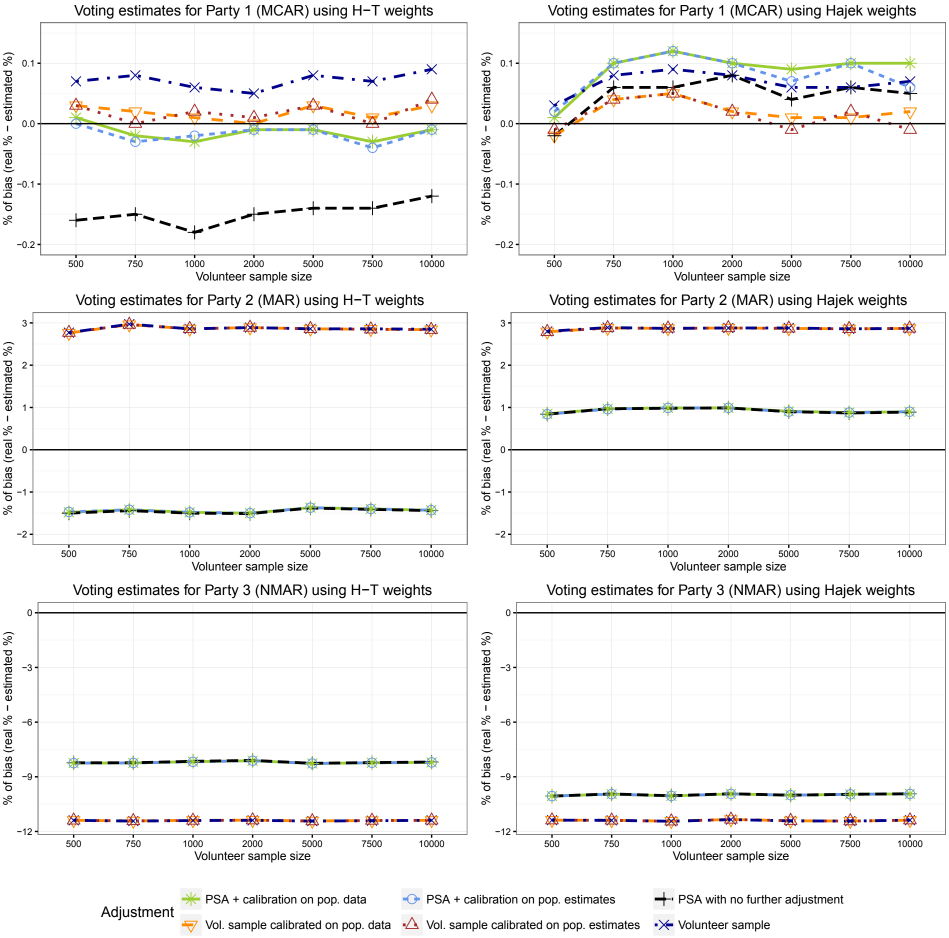
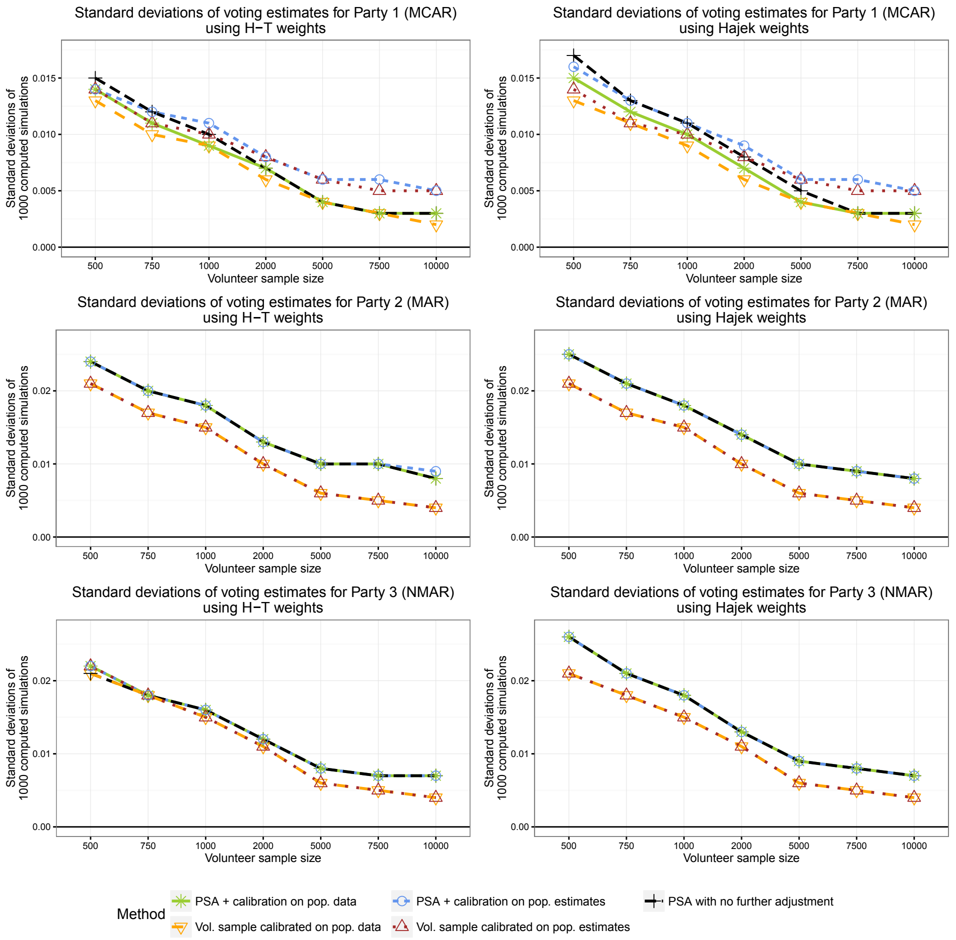


Figure 7: Bias of each method in voting intention estimations by party in Situation 4


Standard deviations of voting estimates for Party 1 (MCAR) using H−T weights

Standard deviations of voting estimates for Party 1 (MCAR) using Hajek weights

Standard deviations of voting estimates for Party 2 (MAR) using H−T weights

Standard deviations of voting estimates for Party 2 (MAR) using Hajek weights

Standard deviations of voting estimates for Party 3 (NMAR) using H−T weights

Standard deviations of voting estimates for Party 3 (NMAR) using Hajek weights

Standard deviations of
1000 computed simulations
0.000
0.005
0.010
0.015
0.00
0.01
0.02
500
750
1000
2000
5000
7500
10000
Volunteer sample size
Method
PSA + calibration on pop. data
PSA + calibration on pop. estimates
PSA with no further adjustment
Vol. sample calibrated on pop. data
Vol. sample calibrated on pop. estimates

Figure 8: Standard deviation of voting intention estimations by party provided by each method in Situation 4

Figure 7 and 8 summarize results for Situation 4. The differences between weighting methods disappear in the MCAR case but remain in the MAR and NMAR cases. In addition, no reduction in bias could be attributed to the calibration of the sample, in contrast with Situation 3, where calibration resulted in less biased estimates in all cases. Regarding standard deviations, the most remarkable result in this situation is the increase in variance that calibration produces in this situation.

# 4 Application study

## 4.1 Data description

The probabilistic sample data for the application case was obtained through a survey conducted amongst the students of the University of Granada, Spain (UGR) in 2015, with a sample size of n = 856 participants. Respondents were recruited through face-to-face interviews following a cluster sampling scheme in three phases, in which Faculties were the primary units, degrees were the secondary units, and academic years were the tertiary units. A total of 34 clusters were randomly drawn from the population following this design. Sampling error was estimated at ± 3.3% given the sample size and a confidence level of 95%. Respondents had to fill questionnaires which included several screening instruments for certain kinds of abuse or dependency, including the Cannabis Abuse Screening Test (CAST) and the Severity of Dependence Scale (SDS), which were both validated for the sample. The questionnaire also measured the age and gender of the participants.

The non-probabilistic sample used in this application case came from a survey performed in 2017 by students of the UGR amongst their peers, with a sample size of n = 341 participants. Respondents were recruited following a snowball sampling scheme in online social networks, and filled the questionnaire using an online platform (Google Drive$^{TM}$). The questionnaire included the CAST and the SDS, as well as questions regarding the age and gender of the respondents. The sampling method implied an Internet connection from the respondent and a certain willingness to volunteer in the survey, meaning selection bias came from the same sources than in most of the online non-probabilistic surveys.

The aim of the application was to estimate the SDS mean score for the non-probabilistic sample using the aforementioned correction techniques. Given that SDS scores were provided only for cannabis users in both samples, the original sample sizes dropped out to n = 115 participants for the probabilistic survey and n = 87 for the non-probabilistic survey.

## 4.2 Results

The probabilistic sample was used to estimate the total number of cannabis users in the UGR by age groups and gender. These estimates were used as population totals in calibration, in reference to the simulation study results which shown no difference, in terms of bias reduction, between using actual population totals or their estimates. However, this meant that only age and gender could be used as calibration variables. On the other hand, PSA could be performed using age, gender and CAST scores. Differences in data for the three variables between both samples can be consulted in Table 3.

Table 3: Means and relative frequencies of each sociodemographic level in the studied samples, and p-values for tests of independence or difference in means performed on each variable

| Variable | Level | Probab. sample | Non-probab. sample | p-value |
|---|---|---|---|---|
| Gender | | | | |
| | Male | 51.30 % | 74.71 % | 0.001[a] |
| | Female | 48.70 % | 25.29 % | |
| Age | | | | |
| | 18 or younger | 13.91 % | 16.09 % | 0.425[b] |
| | 19 | 13.91 % | 18.39 % | |
| | 20 | 9.57 % | 12.64 % | |
| | 21 | 20.87 % | 10.34 % | |
| | 22 | 12.17 % | 14.94 % | |
| | 23 or older | 29.57 % | 27.59 % | |
| CAST score | | | | |
| | Mean score | 4.435 | 5.322 | 0.167[c] |

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

[b] Pearson's Chi-squared test

[c] Welch two sample t-test

The difference in gender proportions between both samples is statistically significant (p = 0.001205), hence it can be assumed that the frames from which samples were withdrawn had different gender proportions. However, this assumption cannot be made for any of the other variables; no practical or statistical significance was found in the difference between samples. These results are an evidence of the lack of discriminant power of PSA potential covariates, thus the propensity of belonging to any of both samples might be much less explanatory.

Estimates of the SDS mean score were computed for each possible combination of techniques (no adjustment, calibration, PSA, and PSA with calibration), auxiliary variables and PSA covariates. Hajek estimator weights were computed in PSA considering the small number of covariates to be used in several combinations, which might not allow to properly allocate the propensity in groups. In each case, Jackknife leave-one-out was performed in order to compute an unbiased estimate of the standard error committed by each method. Results are presented in Table 4, along with the relative difference (in percentage) between each estimate and the mean SDS score provided by the probabilistic sample.

Table 4: Estimated SDS mean, standard error and difference with the mean estimated with the probabilistic sample by method, calibration auxiliary variables, and PSA covariates

| Method | Calibration aux. variables | PSA covariates | Mean SDS score<br>Estimated | <br>Std. Err. | <br>Dif. |
|---|---|---|---|---|---|
| **Reference sample** | | | | | |
| Unweighted | | | 6.261 | 0.199 | |
| **Volunteer sample** | | | | | |
| Unweighted | | | 7.264 | 0.272 | 16.03 % |
| Calibration | | | | | |
| | Sex | | 7.004 | 0.253 | 11.87 % |
| | Age | | 7.206 | 0.276 | 15.09 % |
| | Sex and age | | 6.904 | 0.253 | 10.26 % |
| PSA (Hajek) | | | | | |
| | | Sex | 6.939 | 0.252 | 10.84 % |
| | | Age | 7.349 | 0.286 | 17.39 % |
| | | CAST | 6.986 | 0.246 | 11.58 % |
| | | Sex, age | 6.997 | 0.266 | 11.76 % |
| | | Sex, CAST | 6.790 | 0.238 | 8.46 % |
| | | Age, CAST | 6.971 | 0.251 | 11.34 % |
| | | Sex, age, CAST | 6.742 | 0.247 | 7.68 % |
| PSA (Hajek) + calibration | | | | | |
| | Sex | Sex | 7.311 | 0.278 | 16.77 % |
| | | Age | 7.007 | 0.253 | 11.92 % |
| | | CAST | 7.028 | 0.253 | 12.25 % |
| | | Sex, age | 7.323 | 0.280 | 16.97 % |
| | | Sex, CAST | 7.311 | 0.278 | 16.78 % |
| | | Age, CAST | 7.052 | 0.254 | 12.63 % |
| | | Sex, age, CAST | 7.331 | 0.281 | 17.10 % |
| | Age | Sex | 7.182 | 0.283 | 14.70 % |
| | | Age | 7.126 | 0.264 | 13.82 % |
| | | CAST | 7.239 | 0.278 | 15.62 % |
| | | Sex, age | 7.086 | 0.270 | 13.19 % |
| | | Sex, CAST | 7.195 | 0.282 | 14.92 % |
| | | Age, CAST | 7.136 | 0.261 | 13.97 % |
| | | Sex, age, CAST | 7.086 | 0.266 | 13.18 % |
| | Sex and age | Sex | 7.216 | 0.283 | 15.26 % |
| | | Age | 6.837 | 0.243 | 9.20 % |
| | | CAST | 6.955 | 0.254 | 11.09 % |
| | | Sex, age | 7.136 | 0.272 | 13.97 % |
| | | Sex, CAST | 7.233 | 0.283 | 15.53 % |
| | | Age, CAST | 6.875 | 0.240 | 9.81 % |
| | | Sex, age, CAST | 7.145 | 0.269 | 14.12 % |

In this application, reweighting with PSA and a Hajek-type estimator is the less biased alternative when using gender, age and CAST score as PSA covariates. When using only gender and CAST scores, the estimator achieves the minimum standard error within all the alternatives. Overall, estimates reweighted with PSA or PSA and calibration to gender and age presented the best results, both in terms of least difference with the reference sample value and least standard error according to Jackknife.

# 5 Discussion and Conclusions

In the last years we are witnessing a strong development of online research methods in general and web surveys specifically. Web surveys are a very attractive option because fieldwork costs are rather low when compared with other modes as mail, telephone and face to face. In addition to cost-effectiveness, there are other reasons that explain why the market research industry has decidedly embraced web surveys in the last years such as the speed of data collection and the advantages associated with the computerization of the questionnaire and self-administration. However, currently the web survey mode has some limitations to adequately represent the general population. In spite of the fast adoption of the Internet in the last decades, the number of non-users is still important in most countries. Moreover, non-internet users differ significantly from those who have access and use this technology. As a result, web surveys that fail to include non-internet users are at a high risk of incurring in coverage bias. A second problem that hinders the use of probability sampling in web surveys of the general population is the lack of a proper sampling frame.

In this paper we have focused on the problem of the the lack of coverage of nonprobabilistic samples. It is obvious that such a problem can be responsible for a large increase in the bias of the final results. Various correction techniques, such as calibration and Propensity Score Adjustment or PSA, can be applied to remove the bias. This study attempts to analyze the efficiency of correction techniques in multiple situations, applying a combination of PSA and calibration.

The simulation study, which is a technique widely used when studying methods to improve the estimates provided by problematic surveys and particularly calibration or PSA (Lee (2006); Lee and Valliant (2009); Kim and Park (2009); Bethlehem (2010)), is performed in this work with several limitations, such as the variables selected for PSA and calibration and the diversity among possible situations.

Some of the results presented in this work successfully reproduce relevant findings of the existing literature. For example, it is proved in Bethlehem (2010) that bias can be highly reduced through calibration with the right covariates when the non-response due to volunteering has a MAR scheme, while it cannot be equally done in NMAR situations. This is similar to the results obtained in the simulation study; PSA achieves an improvement in the amount of bias much higher for MAR than for NMAR, but as a difference, the right covariates were used for PSA this time rather than for calibration. As a result, calibration fails to remove any bias if not combined with PSA. These results can be linked to Lee (2006), where it was stated that it is critical to add covariates related to the objective of the study, in order to make PSA useful. These findings are relevant in the sense of finding a procedure to remove coverage error when calibration with covariates is not possible; however, results also show that using estimates of the population totals does not cause any significant difference in final results, therefore the usage of the reference survey to estimate population totals of covariates might be considered for calibration purposes.

In addition, it is worth to note that this work introduces the comparison of the efficiency of Horvitz-Thompson and Hajek weights for PSA, a duality proposed in Schonlau and Couper (2017). Results of this study conclude that a difference in efficiency can be made between both approaches only if the right covariates and calibration totals have been chosen previously, and in fact the individual observed differences in weights computed in the simulation study are negligible. This could be explained by the fact that the strata formed with the propensity scores are thought to have individuals whose propensity score is very similar between them, something feasible given the features of the logistic regression model used for that purpose. Under these circumstances, it is very likely that stratification makes no effect in the computation of final weights. On top of that, PSA weights were subsequently used as original calibration weights, contributing to dilute even more the difference between the former.

Finally, the application of the developed adjustment methods in a specific volunteer survey reflects the conclusions of several studies performed in the past on PSA (Lee (2006); Valliant and Dever (2011)) that the choice of covariates used for the PSA plays a fundamental role on its further efficiency. However, as it happens in most of health-related surveys, this application is limited by the fact that there are no population totals that estimates can be compared with. Further studies should take into account the availability of population counts in their earlier research steps.

On the other hand web surveys, as any other survey, suffer from non-response even if the use of responsive or adaptive design features account for participation rates. Non sampling errors are particularly important when the investigator has to gather information concerning highly personal, sensitive, stigmatizing and perhaps incriminating issues such as abortion, drug addiction, HIV/AIDS infection status, duration of suffering from a disease, sexual behaviour... In these situations, collecting data by means of survey modes based on direct questioning methods of interview is likely to encounter two serious problems: (i) participants in the survey may deliberately release untruthful or misleading answers, or (ii) participants may refuse to respond (”unit nonresponse” or ”item nonresponse”) due to the social stigma or because they feel threatened by such inquiries and fear that their personal information may be released to third parties for purposes other than those of the survey.

A considerable limitation of the presented approach could be the Big Data issues that may arise when the volume of data gets larger. This is a feasible situation in Internet surveys, given that their characteristics allow for an important number of respondents to take part on them. The main potential limitation of PSA under these circumstances could be related to the adequacy of logistic regression as a predictor for propensity scores, as they would tend to oversimplify the actual relationships between covariates and target variables. The usage of some alternatives to these models, such as Machine Learning algorithms (e. g. classifiers), should be considered in future research in the area.

## Acknowledgements

The authors thank the valuable comments and suggestions given by two anonymous reviewers. This study was partially supported by the Spanish grant MTM 2015-63609-R.

## References

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review* 78(2), 161–188.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika* 63, 615–620.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics* 24(2), 295–313.

Couper, M. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly* 64(4), 464–494.

Couper, M. (2017). Developments in Survey Collection. *Annual Review of Sociology* 43, 121–145.

Couper, M., Kapteyn, A., Schonlau, M. and Winter, J. (2007). Noncoverage and Non-response in an Internet Survey. *Social Science Research* 36, 131–148.

Couper, M. and Peterson, G. (2017). Why Do Web Surveys Take Longer on Smartphones? *Social Science Computer Review* 35(3), 357-377.

Dever, J. A., Rafferty, A. and Valliant, R. (2008). Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? *Survey Research Methods* 2(2), 47–62.

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

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

Díaz de Rada, V. and Domínguez, J. A. (2015). The quality of responses to grid questions as used in Web questionnaires (compared with paper questionnaires). *International Journal of Social Research Methodology* 18(4), 337–348.

Díaz de Rada, V. and Domínguez, J. A. (2016). Mail survey abroad with an alternative web survey. *Quality and Quantity* 50(3), 1153–1164.

Elliott, M. R. and Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science* 32(2), 249–264.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research* 21(1), 111–121.

Kim, J. K. and Park, M. (2009). Calibration estimation in survey sampling. *International Statistical Review* 78(1), 21–39.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics* 22(2), 329–349.

Lee, S. and Valliant, R. (2009) Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research* 37(3), 319–343.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Manfreda, K. L., Berzelak, J., Vehovar, V., Bosnjak, M. and Haas, I. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research* 50(1), 79–104.

Martínez, S., Rueda, M., Arcos, A. and Martínez, H. (2010). Optimum calibration points estimating distribution functions. *Journal of Computational and Applied Mathematics* 233, 2265–2277.

Mei, B. and Brown, G. (2017). Conducting Online Surveys in China. *Social Science Computer Review* 0894439317729340.

National Institute of Statistics (2016). Población (espanoles/extranjeros) por edad (grupos quinquenales), sexo y ano. Retrieved from http://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p08/l0/&file=02002.px (Accessed 20 March 2018).

National Institute of Statistics (2017a). Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares. Retrieved from http://www.ine.es/prensa/tich2017.pdf (Accessed 20 March 2018).

National Institute of Statistics (2017b). Espana en Cifras 2017. Retrieved from http://www.ine.es/prodyser/espacifras/2017/index.html (Accessed 20 March 2018).

National Institute of Statistics (2017c). Nivel de formación de la población adulta (de 25 a 64 anos). Retrieved from http://www.ine.es/ss/Satellite?c=INESeccionC&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&cid=1259925481659&L=01 (Accessed 20 March 2018).

Pew Research Center (2017). Demographics of Internet and Home Broadband Usage in the United States. Retrieved from http://www.pewinternet.org/fact-sheet/internet-broadband/(Accessed 20 March 2018).

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1), 41–55.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation With Adjusted

Weights and Multiple Imputations. *Journal of Business & Economic Statistics* 4(1), 87–94.

Rueda, M., Sánchez-Borrego, I., Arcos, A. and Martínez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika* 71, 33–44.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33(2), 99–119.

Schonlau, M. and Couper, M. (2017). Options for Conducting Web Surveys. *Statistical Science* 32(2), 279–292.

Schonlau, M., van Soest, A., Kapteyn, A. and Couper, M. (2009). Selection Bias in Web Surveys and the Use of Propensity Scores. *Sociological Methods & Research* 37(3), 291–318.

Pasadas-del-Amo, S. (2018). Cell Phone-only Population and Election Forecasting in Spain: The 2012 Regional Election in Andalusia. *Revista Espanola de Investigaciones Sociológicas (REIS)* 162, 55–72.

Taylor, H. (2000). Does Internet research work? *International Journal of Market Research* 42(1), 51–63.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W. and Terhanian, G. (2001). 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.

Valliant, R. and Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research* 40(1), 105–137.

# Appendix A2

## Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys

Ferri-García, Ramón; Rueda, María del Mar (2020)
Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys
*PLoS ONE*, vol. 15, number 4, p. e0231500
DOI: 10.1371/journal.pone.0231500

PLOS
ONE
A peer-reviewed, open access journal

| MULTIDISCIPLINARY SCIENCES | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 2.740 | 27/71 | Q2 |

## Abstract

Modern survey methods may be subject to non-observable bias, from various sources. Among online surveys, for example, selection bias is prevalent, due to the sampling mechanism commonly used, whereby participants self-select from a subgroup whose characteristics differ from those of the target population. Several techniques have been proposed to tackle this issue. One such is Propensity Score Adjustment (PSA), which is widely used and has been analysed in various studies. The usual method of estimating the propensity score is logistic regression, which requires a reference probability sample in addition to the online nonprobability sample. The predicted propensities can be used for reweighting using various estimators. However, in the online survey context, there are alternatives that might outperform logistic regression regarding propensity estimation. The aim of the present study is to determine the efficiency of some of these alternatives, involving Machine Learning (ML) classification algorithms. PSA is applied in two simulation scenarios, representing situations commonly found in online surveys, using logistic regression and ML models for propensity estimation. The results obtained show that ML algorithms remove selection bias more effectively than logistic regression when used for PSA, but that their efficacy depends largely on the selection mechanism employed and the dimensionality of the data.

# Introduction

One of the main drawbacks of online surveys is the selection bias [1] that may be introduced in their use. This problem occurs when the population sample used differs from the non-observed population in such a way that the sample results cannot be extrapolated to the full population. In online surveys, samples are often drawn from volunteer participants, for reasons of time and financial economy, making this population nonprobabilistic and therefore unsuitable for the usual sampling methods employed for inference and estimation. Assuming that some groups are more likely than others to participate, volunteer samples present an inherent selection bias. Hence, determining optimum probabilistic sampling conditions in an online survey is not a trivial undertaking. As [2] state, probabilistic online frames can only be used when the population of interest is narrow (the members of well-defined organizations); evidently, if the target population is not properly defined a reliable sampling frame of internet users may not be achieved. Internet access is often associated with sociodemographic variables related to the variables of interest in a given study ([3]). For example, according to [4], the internet penetration rate in Spain is above 90% of the population in all population groups aged under 54 years; however, among persons aged 65 to 74 years, the penetration rate is only 43.7%. In consequence, the potentially covered population (as defined in [1]) is immediately subjected to a selection bias, which cannot be completely excluded by the usual reweighting methods ([5]; [6]).

In recent years, propensity score adjustment (PSA) has increasingly been used as a means of correcting selection bias in online surveys. This method, first proposed by [7], was originally intended to correct the bias introduced by factors associated with exposure (group allocation) and outcome in the experimental design, and studies have demonstrated its effectiveness in this regard ([8]; [9]). PSA, like most adjustment instruments in population sampling, is based on the use of auxiliary information. However, in addition to the nonprobabilistic volunteer sample, it also requires the availability of a probabilistic reference sample. This is usually obtained from a survey focused on a different subject area. Accordingly, it does not measure the present variable or variables of interest, but rather a set of covariates that have also been recorded or the nonprobabilistic sample. The reference survey does not have to address the same research questions, but it should be well conducted and avoid all sources of bias as much as possible.

The efficacy of PSA at removing selection bias from online surveys has been discussed in numerous studies. However, its performance depends on the covariates chosen. Moreover, the use of PSA generally increases the sampling variability of the estimators with respect to the unweighted case ([10]; [11]).Therefore, PSA weighting should be complemented with further calibration adjustments using complementary variables to make estimates less biased ([11]; [12]).

Propensity scores in PSA are usually estimated using logistic regression models, where the target variable is a binary indicator that takes 1 if an individual belongs to the nonprobabilistic sample and 0 otherwise. This approach is equivalent to estimating the probability of an individual volunteering to participate in a survey, given a specific set of covariates. Logistic regression provides estimates that are robust, i.e. they remain stable when new data are incorporated, and simple to implement in most statistical packages. However, they also present certain drawbacks that should be taken into account. Thus, in logistic modelling it is assumed that the log-odds risks have a linear relationship with the covariates ([13]). In the online survey context, this assumption could easily fail to hold, especially with larger samples and a greater number of covariates.

Alternatives to logistic regression in PSA have appeared in parallel with the development of machine learning (ML) classification algorithms. A vast and still-increasing number of ML approaches provide the raw probabilities of occurrence of a given class, both black-box and interpretable, the application of which in PSA has mainly been studied with respect to experimental design. Research into interpretable algorithms for PSA has focused on classification and regression trees (CART) ([16]), concluding that these decision trees provide less biased effect estimates, even under conditions of non-additivity and non-linearity ([17]; [18]). In this respect, [19] examined a special case of discriminant analysis, selecting the best classification tree in terms of optimality.

Among the black-box alternatives that have been developed in the field of ML, neural networks and bagging/boosting algorithms have attracted much attention. Neural networks are discussed in [15] as a potential replacement for logistic regression in PSA, but to our knowledge, they have only been successfully applied in [17]. In addition, bagging algorithms such as Random Forest ([20]) and boosting algorithms such as the Gradient Boosting Machine (GBM) ([21]) have been included in several studies. It has been suggested that the use of GBM could provide more stable weights and greater bias reduction than is the case with logit models ([22]) and multinomial models ([23]), at the cost of a minimal increase in variance. Similarly, the incorporation of Random Forest into PSA may also reduce the level of bias in the estimates obtained, compared to logistic models ([18] [24]) and classification trees ([25]). GBM has been successfully applied in real experiments with propensity scores ([26], [27], [28]). In this field, too, [29] studied the performance of boosted CART (GBM), but found that the performance of each algorithm was strongly dependent on the scenario. Finally, [30] analysed the efficiency of propensity estimation, using Random Forests for matching, taking into account the existence of missing data from the predictors, and reported good results for group balancing.

An interesting case of black-box algorithms is described by [31]. These authors used the Super Learner paradigm proposed by [32], and estimated propensity scores by choosing the best algorithm, in terms of goodness-of-fit, from a set of ML classification algorithms, including Bayesian Generalised Linear Models, Support Vector Machines, Multivariate Adaptative Regression Splines and k-Nearest Neighbours, apart from those mentioned above. This study showed that overall efficiency was dependent on the underlying covariate structure, but that PSA, using the Super Learner strategy, presented good balancing properties.

In recent survey research, ML algorithms have been widely studied in the probability sampling context ([33]; [34]; [35]; [36]; [37]). In nonprobability sampling, however, PSA has mainly been used in nonresponse propensity adjustments. The PSA procedure for addressing the question of nonresponse bias, which was first developed by [38], usually follows the same steps as in dealing with selection bias, but some alternatives to logistic regression have been proposed. Thus, [39] used local polynomial regression models to adjust nonresponse propensity estimates, in a paper extending their previous work on propensity estimates via kernel regression. Further details of this method are discussed by [40]. These models provide better estimates of propensity, in terms of likelihood, and lower variance than is the case with logistic regression models, provided that the polynomial degree is properly specified. Applications of ML algorithms in PSA for nonresponse propensity have been studied for classification and regression trees ([41]) and Random Forests ([42]); their ability to reduce nonresponse bias, in comparison with logistic regression, depends on the covariates available and on the complexity of the relationships. These techniques for modelling nonresponse propensity are also addressed by [43].

In the present paper, the ML approach is extended to the question of reducing selection bias, considering various online survey scenarios that are subject to selection bias and examining how PSA may reduce this bias, according to the algorithm used to compute the propensity estimates. The study method and the ML methods used are described in detail, after which we present a simulation study based on artificial and real-world data. The implications of these results are then discussed, and in the final section we suggest related lines of work for future research.

## Propensity Score Adjustment (PSA) for volunteer online samples

The procedure to perform Propensity Score Adjustment for removing volunteer bias in online surveys can be described as follows: let $s_v$ be a volunteer nonprobabilistic sample of size $n_{vs}$, 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 $i$-th individual. Note that each element in both samples has a base weight associated, say $d_j^v$, $j = 1, ..., n_{vs}$ for the volunteer sample and $d_k^r$, $k = 1, ..., n_{rs}$ for the reference sample (usually, $d_k^r = 1/\pi_k$). Covariates $\mathbf{X}$ used to adjust the propensity scores have been measured on both samples, while the variable of interest $y$ has been measured only in the volunteer sample, and the probabilistic sample cannot be directly used for its estimation as a result. Let $z$ be a binary variable which measures whether a participant of the complete sample $s = s_r \cup s_v$ belongs to $s_r$ or $s_v$.

$$z_i = \begin{cases} 0 & i \in s_r \\ 1 & i \in s_v \end{cases}, \quad i = 1, ..., n, n = n_{vs} + n_{rs} \tag{1}$$

Let $\pi(\mathbf{x}_i)$ be the propensity score for participant $i$ conditional on his/her covariates' value $\mathbf{x}_i$. $\pi(\mathbf{x}_i)$ reflects the probability of $z_i = 1$ given the set of covariates $\mathbf{X}$. The reference sample is assumed to suffer from a small selection bias or no bias at all, and can be used to generate a reliable estimate of the covariates' distribution in the target population. This information could be used to calculate which types of individuals are more or less prone to participate in an online survey. The above-mentioned propensity scores, $\hat{\pi}(\mathbf{x}_i)$, are often estimated using a logistic regression model which can be described as in Eq. 2.

$$\hat{\pi}(\mathbf{x}_i) = \frac{1}{e^{-(\gamma^T \mathbf{x}_i)} + 1} \tag{2}$$

where $\gamma$ is the vector of regression coefficients obtained in the modelling process. The original online sample is reweighted using the propensity estimates to take into account the information on selection bias provided by PSA. This procedure can be performed using weights for either the Horvitz-Thompson or the Hajek estimators; the procedure for the Horvitz-Thompson-type weights is described in [10] and [11] and can be summarised as follows. The combined sample is sorted and then divided into $C$ classes ([44] recommend the use of five classes) according to each individual's propensity score. An appropriate adjustment factor $f_c$ is obtained using Eq. 3

$$f_c = \frac{\sum_{k \in s_r^c} d_k^r / \sum_{k \in s_r} d_k^r}{\sum_{j \in s_v^c} d_j^v / \sum_{j \in s_v} d_j^v} \tag{3}$$

where $s_r^c$ and $s_v^c$ are individuals from the reference sample and the volunteer sample respectively, belonging to the $c$-th class. The new weights $w$ for individuals in the volunteer sample are then calculated as follows:

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

Hajek-type weights can be calculated as described in [2] and according to Eq. 5. In this case, the weights adjust the volunteer sample to the population of the probabilistic sample, $U_r$, rather than the complete population $U$.

$$w_j = \frac{1 - \hat{\pi}(\mathbf{x}_j)}{\hat{\pi}(\mathbf{x}_j)} \tag{5}$$

According to [45], the difference between the two approaches to the final estimates depends both on the discreteness of the support of the covariates and on the selection mechanism used (i.e., whether or not it is related to the target variable).

## Machine Learning classification algorithms for propensity score estimation

As described above, various alternatives to logistic regression can be used in propensity score estimation, leading to different formulas to obtain $\hat{\pi}(\mathbf{x}_i)$. In this section, we present some formulas associated with the application of some algorithms commonly used in PSA literature, together with other techniques frequently seen in data mining ([49]), namely decision trees, Random Forests, GBM, k-Nearest Neighbours and Naïve Bayes.

Decision trees can be defined as a set of rules organised in a hierarchical structure, starting from an initial node that represents the complete dataset. To predict a given individual, the dataset is split into different subsets according to a rule based on an input predictor variable. Each subset can also be split, successively, until a convergence criterion is met; then, the rule stops increasing in complexity, and a terminal node is reached. Any input individual for the decision tree will meet the criteria of a rule specified for it, and thus predicted according to data from the individuals meeting the rule criteria. In our study, the algorithms used for tree building involve C4.5, C5.0 ([50]) and CART ([16]). They differ in some aspects of tree building, such as the rule pruning and complexity, but for brevity these questions are not addressed in the present paper.

This approach can be used to obtain the probabilities of the input individuals of a decision tree belonging to any given class. In this context, these probabilities represent the individuals' propensity to participate in an online survey, where $z$ represents the binary target variable. Let $J_1, \ldots, J_k$ be the set of rules (terminal nodes) of a decision tree; each rule represents a multidimensional range for each covariate, say: $J_i = \{\mathbf{X} \in B_i\}$ where $B_i \in \mathbb{R}^p$, and where $p$ is the number of covariates. Let

$n(s_v^{J_i})$ and $n(s^{J_i})$ be the number of volunteer sample and combined sample members, respectively, which meet the criteria of the $i$th terminal node. The formula for estimating propensity scores for an individual $i$ using decision trees is described in Eq. 6.

$$\hat{\pi}(\mathbf{x}_i) = \begin{cases} \frac{n(s_v^{J_1})}{n(s^{J_1})} & \{i \in s / \mathbf{x}_i \in J_1\} \\ \dots & \dots \\ \frac{n(s_v^{J_k})}{n(s^{J_k})} & \{i \in s / \mathbf{x}_i \in J_k\} \end{cases} \tag{6}$$

In the case of Random Forests, propensities are estimated by averaging the number of times that an input individual is classified in the class representing the presence (often denoted as ”1”) through a set of $m$ trees known as *weak classifiers*. Input variables for each tree are randomly selected, in subsets of fixed size, from the available covariates. Therefore, the propensity score estimation can be reformulated as in Eq. 7.

$$\hat{\pi}(\mathbf{x}_i) = \frac{\sum_{j=1}^{m} \phi_j(\mathbf{x}_i)}{m}, \ \phi_j(\mathbf{x}_i) = \begin{cases} 1 & \{i \in s / \mathbf{x}_i \in J_{pr}\} \\ 0 & \{i \in s / \mathbf{x}_i \in J_{ab}\} \end{cases} \tag{7}$$

where $J_{ab}$ and $J_{pr}$ represent the set of terminal nodes where individuals from the volunteer sample are minority and majority, respectively. In other words:

$$J_{pr} = \{J_l, l = 1, ..., k : \frac{n(s_v^{J_l})}{n(s^{J_l})} \geq 0.5\} \tag{8}$$

$$J_{ab} = \{J_l, l = 1, ..., k : \frac{n(s_v^{J_l})}{n(s^{J_l})} < 0.5\} \tag{9}$$

Note that in the cases where the volunteer and the reference sample are very unbalanced in size, the propensity scores may be exactly zero or one for some individuals and in such cases cannot be properly applied. In some studies, adjustments have been made in order to avoid this situation; for instance, [42] applied a $(1000 \cdot x + 0.5)/1001$ transformation to move the propensities away from zero and one.

For $k$ nearest neighbours, computing the propensity score estimates involves a distance function $d$ which measures the closeness of each data point to a given individual $i$ using covariates $\mathbf{X}$. This distance allows the $n-1$ individuals to be rearranged as $\mathbf{x}_{(1)}, ..., \mathbf{x}_{(n-1)}$, where $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(n-1)}$ represent, respectively, the covariates of the closest and the furthest individual from $i$ according to $d$. As the target variable is binary, the propensity scores can be estimated with the following formula:

$$\hat{\pi}(\mathbf{x}_i) = \frac{\sum_{j \in s / d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k)})} z_j}{k} \tag{10}$$

Application of the formula shown in Eq. 10 is equivalent to calculating the proportion of individuals from the volunteer sample out of the $k$ nearest neighbours to the individual $i$. The number of neighbours $k$ is arbitrary, meaning that $k = 1$ or even a small enough $k$ will provide probabilities of zero or one.

Estimation of the propensity scores using the Naïve Bayes algorithm is based on the Bayes formula, derived from the observed probabilities of participants belonging to the volunteer sample and the occurrence of a given vector for $\mathbf{X}$, that is, the values of the covariates for a given individual $i$.

$$\hat{\pi}(\mathbf{x}_i) = \frac{P(z_i = 1)P(\mathbf{X} = x_i | z_i = 1)}{P(\mathbf{X} = \mathbf{x}_i)} \tag{11}$$

If variables with very rare classes or presenting high cardinality are used as covariates, the propensity estimates might present values significantly far from the real propensity.

Finally, when using a GBM algorithm ([21]), the formula for propensity score estimation has the same structure as that used in logistic regression, but is based on a different parametrisation:

$$\hat{\pi}(\mathbf{x}_i) = \frac{1}{e^{-w^T J(\mathbf{x}_i)} + 1} \tag{12}$$

where $J(\mathbf{x}_i)$ represents a matrix of terminal nodes of $m$ decision trees (the number of trees used is decided by the user but should be correlated with the sample size, as in Random Forests) and $w$ is a vector representing the weights of each tree. The development of trees in $J(\mathbf{x}_i)$ is achieved through an iterative process minimising of the specified loss function for a small sample of the input dataset (which, in this context, is assumed to be the combined sample $s$) selected for testing purposes.

# Simulation study

## Artificial data

To evaluate the performance of classification algorithms applied under different circumstances for PSA, we conducted an experiment using a fictitious population of voters. This population was used originally by [45], following an experiment by [5] to measure the efficiency of adjustments in selection bias. For the present study, minor changes were made to the gender distributions by age so that a proper Missing Completely At Random (MCAR) situation could be simulated. This population and the experiment are detailed below.

The fictitious population had a total size of N = 50,000 individuals. The study aim was to estimate the fraction of votes obtained by each of three fictitious parties, Party 1, 2 and 3, in a hypothetical election. Four sociodemographic variables – age, nationality, gender and education – were measured in each sample and used as covariates for the PSA models.

- The age variable was determined by applying the following transformation of a simulated Beta distribution: $Age = 82x + 18, x \sim \beta(2,3)$. The resulting age pyramid was similar to the real-world case in Spain ([46]).

- In the study population, 15% were non-natives aged under 35 years, 10% were non-natives aged 35 to 65 years, and 2.5% were non-natives aged over 65 years, which is similar to the nationality distribution by ages in Spain ([47]).

- The probability of the individual being male or female was identical, for the whole population, in contrast to the experiment performed by [45].

- Education was constrained to be dependent on the age strata (the same strata as for the Nationality variable), in order to make it similar to the pattern of education levels among Spanish adults ([48]).
  - Of the individuals aged under 35 years, 35% had only primary education, 20% had secondary education and 45% had higher education.
  - Of the individuals aged 35 to 65 years, 45% had only primary education, 25% had secondary education and 30% had higher education.
  - Of the individuals aged over 65 years, 80% had only primary education, 10% had secondary education and 10% had higher education.

In addition, internet access was made dependent on age and nationality. Among non-natives, internet access was available to 20% of those aged under 35 years, but only to 10% of those aged 35 to 65 years and to 0% of those aged over 65 years. In contrast, for natives the corresponding values were 90%, 70% and 50% respectively.

The probabilities of a person voting for Party 1, 2 or 3 were considered in relation to the above variables. Party 1 would attract the votes of 20% of the female population, but the men would not vote for it at all. As internet access did not depend on gender, measuring the proportions of the population who would vote for Party 1 could be considered an example of a Missing Completely at Random (MCAR) selection mechanism. For Party 2, the voting probability increased in line with the voter's age; among the population as a whole, nobody aged under 35 years would vote for this party, while 40% of those aged 35 to 65 years would do so, as would 60% of those over 65 years old. The above-mentioned relationship between age and internet access means that the measurement of voting intentions for Party 2 is also subject to a Missing At Random (MAR) selection mechanism. Finally, voting intentions for Party 3 depended on both age and internet access: thus, 10% of individuals with no internet access (regardless of their age) would vote for this party, while among those with internet access, the party would attract the votes of 60% of those under 35 years, 40% of those aged 35 to 65 years, and 20% of those aged over 65 years. These relationships mean that the measurement of voting intentions for Party 3 is subject to a Not Missing At Random (NMAR) selection mechanism, as the target variable is in fact related to the selection variable.

The distribution of values for the population as a whole and for each of the subpopulations, with and without access to internet, is shown in Table S1. As expected, there is a slight divergence in voting intentions for Party 2 between the population as a whole and those with internet, and a strong divergence in this respect for Party 3. Persons with internet were more likely to have completed a course of higher education, were on average two years younger and were five times less likely to be non-native. However, differences in gender were negligible, and so voting intentions for Party 1 were barely affected.

To estimate voting percentages for each party, we selected a convenience sample from the population with internet access, and a reference sample from the full population. The reference sample was drawn by simple random sampling without replacement (SRSWOR), and three different sampling schemes were tested to select the convenience sample:

1. SRSWOR from the whole internet population.

2. Sampling from the whole internet population with unequal self-selection probabilities, obtained by the following formula:

$$\pi_i = \frac{1}{1+e^{-1+0.05\cdot\mathrm{Age}_i}}, i = 1,2,\ldots,31,881 \tag{13}$$

   where $\mathrm{Age}_i$ is the age of the $i$-th individual of the internet population.

3. Sampling from the whole internet population with unequal self-selection probabilities, obtained by the following formula:

$$\pi_i = \frac{1}{1+e^{1-sin(\mathrm{Age}_i/20)}}, i = 1,2,\ldots,31,881 \tag{14}$$

   where $\mathrm{Age}_i$ is the age of the $i$-th individual of the internet population.

The formulas for the inclusion probabilities in schemes 2 and 3 were tested to evaluate how ML algorithms perform in comparison with logistic regression when the relationship between the covariates and the selection probability (which we assume to be the self-selection probability) can be modelled using the logit function, with either linear or nonlinear relationships. The experiment was replicated varying the convenience sample size across $n_{vs}$ = 500, 750, 1,000, 2,000, 5,000, 7,500 and 10,000, and the size of the reference survey was established at 500 individuals for each replication. The replication results were obtained by averaging the bias and calculating the MSE of the estimates in 500 simulations. The mean bias of each replication was obtained according to Eq. 15:

$$\mathrm{Bias}_k = \frac{\sum_{m=1}^{500} \hat{p}_m^k}{500} - p^k \tag{15}$$

where $\hat{p}_m^k$ is the proportion of voters for Party $k$ estimated in the $m$-th simulation and $p^k$ is the real proportion of voters for Party $k$. The MSE for the estimators in each replication was obtained directly from the estimates, as in Equation 16:

$$\mathrm{MSE}_k = \frac{\sum_{m=1}^{500}(\hat{p}_m^k - \overset{\frown}{\bar{p}}^k)^2}{499} + (\mathrm{Bias}_k)^2 \tag{16}$$

where $\overset{\frown}{\bar{p}}^k$ is the mean of the estimates for the proportion of voters for Party $k$.

## Real data

A set of real data was analysed to determine the usual patterns observed in real applications. This real-data approach is commonly employed in studies of PSA ([10]; [11]; [12]).

The dataset used to simulate a pseudo-population was obtained from the microdata of the 2012 edition of the Life Conditions Survey (known by the Spanish acronym, ECV) ([51]). This annual survey is conducted face-to-face by the Spanish Institute of Statistics, targeting the entire Spanish population aged 16 years or older. The primary unit considered is the household, and the secondary units are the members of the household. The variables considered include income, poverty, equality, employment and household living conditions. In 2012, 12,714 households were surveyed, providing a sample population of 33,573 individuals. In this study, the full sample had to be preprocessed before the simulation due to the considerable volume of missing data. After this filtering process, the size of the pseudo-population (that is, the full filtered dataset) was N = 28,210, and 61 variables were identified as potential covariates for the PSA models.

The convenience and reference samples were selected by SRSWOR from the volunteer (internet) population and the full population, respectively. The identifying variable for volunteers and non-volunteers was the presence of a computer in the household. According to the 2012 Spanish Survey on Equipment and Use of Information and Communication Technologies in Households ([52]), 90.1% of persons who had a computer in their household also had internet access, and 98.3% of those with internet access at home also had a computer. Therefore, we believe it reasonable to assume that taking the presence of a computer in the household as the selection variable is a very good proxy of a variable measuring internet access in the household. Two target variables were considered:

- The proportion of the population whose self-reported health was poor (those who responded ”poor ” or ”very poor” to the question regarding their general state of health.

- The proportion of the population living in a household with more than two members.

The experiment was replicated 500 times, varying the size of the convenience sample across $n_{vs}$ = 500, 750, 1,000, 2,000, 5,000 while the reference sample size was maintained at $n_{rs}$ = 500, and considering the following groups of covariates:

- Group 1: Nine covariates measuring region, size of home town/city, gender, marital status, nationality, country of origin and education level (both achieved and currently studying).

- Group 2: All the covariates in Group 1 plus five health-related variables, namely chronic diseases, presence of disability, and lack of access to medical and/or dentistry services (and reasons for this lack).

- Group 3: All the covariates in Group 1 plus eleven poverty-related variables, namely delays in bill payment, incidence of bills on the household economy, difficulty in living within household income, ability to acquire certain household goods, possession of electrical appliances, income needed to live without financial difficulty and calculated indicators of poverty risk and material scarcity.

- Group 4: All 61 potential covariates. All of the above variables plus working conditions, care provision, energy poverty and household conditions and expenditure.

The S1 Dataset includes the full dataset used to perform these analyses.

## Algorithms and parameter tuning

The procedure in both simulations was the same: in each of the 500 simulations, convenience and reference samples were selected, PSA was applied to reweight the convenience sample using Hajek-type weights, and the population parameter was estimated using the convenience sample with PSA. Measures of bias and MSE for each scenario, algorithm and $n_{vs}$ were estimated as in (15) and (16). This procedure was implemented in the statistical software R ([53]) using the packages *RWeka* ([54]; [55]),*C50* ([56]), *rpart* ([57]), *randomForest* ([58]), *e1071* ([59]) abd *gbm* ([60]). Packages *ggplot2* ([61]), *xlsx* ([62]), *gridExtra* ([63]) and *RColorBrewer* ([64]) were used to generate the figures illustrating the results.

Propensity scores were calculated in each case using logistic regression, C4.5, C5.0, CART, k-nearest neighbors, Naïve Bayes, Random Forest, and GBM. For exploratory purposes, in the artificial data simulation the parameter configuration of each algorithm was selected on a grid, as follows:

- Decision trees (C4.5, C5.0 and CART) were applied taking 0.1, 0.25 and 0.5 as confidence values for pruning, and 0.5%, 1% and 5% of the dataset as the minimum number of observations per node.

- K-Nearest Neighbours was applied taking $k = 3,5,7,9,11,13$.

- Naïve Bayes was applied with a slight Laplace smoothing for the values $0,1,2,5,10$.

- Random Forests were generated with 500 trees and $1,2,4$ sampled variables for each tree.

- GBM was applied with interaction depths of 4, 6 and 8, and learning rates of 0.1, 0.01 and 0.001.

The impact of tuning parameters in PSA is still poorly understood, and optimality criteria are lacking. In this context, goal of classification algorithms is not to achieve greater accuracy but a higher likelihood for the propensity of volunteer participation in an online survey ([11]).

Parameter tuning was implemented for real data simulation. Thus, 10 times repeated 10-fold cross-validation was performed for each scenario, algorithm and $n_{vs}$ using the *caret* package in R ([65]), except for the CART algorithm, for which the cross-validation was coded separately, as *caret* does not allow us to refine the minimum number of observations per node. Log-Loss optimisation was used, as this metric bettter explains the deviation of estimated propensities from real participation. The parameter grids were as described above, with the following exceptions: the sampled variables for the Random Forest trees were taken as $\sqrt{p}$, $p/2$ and $p$, where $p$ is the number of covariates. In C5.0, we did not optimise the confidence value for pruning and minimum number of observations in the nodes. The optimal values obtained for C4.5 were used in C5.0, because the two algorithms are closely related and likely to behave in a similar way. On the other hand, the trials, type of model (rule-based or tree-based) and winnowing (feature selection) were tuned in C5.0. The results obtained are summarised in Table S2.

## Results

### Artificial data

Tables S3 and S4 show the bias and MSE results, respectively, obtained from using PSA with ML algorithms and SRSWOR from the internet population to build the convenience sample. There are small differences in bias reduction between C4.5, C5.0 and CART, especially for larger volunteer sample sizes. For Party 1, these algorithms outperform logistic regression only when the volunteer sample size is small, converging to the unadjusted case for larger samples. For Parties 2 and 3, the three algorithms are only better than unadjusted estimations when the sample sizes are balanced, but they never improve on PSA estimates using logistic regression. The MSE estimators with C4.5, C5.0 and CART also converge to the unadjusted case, which is smaller than PSA with logistic regression for Party 1 but greater for Parties 2 and 3. The parameter tuning of decision trees (with any algorithm: CART, C4.5 or C5.0) has no significant effect on bias removal, although greater confidence in the pruning appears to be slightly advantageous in the case of Parties 2 and 3 if the C4.5 algorithm is used and the sample sizes are relatively balanced.

The use of the k-NN classifier yields less biased estimates than that of baseline logistic regression for all of the missing data mechanisms considered. Thus, PSA with k-NN provides estimates that are less biased, on average, than the unadjusted estimates and the default-PSA reweighted estimates for Party 3. For Party 2, reweighting with PSA using k-NN transforms the bias in the opposite direction to the original bias; however, this bias is lower than that produced by the PSA estimates with logistic regression in absolute terms. For Party 1, k-NN provides less biased estimates than logistic regression but a low value for $k$; moreover, larger sample sizes are required. When estimating the likelihood of an individual voting for Party 1 or 2, the estimator MSE is smaller when PSA is used with k-NN rather than logistic regression, for larger numbers of neighbours and balanced sample sizes. When this is done for Party 3, the MSE of PSA with k-NN is significantly lower than for PSA with logistic regression, although large values of $k$ are required for full efficiency.

Application of the Naïve Bayes classifier in PSA produces a substantially greater reduction in bias than when PSA is performed with logistic regression, but only for the case of Party 2. For the other two parties, Naïve Bayes does not outperform logistic regression in PSA in terms of estimation bias except when samples are balanced and larger integers are used for Laplace smoothing. In addition, the MSE of the estimators is smaller with Naïve Bayes when the sample sizes are balanced and Laplace smoothing uses larger integers. The improvement, however, is rather limited.

Propensity estimation with the Random Forest algorithm is only advantageous in terms of bias removal in the estimations for Party 3, in which case the Random Forests algorithm achieves the highest bias reduction of all the classifiers reviewed. This is an important finding, as this missing data mechanism is particularly troublesome and, moreover, is commonly encountered in real data. The results for the MSE estimators under PSA with Random Forests show that this value may be only half that obtained with PSA and logistic regression for Party 3. The number of candidate variables for tree growing provides better results, remaining low for balanced sample sizes, but high for larger samples.

Finally, the efficiency of PSA reweighting with GBM is crucially dependent on the parameter configuration employed. For all kinds of missing data mechanisms, PSA with GBM removes bias more effectively when the learning rate is relatively low; thus, for Party 2, the bias reduction is almost complete. The MSE of the estimators reveal that GBM for PSA is advantageous for Parties 1 and 2 if parameter fitting is adequate (lower learning rates for Party 1, higher ones for Party 2), and significantly advantageous for Party 3 when the learning rate is high. The effects of interaction depth are mainly apparent with larger volunteer sample sizes, and greater interaction depths provide estimations with lower levels of bias and MSE.

Tables S5 and S6 show the results obtained from using PSA with ML algorithms with unequal selection probabilities in the internet population, following the logistic formula described in Eq. 13 for convenience sampling, for bias and MSE, respectively. As in the previous scenario, bias reduction with PSA using decision trees (C4.5, C5.0 or CART) converges to the unadjusted case as the convenience sample size increases. The best results are provided by C4.5 trees, but these are still much worse than those obtained with logistic regression. Regarding MSE, the lack of variability produced by the inability of decision trees to grow in samples with a large fraction of volunteer respondents leads them to have a smaller error than is the case with logistic regression in the estimation of intentions to vote for Party 1, especially with CART. Parameter tuning has a noticeable (albeit small) effect only when the samples are relatively balanced in size: with C4.5, higher confidence in pruning leads to better results, while with CART the opposite is true.

Using the k-NN algorithm in PSA produces a greater bias reduction than that of logistic regression for Parties 2 and 3, provided the number of neighbours, $k$, and the sample size are sufficiently large. The increase in variability provoked by the use of this algorithm makes the MSE slightly higher than with logistic regression in the intention to vote for Party 2. However, this is not the case regarding Party 3, where PSA with k-NN provides estimates with less error. In the case of Naïve Bayes, and regardless of the Laplace smoothing used, the bias and MSE are greater for Party 2 than with logistic regression, but these values are smaller for Party 3. Comparatively, Naïve Bayes in PSA provides estimates which produce a smaller error than either logistic regression, decision trees or k-NN.

The bias removal provided by PSA with Random Forest is strongly dependent on the size of the convenience sample and the number of variables sampled to create the trees. The bias for Party 3 is close to zero when the convenience sample is around four times larger than the reference sample and only two variables are sampled. If four variables are sampled, the bias reduction is greatest when the sample is 10 to 15 times greater. These results show that the Random Forest algorithm again provides the best MSE results in estimating voting probabilities for Party 3.

The GBM algorithm applied in PSA for sampling with unequal selection probabilities produces a very similar situation to SRSWOR, except that efficiency decreases in line with the size of the convenience sample size. When comparing the MSE in the voting estimation for Party 2 with that of k-NN and logistic regression, GBM is poorer with small sample sizes but better with larger ones. Accordingly, GBM is the best option for estimating voting intentions for Party 2 when a large convenience sample is available.

Tables S7 and S8 show the results for unequal selection probabilities in the internet population following the logistic formula described in Eq. 14 for convenience sampling, for bias and MSE, respectively. The performance of all the algorithms, taking into account that the amount of inherent bias is smaller, is very similar to the previous case. Among the differences observed, it should be noted that bias reduction is worse with k-NN (especially in estimating voting intentions for Party 2, for which this algorithm performs no better than logistic regression) and that Naïve Bayes performs better for Party 2 but worse for Party 3.

Table S9 summarises the bias and MSE measures obtained for each algorithm and selection mechanism, revealing certain characteristic patterns. For Party 1, while Naïve Bayes provides the lowest mean bias and is the best adjustment more frequently than the other algorithms, decision trees are better choices in terms of MSE, especially CART. For Party 2, bias reduction is dominated by PSA with k-NN and GBM but the former is surpassed by GLM in terms of MSE. Finally, Random Forest seems to be the best algorithm for PSA regarding voting intentions for Party 3, both in terms of bias and MSE. In general, ML algorithms (except for decision trees) produce the largest reductions in bias and, in many cases too, the lowest MSE.

### Real data

Table S10 show the results obtained for the bias present in estimating the fraction of the population who perceive their health to be poor. The table rows show the estimations obtained after PSA reweighting with the four covariate groups. These results clearly reflect the importance of the variables regarding PSA efficiency; when only demographic variables are included, PSA with Naïve Bayes provides the least biased estimates for all sample sizes, but also greater variance than the other methods and hence a larger MSE. In consequence, PSA with logistic regression provides the smallest error term. The bias is smaller when variables related to the outcome (health) or the exposure (poverty) are included in the models, with the former group leading to greater reductions in bias and MSE, but in this respect the situation for algorithms is unaffected. However, when all available covariates are used, PSA with Naïve Bayes appears to produce high levels of bias, while PSA with logistic regression is almost unbiased for large volunteer sample sizes, at the cost of high variance. As a result, MSE values are poor for PSA with logistic regression, while decision trees (for small $n_{vs}$) and bagging/boosting algorithms (for large $n_{vs}$) have the smallest term of error. The estimation with the lowest MSE was achieved using PSA with logistic regression together with demographic and health-related predictors, followed by PSA with GBM using all available predictors .

Table S11 shows the bias estimates for the fraction of households with more than two members, after PSA reweighting. It is noticeable that PSA with Random Forest removes most of the original bias as the volunteer sample size increases when only demographic variables are used, to the point that the MSE of the estimates obtained by PSA with Random Forest with $n_{vs}$ is the lowest of all those observed during the experiment. This pattern continues when health-related variables are added, although the bias of the estimates increases. On the other hand, if small volunteer samples are used, PSA with logistic regression provides the estimates with the smallest error term, and this true for all sample sizes when poverty covariates are used. When all covariates are included, a similar pattern is observed: thus, decision trees and GBM (with the latter providing the second-lowest MSE of the experiment), are the best algorithms for PSA when small and large sample sizes, respectively, are available.

## Discussion

New technologies have had a profound impact on surveying techniques worldwide. This impact is especially significant for social and political surveys, and most particularly for market research surveys, where the speed increases and cost reductions achieved with new technologies have radically changed the ways in which data are compiled. While in many cases the public sector continues to conduct interviews face-to-face and/or via telephone landlines, private companies are using mobile phones, tablets and the web, as standalone or combined strategies, thus obtaining data from volunteer participants. On the other hand, the results obtained with such nonprobability surveys present various problematic issues, notably the absence of a sample design assigning weights to the sample units, the presence of frame coverage issues and the risk of nonresponse bias. Although many statistical methods have been proposed to alleviate the problems of noncoverage and nonresponse, the question of nonrandomness in the sample is more complex and has not been thoroughly addressed.

In this respect, [66] reviewed existing inference methods to correct for selection bias in nonprobability samples. These authors considered a situation where only a nonprobability sample is available and compared a range of predictive inference methods (pseudo-design-based and model-based) in a general framework. The conclusion drawn from this study was that machine learning methods should be incorporated to address the problem of misrepresentation in nonprobability samples.

The present study considers another class of methods that may be used to correct selection bias in volunteer online surveys, which combine a nonprobability sample with a reference sample in order to construct propensity models. Our analysis compares logistic regression and ML classification algorithms for propensity estimation to determine the extent to which ML may be considered a viable alternative. ML algorithms present certain advantages over logistic regression; for example, they present greater flexibility, and do not require the analyst to specify a model with its interactions on nonlinear relationships, as ML is capable of capturing these relationships in the data learning procedure. Our study considers situations with few and with many covariates, for three different missing-data mechanisms influencing the selection process, and for different parameter configurations in the classifiers. To our knowledge, the only previous studies of the efficiency of classifier parameter tuning in PSA are those of [17] and [18] for decision trees, and to a more limited extent than in the present case. In addition, [42] alluded to some preliminary tests for Random Forest parameters, suggesting that optimum parameter selection would improve the estimations achieved.

The results we present show that most of the algorithms evaluated may provide a valid alternative to logistic regression in PSA if circumstances make the latter inappropriate. The C4.5 and C5.0 algorithms for decision trees are particularly useful when reference and volunteer sample sizes are balanced and the variables are numerous. Decision trees can be considered as variable selectors, as they automatically select subsets of optimal variables for classification, which is advantageous when the dimensionality is high ([67]). However, they also increase estimation variance when used for PSA, especially when there are significant nonlinear relationships between variables and the sample size is small ([18]; [29]).

The k-Nearest Neighbours (k-NN) algorithm is another useful alternative to logistic regression in PSA if the number of covariates available is low, especially with NMAR selection. However, as the dimensionality increases, k-NN becomes less efficient than other approaches. Its behaviour in both low and high-dimensional contexts was studied by [68], who concluded that higher dimensionality results in more concentrated distances, which makes k-NN less explicative of the actual class of an individual.

Our evaluation of Naïve Bayes in PSA for controlling selection bias revealed the existence of certain very clear patterns. When used with balanced sample sizes and few covariates, and not presenting rare or infrequent values, this algorithm provides smaller MSE values. In any other case, although PSA with Naïve Bayes behaves in an unstable way, simulations for NMAR using real data show that MSE is also substantially reduced. Ideally, Naïve Bayes should be employed with discrete input variables, as the probability computation performed by the algorithm is based on cross-tabulations. In addition, Naïve Bayes assumes independence between the variables, which may not be realistic in a high dimensional context due to the redundancy and noise issues that often arise (see [67]).

The application of bagging and boosting algorithms produced interesting results. Random Forest, which has been widely tested for PSA ([18]; [24]; [23]; [25]; [31]; [42]; [30]), achieved the largest bias reduction when the selection mechanism was NMAR, both for simulated data (under the condition of sample balancing) and with real data. However, its application presented several drawbacks, especially the fact that it is very prone to overfit propensity estimates on the data, as was apparent in the MSE of the Random Forest estimates with PSA, which tended to decrease and stabilise as the volunteer sample size increased. This pattern of behaviour has been reported previously by [25] for treatment effect estimates, and by [42], who observed an increase in variance when Random Forests of classification trees were used. On the other hand, the GBM, also referred to in the literature as boosted CART ([18]; [31]), provided weights that resulted in more stable behaviour of the estimates, as has also been noted previously ([18]; [23]). The GBM is efficient if the parameters are correctly tuned and the covariates are sufficiently discriminant. In this respect, [22] proposed a default parameter configuration for the GBM with low interaction depth and shrinkage. In the simulated data example described, better results were obtained with greater interaction depths. On the other hand, the best results with artificial data simulation were obtained when the learning rate was maximal; this parameter is related to overfit, and therefore should not produce a different pattern of behaviour in other situations. Nonetheless, further research is needed on the question of GBM parameter fitting. Finally, let us note that PSA with GBM in the real data simulation provided the best results in terms of MSE, for a large volunteer sample and when all available covariates were used.

## Conclusion

Our study findings support the use of ML algorithms as an alternative to PSA for reducing or eliminating selection bias in online surveys, although logistic regression is also shown to be a robust, reliable technique for propensity estimation. The efficiency of ML algorithms is closely related to the type of data considered and therefore no single approach is optimum for every case. We provide evidence with respect to MCAR, MAR and NMAR selection mechanisms, and for situations of low or high dimensionality. When selection follows a MCAR scheme, CART and GBM are the best alternatives, although the other ML algorithms tested, except Random Forest, also improve upon the results obtained by PSA with logistic regression, especially as the volunteer sample size increases. With MAR or NMAR selection, logistic regression generally provides good adjustments, especially when the dimensionality is low and the covariates are not very discriminant. However, if more covariates are available, logistic regression tends to destabilise and the MSE increases, despite its improved performance in bias removal; in this case, GBM, k-NN, decision trees and Random Forests all represent good alternatives. Random Forests provides good results when the data are MCAR, even if covariates are nonsignificant, although more research is needed on the possible incidence of overfitting on the final results obtained. The presence of balancing and overfitting issues suggests that data preprocessing should be a key step in the estimation of propensity scores, as observed previously by [19]. We recommend that further studies should consider the application of data preprocessing techniques such as noise filtering, sample balancing or feature selection (see [69]) before PSA application, and also take into account the effects of dimensionality when designing simulation experiments or applications.

In general, our findings support the view given in [66] that ML methods can usefully be used to remove selection bias when dealing with non-probability samples. Prior research has shown that PSA successfully removes bias in some situations but at the cost of increasing the variance of the estimates ([10]; [11]). The technique proposed by [11] and [12], applying a combination of PSA and calibration, may represent a good alternative in such situations. The behaviour of ML methods when both PSA and calibration are applied is currently under study.

## Supporting information

**S1 Table. Summary statistics of the simulated population and the subpopulations with and without internet access.**

**S2 Table. Optimal parameters for each algorithm given a volunteer sample size and a group of covariates in the real data simulation, obtained with a 10 times repeated 10-fold cross-validation.**

**S3 Table. Bias on the estimation of vote intention with unequal selection probabilities for the convenience sample based on SRSWOR from the internet population.**

**S4 Table. Mean Square Error (MSE) in the estimation of vote intention with unequal selection probabilities for the convenience sample based on SRSWOR from the internet population.**

**S5 Table. Bias in the estimation of vote intention with unequal selection probabilities for the convenience sample based on the logistic formula.**

**S6 Table. Mean Square Error (MSE) in the estimation of vote intention with unequal selection probabilities for the convenience sample based on the logistic formula.**

**S7 Table. Bias in the estimation of vote intention with unequal selection probabilities for the convenience sample based on the logistic formula with a sine transformation.**

**S8 Table. Mean Square Error (MSE) in the estimation of vote intention with unequal selection probabilities for the convenience sample based on the logistic formula with a sine transformation.**

**S9 Table. Mean and median of bias (absolute values) and MSE of estimates using PSA for each algorithm, and number of times its estimates are among the best (absolute bias or MSE less than 1% greater than the minimum value).**

**S10 Table. Bias and MSE prevalence estimates of self-reported "poor health" status after reweighting with PSA using classification algorithms.**

**S11 Table. Bias and MSE of the estimates of the fraction of households with more than two members after reweighting with PSA using classification algorithms.**

**S1 Data. Datasets used in the simulation study.**

## Acknowledgments

This study was partially supported by Ministerio de Economía y Competitividad, Spain [grant number MTM2015-63609-R] and, in terms of the first author, a FPU grant from the Ministerio de Ciencia, Innovación y Universidades, Spain (FPU17/02177).

# References

[1] Elliott MR and Valliant R. Inference for Nonprobability Samples. *Stat Sci* 2017; 32(2):249-264.

[2] Schonlau M and Couper M. Options for Conducting Web Surveys. *Stat Sci* 2017; 32(2):279-292.

[3] Couper M, Kapteyn A, Schonlau M and Winter J. Noncoverage and Nonresponse in an internet Survey. *Soc Sci Res* 2007; 36:131-148.

[4] National Institute of Statistics. Survey on Equipment and Use of Information and Communication Technologies in Households. 2017.

[5] Bethlehem J. Selection Bias in Web Surveys. *Int Stat Rev* 2010; 78(2):161-188.

[6] Dever JA, Rafferty A and Valliant R. Internet surveys: Can statistical adjustments eliminate coverage bias?. *Surv Res Methods* 2008; 2(2):47-62.

[7] Rosenbaum PR and Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; 70(1):41-55.

[8] Taylor H. Does internet research work? *Int J Market Res* 2000; 42(1):51-63.

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

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

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

[12] Valliant R and Dever JA. Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociol Method Res* 2011; 40(1):105-137.

[13] Agresti A. *An introduction to categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons, 2007.

[14] D'Agostino, RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; 17:2265-2281.

[15] Westreich D, Lessler J and Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol* 2010; 63:826-833.

[16] Breiman L, Friedman J, Olshen R and Stone C. *Classification and regression trees*. Belmont, California: Wadsworth, 1984.

[17] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ and Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 2008; 17(6):546-555.

[18] Lee, BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2010; 29(3):337-346.

[19] Linden A and Yarnold PR. Using classification tree analysis to generate propensity score weights. *J Eval Clin Pract* 2017; 23(4):703-712.

[20] Breiman L. Random forests. *Mach Learn* 2001; 45(1):5-32.

[21] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 1189-1232.

[22] McCaffrey DF, Ridgeway G and Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 2004; 9(4):403-425.

[23] McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R and Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013; 32(19):3388-3414.

[24] Austin PC. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivar Behav Res* 2012; 47(1):115-135.

[25] Watkins S, Jonsson Funk M, Brookhart MA, Rosenberg SA, O'Shea TM and Daniels J. An Empirical Comparison of Tree Based Methods for Propensity Score Estimation. *Health Serv Res* 2013; 48(5):1798-1817.

[26] Harder VS, Morral AR and Arkes J. Marijuana use and depression among adults: Testing for causal associations. *Addiction* 2006; 101(10):1463-1472.

[27] Harder VS, Stuart EA and Anthony JC. Adolescent cannabis problems and young adult depression: male-female stratified propensity score analyses. *Am J Epidemiol* 2008; 168(6):592-601.

[28] Harder VS, Stuart EA and Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods* 2010, 15(3):234.

[29] Wyss R, Ellis AR, Brookhart MA, Girman CJ, Jonsson Funk M, LoCasale R and Stürmer T. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am J Epidemiol* 2014; 180(6):645-655.

[30] Zhao P, Su X, Ge T and Fan J. Propensity score and proximity matching using random forest. *Contemp Clin Trials* 2016; 47:85-92.

[31] Pirracchio R, Petersen ML and Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2014; 181(2):108-119.

[32] Van Der Laan MJ, Polley EC and Hubbard AE. Super Learner. *Stat Appl Genet Mo B* 2007; 6(1):1-21.

[33] Montanari GE and Ranalli MG. Multiple and ridge model calibration. In: *Proceedings of Workshop on Calibration and Estimation in Surveys*, Ottawa, Canada, October 31-November 1 2007.

[34] Baffetta F, Fattorini L, Franceschi S and 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(3):463-475.

[35] Tipton J, Opsomer J and Moisen G. Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sens Environ* 2013; 139:130-137.

[36] Wang JC, Opsomer JD and Wang H. Bagging non-differentiable estimators in complex surveys. *Surv Methodol* 2014; 40:189-209.

[37] Breidt J and Opsomer J. Model-assisted survey estimation with modern prediction. *Stat Sci* 2017; 32(2):190-205.

[38] David M, Little RJA, Samuhel ME and Triest RK. Nonrandom nonresponse models based on the propensity to respond. In: *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 1983, pp. 168-173.

[39] Da Silva, DN and Opsomer, JD. Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Surv Methodol* 2009; 35(2):165-176.

[40] Da Silva, DN and Opsomer, JD. A kernel smoothing method of adjusting for unit non response in sample surveys. *Can J Stat* 2006; 34(4):563-579.

[41] Phipps P and Toth D. Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Ann Appl Stat* 2012; 6(2):772-794.

[42] Buskirk TD and Kolenikov S. 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* 2015; 17.

[43] Valliant R and Dever JA, Kreuter, F. *Practical tools for designing and weighting survey samples*. New York: Springer, 2013.

[44] Cochran, WG. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics* 1968; 24(2):295-313.

[45] Ferri-García R and Rueda MM. Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Stat Oper Res T* 2018; 1(2):159-182.

[46] National Institute of Statistics. España en Cifras 2017. 2017b.

[47] National Institute of Statistics. Population (Spaniards/Foreigners) by communities, age (five years groups), sex and year. 2016.

[48] National Institute of Statistics. Educational level of the adult population by age groups. CNED-2014. 2017c.

[49] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ and Steinberg D. Top 10 algorithms in data mining. *Knowl Inf Syst* 2008; 14(1):1-37.

[50] Quinlan JR. *C4.5: programs for machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1993.

[51] National Institute of Statistics. Life Conditions Survey. Microdata. 2012.

[52] National Institute of Statistics. Survey on Equipment and Use of Information and Communication Technologies in Households. Microdata. 2012b.

[53] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2018, accessed 9 September 2019).

[54] Hornik K, Buchta C and Zeileis A. Open-Source Machine Learning: R Meets Weka. *Comp Stat* 2009; 24(2):225-232. doi: 10.1007/s00180-008-0119-7

[55] Witten IH and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[56] Kuhn M and Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models. R package version 0.1.2. https://CRAN.R-project.org/package=C50 (2018, accessed 9 September 2019).

[57] Therneau T and Atkinson B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart (2018, accessed 9 September 2019).

[58] Liaw A and Wiener M. Classification and Regression by randomForest. *R News* 2002; 2(3):18-22.

[59] Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0. https://CRAN.R-project.org/package=e1071 (2018, accessed 9 September 2019).

[60] Greenwell B, Boehmke B, Cunningham J and GBM Developers. gbm: Generalized Boosted Regression Models. R package version 2.1.4. https://CRAN.R-project.org/package=gbm (2018, accessed 9 September 2019)

[61] Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.

[62] Dragulescu AA and Arendt C. xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.1. https://CRAN.R-project.org/package=xlsx (2018, accessed 9 September 2019)

[63] Auguie B. gridExtra: Miscellaneous Functions for ”Grid” Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra (2017, accessed 9 September 2019)

[64] Neuwirth E. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer (2014, accessed 9 September 2019).

[65] Kuhn M. caret: Classification and Regression Training. R package version 6.0-81. https://CRAN.R-project.org/package=caret (2018, accessed 9 September 2019).

[66] Buelens B, Burger J and van den Brakel JA. Comparing Inference Methods for Non-probability Samples. *Int Stat Rev* 2018; 86(2):322–343.

[67] García S, Luengo J and Herrera F. *Data preprocessing in data mining*. Switzerland: Springer International Publishing, 2015.

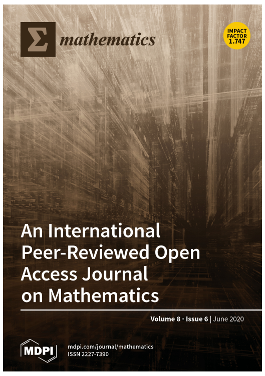
[68] Beyer K, Goldstein J, Ramakrishnan R and Shaft U. When is “nearest neighbor” meaningful?. In: *International conference on database theory*, Jerusalem, Israel, January 10-12, 1999, pp. 217-235. Berlin, Heidelberg: Springer.

[69] Fayyad U, Piatetsky-Shapiro G and Smyth P. From data mining to knowledge discovery in databases. *AI mag* 1996; 17(3):37-37.

# Appendix A3

## Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

Castro-Martín, Luis; Rueda, María del Mar; Ferri-García, Ramón (2020)
Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment
*Mathematics*, vol. 8, number 11, p. 2096
DOI: 10.3390/math8112096


mathematics
IMPACT FACTOR 1.747
An International Peer-Reviewed Open Access Journal on Mathematics
Volume 8 · Issue 6 | June 2020
MDPI
mdpi.com/journal/mathematics
ISSN 2227-7390

| MATHEMATICS | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 1.747 | 28/325 | Q1 |

**Abstract**

This study introduces a general framework on inference for a general parameter using nonprobability survey data when a probability sample with auxiliary variables, common to both samples, is available. The proposed framework covers parameters from inequality measures and distribution function estimates but the scope of the paper is broader. We develop a rigorous framework for general parameter estimation by solving survey weighted estimating equations which involve propensity score estimation for units in the non-probability sample. This development includes the expression of the variance estimator, as well as some alternatives which are discussed under the proposed framework. We carried a simulation study using data from a real-world survey, on which the application of the estimation methods showed the effectiveness of the proposed design-based inference on several general parameters.

# 1 Introduction

Nonprobability samples are increasingly common in empirical sciences. The rise of online and smartphone surveys, along with the decrease of response rates in traditional survey modes, have contributed to the popularization of volunteer surveys where sampling is non-probabilistic. Moreover, the development of Big Data involves the analysis of large scale datasets whose obtention is conditioned by data availability and not by a probabilistic selection, and therefore they can be considered large nonprobability samples of a population [1].

The lack of a probability sampling scheme can be responsible for selection bias. Following the description from [1, 2], we can distinguish the target population, $U_T$, the subpopulation that a given selection method can potentially cover, $U_{pc}$, and the fraction of the subpopulation that is finally covered, $U_{fc}$, and whose individuals might participate in the survey. Selection bias occurs when the characteristics of the individuals in $U_{fc}$ differ significantly from those in $U_T$ in a way that could affect final estimates. Typically, differences between individuals in $U_T$ and individuals in $U_{pc}$ are caused by a lack of coverage induced by the survey administration mode (for example, an online questionnaire cannot be administered to the population without internet access), while differences between $U_{pc}$ and $U_{fc}$ are caused by the variability in the propensities to participate between social-demographic groups (for example, an online questionnaire accesible in a thematic website might only be fulfilled by visitors of the website who have a specific interests that could influence the results).

Following the rise of nonprobability samples, a class of methods for reducing selection bias have been proposed in the last decades. These methods were developed from different perspectives according to the availability of auxiliary information. We can mention calibration, Propensity Score Adjustment (PSA), Statistical Matching and superpopulation modelling as the most relevant techniques to mitigate selection bias produced by coverage and self-selection errors.

Calibration weighting was originally developed by [3] as a method to correct representation issues in samples with coverage or non-response errors. It only requires a vector of auxiliary variables available for each individual of the sample and the population totals of those variables. Calibration is able to remove selection bias in nonprobability samples if the selection mechanism is ignorable [4], and despite being originally developed for parametric estimation, further work [5, 6, 7] has extended calibration to distribution function, quantile and poverty measures estimation.

Propensity Score Adjustment (PSA) and Statistical Matching require, apart from the nonprobability sample, a probability sample to do the adjustments. PSA was originally developed for balancing groups in non-randomized clinical trials [8] and it was adapted for non-response adjustments shortly after [9, 10]. The application of PSA for removing bias in nonprobability surveys was theoretically developed in [11, 12]. Statistical Matching was firstly proposed in [13] and extended in [14] for non-response adjustments. The difference between both methods is the sample used in the estimators: PSA estimates the propensity of each individual of the nonprobability sample to participate in the survey and then this propensity is used to construct the weights of the estimators, while Statistical Matching adjusts a prediction model using data from the nonprobability sample, applies it in the probability sample to predict their values for the target variable $y$ and uses them in the parametric estimators. To the best of our knowledge, PSA and Statistical Matching has not been developed for nonparametric estimation.

Superpopulation modelling requires data from the complete census of the target population for the covariates used in the adjustment, which is assumed to be a realization (sample) of a superpopulation where the (unknown) target values follow a model. It is based on the works by [15, 16], where the main idea is to fit a regression model on the target variable with data from the nonprobability sample, and use the model to predict the values of the target variable for each individual in the population. The prediction can be used for estimation using a model-based approach or some alternative versions such as model-assisted and model-calibrated. LASSO models [17] and Machine Learning predictors [18, 19] have been studied as alternatives to ordinary least squares regression in superpopulation modelling.

The interest of society on poverty and inequality has increased in the last decades given the successive economic cycles and crisis. In such a context, official poverty rates and the percentage of people in poverty (or under a poverty threshold) are some important measures of a country's wealth. The common characteristic of many poverty measures is their complexity. The literature on survey sampling is usually focused on the goal of estimating linear parameters. However, it is usual that the variable of interest in poverty studies is a measure of wages or income, where the distribution function becomes a relevant tool because it is required to calculate the proportion of people with low income, the poverty gap and other measures. Estimators for the cumulative distribution function, quantiles [21, 20] and poverty measures [22] can be found in literature regarding probability samples, but there is hardly any work on the estimation of these parameters when the samples are obtained from volunteers.

In this paper, we aim to develop a framework for statistical inference on a general parameter with non probability survey samples when a reference probability sample is available. After introducing the problem of the mean estimation for volunteer samples in Section 2, in Section 3, we consider the problem of the estimation for a general parameter through general estimating equations. Section 4 presents a new estimator for a general parameter through the use of PSA to estimate the propensity score of each individual in the survey weighted estimating equation and major theoretical results are presented. Results from simulation studies are reported in Sections 5 and 6 presents the concluding remarks.

## 2 Approaches to Estimation of a Mean for Volunteer Online Samples

Let $U_T$ be the target population with $N$ elements and $s_v$ a nonprobability sample drawn from a subset of $U_T$, $U_v$, with a size of $n_v \leq N$. Let $y$ be the target variable of the survey, whose mean in the population $U_T$ is denoted as $\overline{Y}$. The sample estimation of $\overline{Y}$, $\hat{\overline{Y}}$, is done using the Horvitz-Thompson estimator:

$$\hat{\overline{Y}}_{HT} = \frac{\sum_{i \in s_v} w_i y_i}{\sum_{i \in s_v} w_i} \tag{1}$$

where $w$ is a vector of weights that accounts for the lack of representativity of $s_v$ caused by selection bias. If no auxiliary information is given, the weight would be the same for every unit, $w_i = N/n_v$, which requires to assume that the sample was drawn under a simple random sampling scheme. This is a naïve assumption given that $s_v$ is not probabilistic, this is, the probability of being in the sample is unknown and/or null for any of the units in $U_T$.

Let $\mathbf{x}$ be a matrix of covariates measured in $s_v$ along with $y$. If the population totals of the covariates, $\mathbf{X}$, are available, it is possible to estimate the mean using a vector of weights obtained with calibration, $w^{CAL}$. The calibration weights aim to minimize the distance between the original and the new weights

$$\underset{w_i^{CAL}}{minE} \left[ \sum_{i \in s_v} G(w_i, w_i^{CAL}) \right] \tag{2}$$

while respecting the calibration equations

$$\sum_{i \in s_v} w_i^{CAL} \mathbf{x}_i = \mathbf{X}. \tag{3}$$

Some choices for the distance $G(.,.)$ were listed in [3], along with the resulting estimators. Calibration weighting for selection bias treatment was studied in [4], where post-stratification, which is a special case of calibration [23], was used to mitigate the bias caused by different selection mechanisms, showing its efficacy when the selection of the units of $s_v$ is Missing At Random (MAR).

If a reference sample, $s_r$, drawn from the population $U_T$ is available and a number of covariates $\mathbf{x}$ have been measured both in $s_v$ and $s_r$, two procedures can be done to reduce selection bias present in $s_v$. Let $I_v$ be an indicator variable of an element being in $s_v$, this is

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases} \tag{4}$$

Propensity Score Adjustment (PSA) assumes that each element of $U_T$ has a probability (propensity) of being selected for $s_v$ which can be formulated as

$$\pi_i^v = Pr(I_{vi} = 1|\mathbf{x}_i, y_i) \tag{5}$$

where $\pi_i^v$ is the propensity of the $i$-th individual to participate in $s_v$. The random mechanism behind this probability is the selection mechanism that governs the nonprobability sample. If the selection is Missing Completely At Random (MCAR), then $\pi_i^v = Pr(I_{vi} = 1)$ and the selection bias is null, while if the selection is MAR then $\pi_i^v = Pr(I_{vi} = 1|\mathbf{x_i})$ and the selection mechanism is considered ignorable. This does not mean that the selection bias should be ignored but rather it can be treated with the right techniques.

In PSA, we consider the situation where true propensities are not known and therefore have to be estimated; we do it by combining $s_v$ and $s_r$ into a sample. The probability that $I_v = 1$ is then estimated using a prediction model, traditionally a logistic regression one:

$$\hat{\pi}_i^v = \frac{1}{1 + exp\{-\beta \mathbf{x_i}\}} \tag{6}$$

Alternative models, such as non-linear regression and Machine Learning classification algorithms, have been studied in literature as a substitute of logistic regression (see [24] for a review). The resulting propensities can be used to adjust new weights, $w^{PSA}$, with different alternatives:

- A simple inverse probability weighting is proposed by [25]

  $$w_i^{PSA1} = \frac{w_i}{\hat{\pi}_i^v} \tag{7}$$

  which is a similar approach to the formula used in [26]

  $$w_i^{PSA2} = \frac{1 - \hat{\pi}_i^v}{\hat{\pi}_i^v} \tag{8}$$

- Alternatively, individuals of the combined sample ($s_v \cup s_r$) can be grouped in $g$ equally-sized strata of similar propensity scores from which an average propensity is calculated for each group. Let $\overline{\pi}_g$ be the mean propensities of the $g$-th strata. [2] use the means as in (7) to calculate the new weights:

  $$w_i^{PSA3} = \frac{w_i}{\overline{\pi}_{g_i}} \tag{9}$$

  where $g_i$ refers to the strata to which the $i$-th individual of $s_v$ belongs.

- A similar approach can be found in [12], but instead of using the means, a factor is calculated for each strata:
$$f_g = \frac{\sum_{k\in s_{r_g}} \tilde{w}_k / \sum_{k\in s_r} \tilde{w}_k}{\sum_{i\in s_{v_g}} w_i / \sum_{i\in s_v} w_i} \tag{10}$$
where $s_{r_g}$ and $s_{v_g}$ are respectively the individuals from the probability and nonprobability sample that belong to the $g$-th strata, and $\tilde{w}$ is the vector of design weigths of the reference sample. The final weights are obtained by multiplying the original weights and the correction factor:
$$w_i^{PSA4} = w_i \cdot f_{g_i} \tag{11}$$

PSA has been proven to successfully remove selection bias when prognostic covariates are chosen [11] and further adjustments, such as calibration, are applied in the estimations [12, 2, 27]. A recent paper [28] shows a real application of PSA in web panel surveys where the reductions in bias, although present, were not large enough to consider the estimates as unbiased.

As an alternative to PSA, Statistical Matching is another method to mitigate selection bias when a reference sample is available. For the matter, a prediction model for $y$ using $\mathbf{x}$ as the dependent variables is built using data from $s_v$. The model is subsequently applied on the reference sample to obtain the estimates from the predicted values of $y$ in $s_r$, $\hat{y}$:
$$\hat{\hat{Y}} = \sum_{k\in s_r} w_k \hat{y}_k \tag{12}$$

The choice of prediction models has been studied in literature; the usual method is linear regression but other approaches such as donor imputation [13] or Machine Learning algorithms [19, 29] have been listed as alternatives. Under certain conditions, Statistical Matching can reduce bias and mean square error to a greater extent than PSA [29].

When a complete census of the entire target population is available, with information on the covariates present in $s_v$, superpopulation modelling can be applied to remove selection bias [19]. In this paper we consider the case when auxiliary information is available only from a reference probability survey.

## 3 Estimation of a General Parameter by Using PSA

Let $y$ be the variable of interest in a survey and $y_i$ be the value of the $i$-th unit in that variable, $i = 1,\ldots,N$. Suppose we want to estimate a finite population parameter $\theta_N$ of dimension $p \geq 1$ defined as the solution of the census estimating equations:
$$U(\theta_N) = \frac{1}{N}\sum_U u_i(y_i, \theta_N) = \mathbf{0} \tag{13}$$
where $u_i(y_i, \theta_N)$ is be a function of $\theta_N$. Some unidimensional parameters of interest can be:

- the population total $T_y$ for $u_i = (y_i - \theta_N/N)$,
- the population mean $\bar{Y}$ for $u_i = (y_i - \theta_N)$,
- the population distribution function $F_y(t)$ for $u_i = (1(y_i \leq t) - \theta_N)$ with $1(\cdot)$ being the indicator function,
- the finite population quantile of order $j$, $Q_j$ for $u_i = (1(y_i \leq \theta_N) - j$, where $0 < j < 1$,

We denote by $\hat{\theta}$ the solution of the equation:

$$\hat{U}(\theta_N) = \sum_U I_{vi} u_i(y_i, \theta_N)/\pi_i^v = \sum_{s_v} u_i(y_i, \theta_N)/\pi_i^v = \mathbf{0}. \tag{14}$$

It is clear the $E_r(\hat{U}(\theta_N)) = U(\theta_N)$ where $r$ stands for the model of the selection mechanism for the sample $s_v$, this is, the true model that fits propensity scores. If $\pi_i^v$ are known we can get the consistent estimator of $\theta_N$ by solving the equation above. For the study of the properties of this estimator we consider a quasi-probability approach or pseudo-design-based inference ([19]) and we treat the volunteer sample as a realization of a Poisson sampling with probabilities $\pi_i^v$.

For any sample design that verifies certain regularity conditions, the solution to $\widehat{U}(\theta) = \mathbf{0}$ provides a consistent estimator for the parameter $\theta_N$ (see [30]). Poisson sampling verifies these conditions, so that the consistency of the estimator is obtained immediately from the result of [30]. The normality of the estimator is demonstrated by [31], who also obtains the asymptotic variance of the estimator. From said expression and taking into account that in Poisson sampling the extractions are independent and therefore the probability of second order is given by $\pi_{ij}^v = \pi_i^v \pi_j^v$ we can obtain the variance of $\hat{\theta}$:

$$V(\hat{\theta}) = J(\hat{\theta})^{-1} var(\hat{U}(\theta)) J'(\hat{\theta})^{-1} \tag{15}$$

being $J(\theta) = \frac{1}{N}\sum_U \partial u_i/\partial\theta$ and $var(\hat{U}(\theta)) = \sum_U (1 - \pi_i^v) u_i^2/(\pi_i^v)^2$

## 4 Estimation of a General Parameter with Estimated Propensities

The propensity scores $\pi_i^v$ are not known are impossible to estimate using the non-probability sample $s_v$ alone, so additional information must be included. Let $s_r$ be a reference probability sample, of size $n_r$, selected from $U_T$ under a sampling design $(s_d, p_d)$ where the first order inclusion probabilities, $\pi_i^p = \sum_{s_r \ni i} p_d(s_r), i = 1, \ldots, n_r$, are known and non-null.

The covariates of the propensity model $\mathbf{x}$ have been measured both in $s_v$ and $s_r$, while the variable of interest $y$ is only available for those individuals in $s_v$.

Suppose that the propensity scores can be modelled parametrically as

$$\pi_i^v = P(I_{vi} = 1/\mathbf{x}_i) = m(\lambda_o, \mathbf{x}_i) \quad i = 1, \ldots, N \tag{16}$$

for some known function $m(\cdot)$ with second continuous derivatives with respect to an unknown parameter $\lambda_o$.

We estimate the propensity scores by using data of both the volunteer and the probability sample. The maximum likelihood estimator (MLE) of $\pi_i^v$ is $m(\hat{\lambda},\mathbf{x}_i)$ where $\hat{\lambda}$ corresponds to the value of lambda that maximizes the log-likelihood function:

$$l(\lambda) = \sum_U (I_{vi} log(m(\lambda,\mathbf{x}_i)) + (1-I_{vi}) log(1-m(\lambda,\mathbf{x}_i)) =$$

$$\sum_{s_v} log\frac{m(\lambda,\mathbf{x}_i)}{1-m(\lambda,\mathbf{x}_i)} + \sum_U log(1-m(\lambda,\mathbf{x}_i)). \tag{17}$$

As it is usual in survey sampling, we consider the pseudo-likelihood given that some units of the population have not been sampled:

$$\tilde{l}(\lambda) = \sum_{s_v} log\frac{m(\lambda,\mathbf{x}_i)}{1-m(\lambda,\mathbf{x}_i)} + \sum_{s_p} \frac{1}{\pi_i^p} log(1-m(\lambda,\mathbf{x}_i)). \tag{18}$$

We propose thus a two phase procedure in this manner:

Step 1: Calculate $\hat{\lambda}_{pl}$ by solving the score equations:

$$\partial\tilde{l}(\mathbf{x}_i,\lambda)/\partial\lambda = 0$$

Step 2: Calculate $\hat{\theta}_v$ as the solution of the estimating function:

$$\hat{U}_V(\theta) = \sum_U I_{vi} u_i(y_i,\theta)\frac{1}{m(\hat{\lambda}_{pl},\mathbf{x}_i)} = 0 \tag{19}$$

We consider the following asymptotic framework for theoretical development, which is equivalent to the framework in [32]. Let $U_{T\nu}$ be a sequence of finite populations of size $N_\nu$. Each $U_{T\nu}$ has an associated non-probability sample $s_{v\nu}$ of size $n_{v\nu}$ and an associated probability sample $s_{p\nu}$ of size $n_{p\nu}$. We consider that the population size $N_\nu \to \infty$, the nonprobability sample size $n_{v\nu} \to \infty$ and the probability sample size $n_{p\nu} \to \infty$ as $\nu \to \infty$. For notational simplicity the index $\nu$ is suppressed for the rest of the paper. The properties of the estimator $\hat{\theta}_v$ are developed under both the model for the propensity scores and the survey design for the probability sample.

We make the following assumptions:

- A.1. The estimating function $u_i(y_i,\theta,\lambda)$ is twice differentiable with respect to $\theta$ and $\lambda$.
- A.2. The propensities and the sampling design ensure that $\hat{U}_V(\theta) - U(\theta) = O_p(n^{-1/2})$ for any $\theta \in \Theta$.

- A.3. The propensities and the sampling design ensure that $\hat{U}_V(\theta)$ is asymptotically Normal with mean $\mathbf{U}(\theta)$ and entries of the variance at the order $O(n^{-1})$ for any fixed $\theta \in \Theta$.

**Theorem 1**. Under the conditions A.1, A.2 and A.3, $\hat{\theta}_v$ is a consistent and asymptotically normal estimator for $\theta$.

*Proof.* Under assumed conditions,

$\hat{U}_V(\theta) = U(\theta) + O_p(n^{-1/2})$, thus by using the mean value theorem, $\hat{\theta}_v$ has the same asymptotic behaviour that $\hat{\theta}$ which is consistent for $\theta$ and asymptotically normal distributed (see Section 3). □

Variance estimation for $\hat{\theta}_v$ can be handled by combining the two estimating equations, $\tilde{l}$ and $\hat{U}_v$, into a single system as it is done in [33].

The MLE of $\lambda$, $\hat{\lambda}_{pl}$ is the solution to the equations:

$$U_2(\lambda) = \sum_{s_v} \partial log \frac{m(\lambda,\mathbf{x}_i)}{1-m(\lambda,\mathbf{x}_i)}/\partial\lambda + \sum_{s_p} \partial \frac{1}{\pi_i^p} log(1-m(\lambda,\mathbf{x}_i))/\partial\lambda = 0$$

and the PSA estimator of $\theta_N$ is the solution to the estimating equations

$$U_1(\theta,\lambda) = \sum_{s_v} u_i(y_i,\theta)\frac{1}{m(\lambda_{pl},\mathbf{x}_i)} = \sum_{s_v} g_1(y_i,\mathbf{x}_i,\theta,\lambda) = 0.$$

Let $\mathbf{U}(\theta,\lambda) = (U_1'(\theta,\lambda), U_2'(\lambda))'$. Let $\psi = (\theta_N', \lambda_o)'$ be the true parameter values defined through the census estimating equations and $\hat{\psi} = (\hat{\theta}_N', \hat{\lambda}_o{}')'$ the solutions to $\mathbf{U}(\theta,\lambda) = 0$.

We need an additional assumption:

- A.4. The propensities, the sampling design and the estimating function satisfy $\partial\widehat{\mathbf{U}}/\partial\psi = O_p(1)$ and $\partial^2\widehat{\mathbf{U}}/\partial\psi\partial\psi' = O_p(1)$.

**Theorem 2** Under the conditions A.1, A.2, A.3 and A.4, the asymptotic variance-covariance matrix of $\hat{\psi}$ is given by the expression:

$$\mathbf{V}(\hat{\psi}) = \mathbf{H}^{-1} V(\hat{\mathbf{U}}(\theta,\lambda))\mathbf{H}'^{-1} \tag{20}$$

with $\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix}$

$$H_{11} = E\{\frac{\partial}{\partial\theta}U_1(\theta_N,\lambda)\}$$

$$H_{21} = E\{\frac{\partial}{\partial\lambda}U_1(\theta_N,\lambda)\}$$

$$H_{22} = E\{\frac{\partial}{\partial\lambda}U_2(\lambda)\}$$

*Proof.* Since $\hat{\theta}_v$ and $\hat{\lambda}$ are consistent estimator of respective parameters, we can write $\hat{\psi} = \psi + O_p(1)$ and the Taylor series expansion gives:

$$\hat{\psi} = \psi - \mathbf{H}^{-1}\widehat{\mathbf{U}}(\theta,\lambda) + O_p(\|\hat{\psi} - \psi\|^2),$$

Thus the asymptotic variance of $\hat{\psi}$ is given by:

$$\mathbf{V}(\hat{\psi}) = \mathbf{H}^{-1}V(\widehat{\mathbf{U}}(\theta,\lambda))\mathbf{H}'^{-1}.$$

Taking into account the two random mechanisms, and the probabilities of the conditional expectation, we have $V(\widehat{\mathbf{U}}(\theta,\lambda)) = V_pE_r(\widehat{\mathbf{U}}(\theta,\lambda)) + E_pV_r(\widehat{\mathbf{U}}(\theta,\lambda))$ where $r$ stands for the model of the selection mechanism for the sample $s_v$ and $p$ refers to the probability sampling design for $s_p$. □

The asymptotic variance of $\hat{\psi}$ depends on the probability of selecting the sample $s_p$ under the given sampling design and the selection mechanism described by the propensity model. Plug-in estimators can be used to construct variance estimators for all the required components but it is not a simple issue.

In practice, and as described in [7], the use of jackknife [34] and bootstrap techniques [35] in the variance estimation for nonlinear parameters should be more advantageous because of their wide applicability for different cases and conditions. Direct applications of bootstrap methods for estimating the variance-covariance matrix of $\hat{\psi}$ involve solving the equation $\mathbf{U}(\theta,\lambda) = 0$ repeatedly for each bootstrap sample. Multiplier Bootstrap with Estimating Functions was proposed by [36].

## 5 Simulation Study

### 5.1 Data

Data for the simulation study come from a wave of the Spanish Living Conditions Survey collected between 2011 and 2012 [37], which contains an annual thematic module that, in 2012, was dedicated to household conditions. The survey sampling follows a two-phase cluster sampling, where the primary units are the households and the secondary units are their members. In 2012, the final sample included 33,573 individuals. For this study, the dataset was filtered to rule out individuals and variables with high quantities of missing data. After this procedure, the dataset employed as pseudopopulation of the study had a size of $N$ = 28,210 individuals and $p$ = 60 available variables.

From this pseudopopulation, two probability samples of size $n_r$ were drawn according to the following sampling strategies:

- The first sample, $s_{r1}$, was drawn with a stratified cluster sampling, where the strata were defined by the Autonomous Communities (NUTS2 regions) and the clusters were the households, which were drawn with probabilities proportional to the household size. The number of households to be selected, $m$, was estimated dividing $n_r$ by the medium household size in order to reach the aforementioned size of $n_r = 2000$, resulting in $m = 902$ households. The final sample size of $s_{r1}$ was $n_{r1} = 2003$.

- The second sample, $s_{r2}$, was drawn with an unequal probability sampling, where probabilities were proportional to the minimum income of the individual's household to make ends meet (variable HS130 in [37]).

The extraction of the nonprobability sample, $s_v$, was done with unequal probability sampling from the full pseudopopulation, where the probability of selection for the $i$-th individual, $p_i$, was given by the formula:

$$p_i = \frac{1}{1 + exp(-2x_i^1 + 0.2x_i^2 + 0.01x_i^3 + 0.2x_i^{41} + 0.4x_i^{42})} \tag{21}$$

where

- $x_i^1 = 1$ when the $i$-th sampled individual has a computer at home, and $x_i^1 = 0$ otherwise.
- $x_i^2 = 1$ when the $i$-th sampled individual is a man, and $x_i^2 = 0$ otherwise.
- $x_i^3$ is the age (in years) of the $i$-th sampled individual.
- $x_i^{41} = 1$ when the $i$-th sampled individual lives in a medium population density area, and $x_i^{41} = 0$ otherwise.
- $x_i^{42} = 1$ when the $i$-th sampled individual lives in a low population density area, and $x_i^{42} = 0$ otherwise.

The reasoning behind this sampling procedure is to take into account more similar mechanisms to self-selection procedures that take place in real nonprobability surveys.

We have considered three different sample sizes, $n_v$ = 2000, 4000, 6000. 1000 simulation runs were performed for each procedure and sample size, drawing a sample in each run.

## 5.2 Simulation

In each simulation, the parameters to be estimated were the following:

- The Gini coefficient [38], which measures the income inequality, estimated as

$$\widehat{G}_y = \frac{\sum_{k \in s_v} \frac{1}{\pi_k}(2\widehat{F}_y(y_k) - 1)y_k}{\sum_{k \in s_v} y_k/\pi_k}$$

- The proportion of individuals with a disposable income below the at-risk-at-poverty threshold. This measure can be referred to as poverty incidence, poverty proportion, poverty risk or HCI ([39] and is estimated as

$$H\hat{C}I = \frac{1}{N} \sum_{k \in s_v} \frac{1}{\pi_k} I(y < 0.6Q_{0.5})$$

- The interquartile range, estimated as

$$\widehat{IQR} = \frac{\widehat{Q}_{0.75}}{\widehat{Q}_{0.25}}$$

- The interdecile range, estimated as

$$\widehat{IDR} = \frac{\widehat{Q}_{0.9}}{\widehat{Q}_{0.1}}.$$

Every parameter was estimated with and without applying PSA so we could evaluate its performance. In order to estimate the propensities, a logistic regression model was chosen:

$$m(\hat{\lambda}, \mathbf{x}_i) = \frac{exp(\hat{\lambda}^{\mathsf{T}}\mathbf{x}_i)}{1 + exp(\hat{\lambda}^{\mathsf{T}}\mathbf{x}_i)}$$

1000 simulations were executed for each context. The resulting mean bias, standard deviation and Root Mean Square Error were measured in relative numbers to make them comparable across different scenarios. The formulas used for their calculation can be found below:

$$RBias\ (\%) = \left| \frac{\sum_{i=1}^{1000} \hat{\theta}^{(i)}}{1000} - \theta_N \right| \cdot \frac{100}{\theta_N} \tag{22}$$

$$RStandard\ deviation\ (\%) = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\theta}^{(i)} - \hat{\hat{\theta}})^2}{999}} \cdot \frac{100}{\theta_N} \tag{23}$$

$$RMSE\ (\%) = \sqrt{RBias^2 + RSD^2} \tag{24}$$

with $\hat{\theta}^{(i)}$ the estimation in the $i$-th simulation and $\hat{\hat{\theta}}$ the mean of the 1000 estimations.

### 5.3 Results

The relative mean bias of the estimations can be observed in Tables 1–3. We can observe that PSA reduces the bias in all situations, specially in the estimation of HCI. PSA using the reference sample drawn with probabilities proportional to the income, $s_{r2}$, provided much less biased estimates overall.

Table 1: Relative mean bias (%) of each parameter without applying PSA.

| Size | Gini | HCI | IQR | IDR |
|---|---|---|---|---|
| 6000 | 6.7 | 80.4 | 7.9 | 9.4 |
| 2000 | 3.2 | 93 | 3.8 | 3.1 |
| 4000 | 3.1 | 86 | 3.7 | 3 |
| 6000 | 3 | 79 | 3.5 | 3 |

Table 2: Relative mean bias (%) of each parameter applying PSA with the stratified reference sample.

| Size | Gini | HCI | IQR | IDR |
|---|---|---|---|---|
| 2000 | 1.7 | 3 | 2.5 | 1 |
| 4000 | 2.1 | 3.3 | 2.7 | 1 |
| 6000 | 2.2 | 3.1 | 2.7 | 0.9 |

Table 3: Relative mean bias (%) of each parameter applying PSA with the proportional reference sample.

| Size | Gini | HCI | IQR | IDR |
|---|---|---|---|---|
| 2000 | 0.3 | 1.1 | 0 | 0.2 |
| 4000 | 0.1 | 1.3 | 0.1 | 0.3 |
| 6000 | 0 | 1.1 | 0.1 | 0.5 |

The relative standard deviation of the estimations can be observed in Tables 4–6. The standard deviation remained stable across estimates of Gini coefficient, IQR and IDR, even with small gains for the latter when using the reference sample with probabilities proportional to the minimum income to make ends meet, $s_{r2}$, but increased after applying PSA in the estimation of HCI.

Table 4: Relative standard deviation (%) of each parameter without applying PSA.

| Size | Gini | HCI | IQR | IDR |
|---|---|---|---|---|
| 2000 | 1.6 | 0.2 | 2.2 | 4.2 |
| 4000 | 1.1 | 0.3 | 1.5 | 2.9 |
| 6000 | 0.8 | 0.4 | 1.2 | 2.2 |

Table 5: Relative standard deviation (%) of each parameter applying PSA with the stratified reference sample.

| Size | Gini | HCI | IQR | IDR |
|---|---|---|---|---|
| 2000 | 1.7 | 4.1 | 2.7 | 4 |
| 4000 | 1.1 | 2.8 | 1.8 | 2.6 |
| 6000 | 0.9 | 2.2 | 1.4 | 2 |

Table 6: Relative standard deviation (%) of each parameter applying PSA with the proportional reference sample.

| **Size** | **Gini** | **HCI** | **IQR** | **IDR** |
|---|---|---|---|---|
| 2000 | 1.3 | 3.9 | 2.1 | 3.2 |
| 4000 | 0.9 | 2.8 | 1.5 | 2.2 |
| 6000 | 0.8 | 2.3 | 1.2 | 2.3 |

The relative Root Mean Square Error of the estimations can be observed in Tables 7–9. As a result of the stability of standard deviation and the reduction in bias, the RMSE of the estimates of the four parameters has a similar pattern than the observed for bias. Although RMSE is reduced after applying PSA in all cases, PSA was more efficient when the reference sample was drawn with probabilities proportional to the minimum income to make ends meet, $s_{r2}$.

Table 7: Relative RMSE (%) of each parameter without applying PSA.

| **Size** | **Gini** | **HCI** | **IQR** | **IDR** |
|---|---|---|---|---|
| 2000 | 3.6 | 93 | 4.4 | 5.2 |
| 4000 | 3.3 | 86 | 4 | 4.2 |
| 6000 | 3.1 | 79 | 3.7 | 3.7 |

Table 8: Relative RMSE (%) of each parameter applying PSA with the stratified reference sample.

| **Size** | **Gini** | **HCI** | **IQR** | **IDR** |
|---|---|---|---|---|
| 2000 | 2.4 | 5.1 | 3.7 | 4.2 |
| 4000 | 2.4 | 4.3 | 3.2 | 2.8 |
| 6000 | 2.4 | 3.8 | 3 | 2.2 |

Table 9: Relative RMSE (%) of each parameter applying PSA with the proportional reference sample.

| **Size** | **Gini** | **HCI** | **IQR** | **IDR** |
|---|---|---|---|---|
| 2000 | 1.4 | 4.1 | 2.1 | 3.2 |
| 4000 | 0.9 | 3.1 | 1.5 | 2.2 |
| 6000 | 0.8 | 2.5 | 1.2 | 2.4 |

PSA performance could be deeply affected by the selection mechanisms, which could lead to model misspecifications in propensity estimations. To test limitation and robustness of the proposed approach we have repeated the simulation with different patterns of non-response. The selection procedures can be described as follows:

NP1 Simple Random Sampling Without Replacement (SRSWOR) from the population fraction of individuals with a computer at home, $U_v$.

NP2 The probability of selection for the $i$-th individual, is given by

$$p_i = \frac{1}{1+exp(-2x_i^1+0.2x_i^2+0.01x_i^3+0.2x_i^{41}+0.4x_i^{42})} \tag{25}$$

NP3 The probability of selection for the $i$-th individual, is given by

$$p_i = (x_i^3-1925)^3/(1995-1925)^3 \tag{26}$$

NP4 The probability of selection for the $i$-th individual, is given by

$$p_i = 0.35+0.1*x_i^1-cos((2012-x_i^3)/5)/3 \tag{27}$$

The procedure 1 is a typical case of coverage error (which is a type of selection bias itself [1]). The third scheme represents a cubic relationship between age and the probability of selection, with young people being the individuals with the highest probabilities and decreasing as age increases. The last scheme has two components: one dichotomous and the other cosine-shaped.

Tables 10 and 11 show the results of bias and relative ecm for the HCI parameter, where the selection bias of the unweighted estimator is large.

Table 10: Relative mean bias (%) for estimating HCI without and with applying PSA.

| | **Unadjusted** | **PSA with Stratified Sample** | **PSA with Proportional Sample** |
|---|---|---|---|
| NP1 2000 | 93.5 | 1.7 | 4.5 |
| NP1 4000 | 86.9 | 1.8 | 4.5 |
| NP1 6000 | 80.4 | 1.9 | 4.5 |
| NP2 2000 | 93 | 3 | 1.1 |
| NP2 4000 | 86 | 3.3 | 1.3 |
| NP2 6000 | 79 | 3.1 | 1.1 |
| NP3 2000 | 92.9 | 1.3 | 1.3 |
| NP3 4000 | 85.8 | 0.1 | 0.2 |
| NP3 6000 | 78.7 | 0.5 | 0.5 |
| NP4 2000 | 92.8 | 3 | 1.4 |
| NP4 4000 | 85.5 | 3.2 | 1.5 |
| NP4 6000 | 78.3 | 3.2 | 1.4 |

Table 11: Relative RMSE(%) for estimating HCI without and with applying PSA.

| | **Unadjusted** | **PSA with Stratified Sample** | **PSA with Proportional Sample** |
|---|---|---|---|
| NP1 2000 | 93.5 | 3.6 | 5.3 |
| NP1 4000 | 86.9 | 2.8 | 4.9 |
| NP1 6000 | 80.4 | 2.5 | 4.7 |
| NP2 2000 | 93 | 5.1 | 4.1 |
| NP2 4000 | 86 | 4.3 | 3.1 |
| NP2 6000 | 79 | 3.8 | 2.5 |
| NP3 2000 | 92.9 | 9.6 | 8.6 |
| NP3 4000 | 85.8 | 6.4 | 5.6 |
| NP3 6000 | 78.7 | 5 | 4.3 |
| NP4 2000 | 92.8 | 4.7 | 4 |
| NP4 4000 | 85.5 | 4.1 | 3.1 |
| NP4 6000 | 78.3 | 3.6 | 2.4 |

The results show a large decrease in bias and MSE for all response patterns for both PSA methods, which shows the robustness of the adjustment method. The reduction in bias and MSE is different across them. Using PSA with the reference sample drawn under a stratified design, $s_{r1}$, provided less RMSE when the convenience sample was drawn using NP1. On the other hand, PSA using the reference sample drawn with probabilities proportional to the income, $s_{r2}$ provided much less biased estimates overall when the selection mechanism depended on NP2, NP3 or NP4.

## 6 Conclusions

Technological development has made large amounts of inexpensive data (commonly known as Big Data) available for researchers to be used for inference. New survey administration methods have also favoured the rise of data from nonprobability samples. Inferences from Big Data and nonprobability surveys have important sources of error ([4, 28, 24], ...). Given the characteristics of these data collection procedures, selection bias is particularly relevant.

Despite the growing interest raised by nonprobability data (both coming from Big Data or nonprobability surveys), there is still a lack of rigorous theory to make statistical inferences for general parameters through estimating equations. The current paper aims to fill this gap by establishing a theoretical framework for estimation of general parameters with nonprobability samples.

Results observed in our simulation study provide strong evidence on the efficiency of methods based in estimating equations with estimated propensities. However, it must be noted that the efficiency depends on the selection mechanisms of nonprobability samples and the availability of covariates for propensity estimation. In our simulations, results showed that Propensity Score Adjustment is more efficient when the propensity of being in the nonprobability sample is less related to the variable of interest. This behavior has been observed in literature regarding PSA for parametric estimation [11, 24].

We used parametric methods to obtain the estimated propensities but we could use machine learning techniques as regression trees, spline regression, random forests etc. Recently [29, 24] presented simulation studies where decision trees, k-nearest neighbors, Naive Bayes, Random Forest, Gradient Boosting Machine and Model Averaged Neural Networks are used for propensity score estimation. These studies compare the empirical efficiency of the use of linear models and Machine Learning prediction algorithms in estimation of linear parameters, but the theory is more complex and has not yet been developed. Other way to reduce the bias of the PSA estimates is to combine the PSA technique with other techniques as Statistical Matching or calibration. [27] apply a combination of propensity score adjustment and calibration on auxiliary variables in a real volunteer survey aimed to a population for which a complete census was available. [32] propose a doubly robust estimator for population mean estimation by incorporating the model-based estimator framework to PSA methods, improving their efficiency and making it robust to model misspecifications. Further research should focus on extensions of those methods for general parameter estimation.

**Author Contributions:** The authors contributed equally to this work in conceptualization, methodology, software and original draft preparation. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Ministerio de Educación y Ciencia (grant MTM2015-63609-R) and by Ministerio de Ciencia e Innovación (grant PID2019-106861RB-I00-AEI- 10.13039/501100011033). The first author received a FPU grant from the Ministerio de Ciencia, Innovación y Universidades, Spain.

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

## References

[1] Elliott, M.R.; Valliant, R. Inference for Nonprobability Samples. *Stat. Sci.* **2017**, *32*, 249–264.

[2] Valliant, R.; Dever, J.A. Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociol. Method Res.* **2011**, *40*, 105–137.

[3] Deville, J.C.; Särndal, C.E. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382.

[4] Bethlehem, J. Selection Bias in Web Surveys. *Int. Stat. Rev.* **2010**, *78*, 161–188.

[5] Martínez, S.; Rueda, M.; Arcos, A.; Martínez, H. Optimum calibration points estimating distribution functions. *J. Comput. Appl. Math.* **2010**, *233*, 2265–2277.

[6] Martínez, S.; Rueda, M.; Martínez, H.; Arcos, A. Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *J. Comput. Appl. Math.* **2017**, *318*, 444–459.

[7] Martínez, S.; Rueda, M.; Illescas, M. The optimization problem of quantile and poverty measures estimation based on calibration. *J. Comput. Appl. Math.* **2020**, https://doi.org/10.1016/j.cam.2020.113054

[8] Rosenbaum, P.R.; Rubin, D.B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **1983**, *70*, 41–55.

[9] David, M.; Little, R.J.A.; Samuhel, M.E.; Triest, R.K. Nonrandom nonresponse models based on the propensity to respond. In Proceedings of the Business and Economic Statistics Section, American Statistical Association, Toronto, Canada, August 15-18, 1983; 168–173.

[10] Little, R.J. Survey nonresponse adjustments for estimates of means. *Int. Stat. Rev. Int. Stat.* **1986**, *54*, 139–157.

[11] Lee, S. Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J. Off. Stat.* **2006**, *22*, 329–349.

[12] Lee, S.; Valliant, R. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociol. Method. Res.* **2009**, *37*, 319–343.

[13] Rivers, D. Sampling for Web Surveys. In *Presented in Joint Statistical Meetings*; Stanford University and Polimetrix, Inc.: Salt Lake City, UT, USA, 2007.

[14] Beaumont, J.F.; Bissonnette, J. Variance estimation under composite imputation: The methodology behind SEVANI. *Surv. Methodol.* **2011**, *37*, 171–179. Available online: https://es.overleaf.com/project/5eb2de68d45d5000014608e2 (accessed on 19 November 2020).

[15] Hartley, H.O.; Sielken, R.L., Jr. A "super-population viewpoint" for finite population sampling. *Biometrics* **1975**, *31*, 411–422.

[16] Royall, R.M.; Herson, J. Robust estimation in finite populations I. *J. Am. Stat. Assoc.* **1973**, *68*, 880–889.

[17] 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.

[18] Breidt, J.; Opsomer, J. Model-assisted survey estimation with modern prediction. *Stat. Sci.* **2017**, *32*, 190–205.

[19] Buelens, B.; Burger, J.; van den Brakel, J.A. Comparing Inference Methods for Non-probability Samples. *Int. Stat. Rev.* **2018**, *86*, 322–343.

[20] Buskirk, T.D.; Lohr, S.L. Asymptotic properties of kernel density estimation with complex survey data. *J. Stat. Plan. Inference* **2005**, *128*, 165–190.

[21] Francisco, C.A.; Fuller, W.A. Quantile estimation with a complex survey design. *Ann. Stat.* **1991**, *19*, 454–469.

[22] Conti, P.L.; Di Iorio, A.; Guandalini, A.; Marella, D.; Vicard, P.; Vitale, V. On the estimation of the Lorenz curve under complex sampling designs. *Stat. Methods Appl.* **2019**, *29 (1)*, 1–24.

[23] Deville, J.C.; Särndal, C.E.; Sautory, O. Generalized raking procedures in survey sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 1013–1020.

[24] 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.

[25] Valliant, R. Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **2020**, *8*, 231–263.

[26] Schonlau, M.; Couper, M. Options for Conducting Web Surveys. *Stat. Sci.* **2017**, *32*, 279–292.

[27] Ferri-García, R.; Rueda, M.M. Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Stat. Oper. Res. T.* **2018**, *42*, in press.

[28] 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*, 251, doi:10.1186/s12874-020-01134-4.

[29] 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.

[30] Godambe, V.P.; Thompson, M.E. Estimating equations in presence of a nuisance parameter. *Ann. Stat.* **1974**, *2*, 568–571.

[31] Binder, D.A. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *Int. Stat. Rev. Rev. Int. Stat.* **1983**, *51*, 279–292.

[32] Chen, Y.; Li, P.; Wu, C. Doubly Robust Inference With Nonprobability Survey Samples. *J. Am. Stat. Assoc.* **2020**.

[33] Wu, C.; Thompson, M.E. *Sampling Theory and Practice*; Springer Nature: Berlin, Germany, 2020.

[34] Wolter, K.M. *Introduction to Variance Estimation*, 2nd ed.; Springer, Inc.: New York, NY, USA, 2007.

[35] Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26.

[36] Zhao, P.; Haziza, D.; Wu, C. Survey weighted estimating equation inference with nuisance functionals. *J. Econom.* **2020**, *216*, 516–536.

[37] National Institute of Statistics. Living Conditions Survey. Microdata. 2012. https://www.ine.es/en/prodyser/microdatos_en.htm

[38] Handcock, M.S.; Morris, M. *Relative Distribution Methods in the Social Sciences*; Springer Science & Business Media: Berlin, Germany, 2006.

[39] Martínez, S.; Illescas, M.; Martínez, H.; Arcos, A. Calibration estimator for Head Count Index. *Int. J. Comput. Math.* **2020**, *97*, 51–62.

# Appendix A4

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

Ferri-García, Ramón; Rueda, María del Mar (2021)
Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys
*Statistical papers*

Statistical Papers

| STATISTICS & PROBABILITY | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 1.433 | 47/124 | Q2 |

**Abstract**

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

# 1 Introduction

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

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

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

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

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

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

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

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

## 2 Adjustments for nonprobability samples

### 2.1 Calibration

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

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

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

## 2.2 Propensity Score Adjustment

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

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

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

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

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

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

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

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

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

# 3 Data

## 3.1 Artificial data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### 3.2 Real data

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

- 6 target variables: assessment of the current economic situation in Spain and in their own lives (binary, 1 if ”bad” or ”very bad”, 0 otherwise), score on the ideological self-positioning scale (numeric, 1-10), assessment of the central government’s performance(binary, 1 if ”Poor” or ”Very poor”, 0 otherwise), territorial organisation preference (binary, 1 if ”State with no autonomous structures”, 0 otherwise) and national sentiment (binary, 1 if ”Self identification as only Spanish”, 0 otherwise).

- 10 variables to be used as covariates in PSA or calibration variables: frequency of attendance at religious acts, gender, age, education level, socioeconomic status, autonomous community of residence, size of the municipality of residence, nationality, marital status and degree to which voting is expected to change things. Gender, age and size of the municipality were chosen as calibration variables in each simulation run, and were also included as potential covariates for PSA.

- One variable, use of internet in the three months prior to the survey (1 if it was used, 0 otherwise), was taken as a delimiter of the population subset from which nonprobability samples would be drawn. Individuals with a value of 1, but not those with a value of 0, in this variable could belong to the nonprobability sample. The rationale for this delimiter is that it reproduces the conditions that apply in real online surveys, in which people with no internet access cannot be selected to participate.

The pseudopopulation was obtained by bootstrapping the original sample up to $N = 500000$ individuals through simple random sampling with replacement. Prior to the bootstrapping, anyone who did not answer (”Does not know”/”Does not answer”) any of the 17 itemswas excluded, as were the persons who answered

”Other” for education level, or who gave ”Ceuta” or ”Melilla” as their autonomous community of residence. The reason for this filtering process was to remove highly uncommon classes that could produce inconsistencies in a simulated sample and provoke errors in the propensity scoring algorithms. Moreover, the education levels ”No formal education” and ”Primary education” were collapsed into a single class, while missing data in the variable concerning attendance at religious acts was taken as a new class (given that everyone in this group was considered to be atheist or agnostic). After the preprocessing, the sample size before bootstrapping was $n = 2156$.

# 4 Methods

## 4.1 Feature selection algorithms

Feature selection was performed prior to PSA, by considering a predictive model for a target variable with a fixed outcome and exposure variable. In the PSA context, the exposure variable is denoted by $z$,which measures whether an individual has been exposed to the nonprobability sample, and the outcome (whose population values we wish to estimate) is the variable of interest in each case. The following feature selection algorithms were used in the experiment, and their performance was compared to the use of all variables and to the use of the variables provided by Stepwise:

- CFS (Correlation-based Feature Selection) filter with best first search. This algorithm, proposed by Hall (1999), searches the subset of variables which maximises the correlation with the target variable and minimises that between the variables of the subset. Thus, irrelevant and redundant features are discarded from the optimal subset of features for prediction. Note that Pearson’s correlation is used to evaluate the relationships between the variables; if any of the variables within a pair is non-numeric, it is binarised and each of the binary variables is then used separately.

- Chi-square filter. This approach calculates the Cramer’s $V$ value between the target variable and each independent variable, and so the user must define a cut-off point for selection. In our experiment, the cut-off point was the Cramer’s $V$ value with the biggest difference from the $V$ of the next variable in importance (ordered from highest to lowest).

- Gain ratio. This entropy-based filter (Quinlan, 1986) is calculated by dividing the information gain by the entropy of the target variable. The information gain is measured as the difference between the sum of the entropies of the independent and the target variables and the entropy of the target variable after introducing the independent variable into the predictive model (defined as a decision tree). The gain ratio, thus, is a relative continuous measure of the predictive performance of a variable. The cut-off point was chosen in the same way as with the chi-square filter.

- One-R. This algorithm, developed by Holte (1993), is based on very simple rules of association, by which each independent variable is tabulated with the target variable. The number of errors is then determined and interpreted such that higher values represent a stronger predictive power.

- Random Forest importance filter. This algorithm computes the mean importance value across the trees created in a Random Forest model (Breiman 2001) for each independent variable. In our experiment, the importance value taken was the mean decrease in accuracy when the variable was discarded from the Random Forest model.

- Boruta algorithm. This algorithm is based on the Random Forest importance measure, but it considers a set of non-informative variables created from the random shuffling of each independent variable included in the model. As a result, the algorithm selects the variables that have greater importance than non-informative variables. To obtain statistically valid results, the procedure is repeated until every variable has been deemed as ”important” or ”unimportant”. Further details on this algorithm can be consulted in Kursa and Rudnicki (2010).

- LASSO regression (Tibshirani 1996). This regression model performs a variable selection based on introducing penalisation terms into the Ordinary Least Squares equations. As a result, a regression model is provided but only the variables selected have non-zero coefficients. In the present study, we take advantage of the LASSO variable selection technique by extracting the variables with non-zero coefficients and using them as inputs for the propensity estimation models. When all the coefficients of the LASSO model are zero, no PSA is performed and therefore the weights remain unitary.

### 4.2 Estimation with Propensity Score Adjustment and calibration

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

- k-Nearest Neighbours(kNN): $k = 5,7,9$.

- Gradient Boosting Machine (GBM): number of trees $= 50,100,150$, learning rate $= 0.1$, interaction depth $= 1,2,3$.

- Feed-forward neural networks (NN): number of units in the hidden layer $= 1, 3, 5$, weight decay $= 0.1, 0.0001, 0$.

### 4.3 Experiment settings

In both scenarios, the same procedure was followed to measure the effects of variable selection in PSA and calibration on the estimation from nonprobability samples. This procedure, repeated across 200 simulation runs for each dataset (artificial and real), can be sequentially described as follows:

1. Two samples of size n = 1,000 are drawn. The first one, $s_r$, is the probability sample and is drawn by simple random sampling without replacement (SRSWOR) from the full population. The second sample, $s_v$, is the nonprobability sample and is drawn according to the following schemes:

    - Artificial dataset: unequal probability sampling where $\pi$ is the vector of inclusion probabilities, calculated as described in Equation 7.
    - Real dataset: SRSWOR from the subset of the population who had accessed the internet during the three months prior to the survey.

2. Propensity of belonging to $s_v$ is estimated with PSA, using the variable selection algorithms described in Section 4.1 to select the input covariates for propensity prediction models, and the four choices of algorithms described in Section 4.2 to model propensities. We also consider the choice where no variable selection algorithm is applied and all covariates are included in the models.

3. Estimated propensities are transformed into weights using the inverse probability weighting formula $w_i = 1/\pi_i$.

4. Weights are used to estimate the population mean of each target variable with and without applying Raking calibration, on which the propensity weights $w$ obtained in step 3 are used as initial weights.

The resulting 200 estimates of the population mean for each combination of methods are subsequently used to obtain the relative bias of a given combination of methods:

$$RB(\%) = \left| \frac{\sum_{i=1}^{200} \frac{\hat{\bar{y}}_i}{200} - \overline{Y}}{\overline{Y}} \right| \tag{16}$$

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

$$\text{Efficiency}_{k|m,r,v} = \frac{MSE_{k,m,r,v}}{MSE_{\text{All vars.},m,r,v}} \tag{17}$$

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

$$MSE = \text{Bias}^2 + \text{Variance} = \left( \frac{\sum_{i=1}^{200} \hat{\overline{y}}_i}{200} - \overline{Y} \right)^2 + \frac{\sum_{i=1}^{200} \left( \hat{\overline{y}}_i - \frac{\sum_{i=1}^{200} \hat{\overline{y}}_i}{200} \right)^2}{199} \tag{18}$$

An efficiency greater than 1 means that the use of the variable selection method $k$ is inefficient in comparison with using all covariates, while if it remains below 1 the selector $k$ provides more efficient estimates, provided all other adjustments remain equal.

# 5 Results

## 5.1 Artificial data

The relative bias results obtained in the simulation with artificial data are shown in Tables 1 and 2. For the MCAR variables ($y_1$ and $y_2$), variable selection was useful when neural nets were used as the predictive model, although the improvements were not dramatic. The least biased estimates were provided by PSA with kNN using all variables in $y_1$ and variables selected by StepWise in $y_2$, although this result was closely followed by the Gain Ratio, anyother algorithm and no Raking in the latter case. However, the differences are too small to be considered significant.

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

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

| | | Raking = No | | | | Raking = Yes | | | | Raking = No | | | | Raking = Yes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
| | | $y_1$ | | | | | | | | $y_2$ | | | | | | | |
| Exposure | All vars. | 0.287 | 0.255 | 0.178 | 0.325 | 0.301 | 0.273 | 0.192 | 0.373 | 0.082 | 0.083 | 0.082 | 0.084 | 0.084 | 0.085 | 0.086 | 0.086 |
| | Boruta | 0.289 | 0.302 | 0.300 | 0.328 | 0.301 | 0.336 | 0.333 | 0.367 | 0.082 | 0.081 | 0.082 | 0.085 | 0.083 | 0.084 | 0.086 | 0.088 |
| | CFS | 0.302 | 0.312 | 0.346 | 0.315 | 0.296 | 0.305 | 0.319 | 0.300 | 0.083 | 0.081 | 0.083 | 0.083 | 0.082 | 0.083 | 0.076 | 0.085 |
| | Chi-sq. | 0.327 | 0.336 | 0.384 | 0.347 | 0.298 | 0.335 | 0.383 | 0.353 | 0.083 | 0.086 | 0.085 | 0.082 | 0.083 | 0.089 | 0.080 | 0.085 |
| | Gain r. | 0.282 | 0.316 | 0.323 | 0.323 | 0.275 | 0.295 | 0.266 | 0.326 | 0.083 | 0.084 | 0.084 | 0.084 | 0.083 | 0.090 | 0.078 | 0.086 |
| | LASSO | 0.311 | 0.347 | 0.437 | 0.380 | 0.304 | 0.330 | 0.401 | 0.377 | 0.084 | 0.079 | 0.091 | 0.084 | 0.083 | 0.082 | 0.091 | 0.087 |
| | StepWise | 0.285 | 0.319 | 0.309 | 0.342 | 0.299 | 0.339 | 0.293 | 0.340 | 0.083 | 0.083 | 0.091 | 0.086 | 0.084 | 0.089 | 0.095 | 0.092 |
| | OneR | 0.310 | 0.327 | 0.293 | 0.356 | 0.300 | 0.319 | 0.272 | 0.374 | 0.084 | 0.082 | 0.082 | 0.085 | 0.083 | 0.082 | 0.086 | 0.088 |
| | RF imp. | 0.318 | 0.350 | 0.350 | 0.312 | 0.296 | 0.367 | 0.296 | 0.290 | 0.084 | 0.084 | 0.092 | 0.087 | 0.083 | 0.090 | 0.095 | 0.093 |
| Outcome | All vars. | 0.287 | 0.297 | 0.180 | 0.331 | 0.301 | 0.307 | 0.188 | 0.338 | 0.082 | 0.083 | 0.083 | 0.081 | 0.084 | 0.087 | 0.087 | 0.085 |
| | Boruta | 0.361 | 0.347 | 0.348 | 0.375 | 0.352 | 0.330 | 0.325 | 0.321 | 0.082 | 0.084 | 0.090 | 0.088 | 0.085 | 0.083 | 0.094 | 0.095 |
| | CFS | 0.324 | 0.318 | 0.295 | 0.319 | 0.286 | 0.288 | 0.286 | 0.291 | 0.113 | 0.106 | 0.087 | 0.110 | 0.098 | 0.096 | 0.079 | 0.102 |
| | Chi-sq. | 0.336 | 0.323 | 0.297 | 0.325 | 0.287 | 0.273 | 0.287 | 0.284 | 0.081 | 0.080 | 0.082 | 0.081 | 0.083 | 0.081 | 0.083 | 0.082 |
| | Gain r. | 0.328 | 0.323 | 0.296 | 0.322 | 0.286 | 0.285 | 0.286 | 0.285 | 0.077 | 0.078 | 0.082 | 0.078 | 0.083 | 0.083 | 0.083 | 0.083 |
| | LASSO | 0.310 | 0.319 | 0.327 | 0.320 | 0.279 | 0.303 | 0.318 | 0.296 | 0.082 | 0.082 | 0.082 | 0.082 | 0.083 | 0.083 | 0.083 | 0.083 |
| | StepWise | 0.299 | 0.280 | 0.267 | 0.267 | 0.311 | 0.296 | 0.234 | 0.269 | 0.082 | 0.081 | 0.077 | 0.081 | 0.083 | 0.081 | 0.076 | 0.082 |
| | OneR | 0.348 | 0.332 | 0.291 | 0.337 | 0.301 | 0.292 | 0.291 | 0.287 | 0.080 | 0.080 | 0.082 | 0.080 | 0.083 | 0.082 | 0.083 | 0.081 |
| | RF imp. | 0.320 | 0.316 | 0.245 | 0.314 | 0.337 | 0.314 | 0.249 | 0.293 | 0.084 | 0.088 | 0.089 | 0.083 | 0.088 | 0.088 | 0.094 | 0.088 |
| | | $y_3$ | | | | | | | | $y_4$ | | | | | | | |
| Exposure | All vars. | 11.53 | 11.60 | 11.00 | 11.63 | 8.90 | 9.42 | 8.94 | 9.62 | 12.39 | 12.52 | 11.89 | 12.62 | 9.69 | 10.28 | 9.84 | 10.60 |
| | Boruta | 11.53 | 11.65 | 11.23 | 11.75 | 8.91 | 9.47 | 9.27 | 9.78 | 12.39 | 12.56 | 12.06 | 12.69 | 9.69 | 10.36 | 10.08 | 10.73 |
| | CFS | 11.42 | 11.51 | 11.10 | 11.57 | 8.77 | 9.20 | 9.01 | 9.33 | 12.27 | 12.38 | 12.00 | 12.46 | 9.55 | 10.02 | 9.85 | 10.20 |
| | Chi-sq. | 11.48 | 11.58 | 11.26 | 11.68 | 8.77 | 9.30 | 9.22 | 9.52 | 12.34 | 12.48 | 12.10 | 12.57 | 9.55 | 10.16 | 10.00 | 10.36 |
| | Gain r. | 11.38 | 11.50 | 11.16 | 11.56 | 8.78 | 9.23 | 9.11 | 9.35 | 12.25 | 12.37 | 12.01 | 12.45 | 9.56 | 10.04 | 9.86 | 10.20 |
| | LASSO | 11.50 | 11.57 | 11.24 | 11.66 | 8.80 | 9.30 | 9.19 | 9.51 | 12.35 | 12.47 | 12.07 | 12.59 | 9.58 | 10.15 | 9.98 | 10.40 |
| | StepWise | 11.53 | 11.62 | 11.20 | 11.74 | 8.90 | 9.44 | 9.19 | 9.77 | 12.39 | 12.57 | 12.09 | 12.70 | 9.69 | 10.34 | 10.11 | 10.71 |
| | OneR | 11.38 | 11.39 | 11.10 | 11.46 | 8.71 | 9.02 | 8.96 | 9.13 | 12.23 | 12.26 | 11.93 | 12.32 | 9.50 | 9.82 | 9.69 | 9.94 |
| | RF imp. | 11.43 | 11.57 | 11.22 | 11.66 | 8.82 | 9.41 | 9.26 | 9.60 | 12.28 | 12.51 | 12.11 | 12.63 | 9.61 | 10.31 | 10.09 | 10.56 |
| Outcome | All vars. | 11.53 | 11.60 | 10.99 | 11.63 | 8.90 | 9.42 | 8.94 | 9.63 | 12.39 | 12.51 | 11.88 | 12.64 | 9.69 | 10.28 | 9.84 | 10.63 |
| | Boruta | 11.42 | 11.57 | 11.13 | 11.61 | 8.88 | 9.42 | 9.12 | 9.54 | 12.37 | 12.55 | 12.05 | 12.68 | 9.67 | 10.34 | 10.05 | 10.68 |
| | CFS | 11.25 | 11.23 | 10.51 | 11.22 | 8.60 | 8.60 | 8.58 | 8.60 | 12.20 | 12.51 | 12.09 | 12.62 | 9.67 | 10.41 | 10.10 | 10.58 |
| | Chi-sq. | 11.26 | 11.24 | 10.51 | 11.22 | 8.59 | 8.60 | 8.58 | 8.59 | 12.38 | 12.58 | 12.13 | 12.71 | 9.66 | 10.38 | 10.19 | 10.75 |
| | Gain r. | 11.23 | 11.22 | 10.51 | 11.20 | 8.60 | 8.60 | 8.58 | 8.61 | 12.36 | 12.56 | 12.16 | 12.69 | 9.59 | 10.33 | 10.14 | 10.63 |
| | LASSO | 10.46 | 10.45 | 10.45 | 10.45 | 8.55 | 8.55 | 8.55 | 8.55 | 12.39 | 12.58 | 12.18 | 12.72 | 9.58 | 10.29 | 10.13 | 10.61 |
| | StepWise | 11.36 | 11.31 | 10.86 | 11.37 | 8.76 | 8.80 | 8.70 | 8.82 | 12.40 | 12.58 | 12.12 | 12.70 | 9.62 | 10.31 | 10.09 | 10.61 |
| | OneR | 10.95 | 10.97 | 10.72 | 11.02 | 8.72 | 8.93 | 8.74 | 9.03 | 12.39 | 12.60 | 12.13 | 12.71 | 9.69 | 10.41 | 10.18 | 10.76 |
| | RF imp. | 11.23 | 11.33 | 10.99 | 11.41 | 8.85 | 9.18 | 8.97 | 9.27 | 12.38 | 12.58 | 12.17 | 12.71 | 9.61 | 10.36 | 10.18 | 10.67 |

Table 1: Mean Relative Bias of the estimates of population means for variables $y_1, y_2, y_3, y_4$ in the artificial data simulation for each combination of methods. The closer to zero a value, the less biased the mean estimate obtained.

| | | Raking = No | | | | Raking = Yes | | | | Raking = No | | | | Raking = Yes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
| | | $y_5$ | | | | | | | | $y_6$ | | | | | | | |
| Exposure | All vars. | 18.13 | 16.50 | 14.93 | 15.84 | 8.72 | 7.79 | 6.89 | 7.31 | 16.78 | 14.79 | 13.35 | 14.29 | 2.88 | 2.49 | 2.12 | 2.30 |
| | Boruta | 18.13 | 16.40 | 14.98 | 15.73 | 8.73 | 7.66 | 6.90 | 7.14 | 16.78 | 14.73 | 13.35 | 14.24 | 2.89 | 2.45 | 2.15 | 2.26 |
| | CFS | 18.00 | 17.36 | 15.84 | 17.29 | 8.44 | 8.65 | 7.61 | 8.66 | 16.65 | 15.42 | 14.02 | 15.37 | 2.83 | 2.74 | 2.35 | 2.72 |
| | Chi-sq. | 18.11 | 17.02 | 15.62 | 16.90 | 8.33 | 8.20 | 7.29 | 8.10 | 16.85 | 15.25 | 13.93 | 15.20 | 2.76 | 2.59 | 2.26 | 2.56 |
| | Gain r. | 17.97 | 17.27 | 15.84 | 17.21 | 8.54 | 8.63 | 7.61 | 8.58 | 16.52 | 15.29 | 14.02 | 15.28 | 2.88 | 2.74 | 2.39 | 2.71 |
| | LASSO | 18.11 | 16.99 | 15.51 | 16.77 | 8.39 | 8.12 | 7.23 | 7.91 | 16.83 | 15.25 | 13.81 | 15.13 | 2.78 | 2.58 | 2.25 | 2.50 |
| | StepWise | 18.12 | 16.39 | 14.99 | 15.85 | 8.70 | 7.58 | 6.89 | 7.22 | 16.78 | 14.78 | 13.38 | 14.36 | 2.87 | 2.42 | 2.15 | 2.28 |
| | OneR | 17.93 | 17.73 | 16.17 | 17.88 | 8.17 | 9.07 | 7.91 | 9.44 | 16.76 | 15.77 | 14.37 | 15.89 | 2.72 | 2.85 | 2.44 | 2.95 |
| | RF imp. | 18.06 | 16.55 | 15.21 | 16.21 | 8.82 | 7.92 | 7.08 | 7.46 | 16.44 | 14.70 | 13.50 | 14.53 | 2.99 | 2.55 | 2.24 | 2.42 |
| Outcome | All vars. | 18.13 | 16.54 | 14.93 | 15.86 | 8.72 | 7.83 | 6.88 | 7.31 | 16.78 | 14.81 | 13.37 | 14.30 | 2.88 | 2.51 | 2.12 | 2.30 |
| | Boruta | 18.03 | 16.69 | 15.24 | 16.21 | 8.77 | 8.11 | 7.21 | 7.79 | 16.63 | 15.22 | 12.94 | 14.84 | 2.63 | 2.33 | 2.12 | 2.20 |
| | CFS | 18.16 | 16.77 | 15.42 | 16.46 | 8.44 | 7.91 | 7.13 | 7.55 | 15.68 | 15.27 | 14.16 | 15.37 | 3.23 | 3.15 | 2.71 | 3.21 |
| | Chi-sq. | 17.90 | 17.15 | 15.73 | 17.03 | 8.89 | 8.75 | 7.74 | 8.68 | 16.29 | 16.25 | 12.06 | 16.25 | 2.01 | 2.01 | 2.01 | 2.01 |
| | Gain r. | 18.11 | 16.57 | 15.22 | 16.32 | 8.41 | 7.70 | 6.93 | 7.35 | 16.29 | 16.27 | 12.06 | 16.25 | 2.01 | 2.01 | 2.01 | 2.01 |
| | LASSO | 17.84 | 17.86 | 16.31 | 18.04 | 8.63 | 9.49 | 8.19 | 9.77 | 16.74 | 15.88 | 14.45 | 16.00 | 2.72 | 2.89 | 2.48 | 3.01 |
| | StepWise | 17.91 | 17.55 | 15.86 | 17.40 | 8.51 | 9.05 | 7.71 | 8.95 | 16.75 | 15.64 | 14.13 | 15.64 | 2.74 | 2.81 | 2.37 | 2.83 |
| | OneR | 17.47 | 17.56 | 16.17 | 17.64 | 9.08 | 9.57 | 8.33 | 9.69 | 16.74 | 15.88 | 14.47 | 16.02 | 2.72 | 2.89 | 2.48 | 3.01 |
| | RF imp. | 17.67 | 17.78 | 16.37 | 17.96 | 9.13 | 9.74 | 8.49 | 9.97 | 16.29 | 16.28 | 12.06 | 16.25 | 2.01 | 2.01 | 2.01 | 2.01 |
| | | $y_7$ | | | | | | | | $y_8$ | | | | | | | |
| Exposure | All vars. | 10.77 | 10.36 | 9.16 | 10.17 | 7.73 | 6.79 | 6.24 | 6.95 | 15.42 | 15.53 | 14.07 | 15.30 | 11.07 | 11.04 | 10.50 | 11.41 |
| | Boruta | 10.77 | 10.33 | 9.29 | 10.15 | 7.72 | 6.77 | 6.38 | 6.90 | 15.43 | 15.53 | 14.29 | 15.37 | 11.07 | 11.11 | 10.75 | 11.49 |
| | CFS | 10.48 | 10.50 | 9.57 | 10.53 | 6.90 | 7.06 | 6.52 | 7.11 | 15.52 | 15.59 | 14.54 | 15.67 | 10.47 | 10.88 | 10.55 | 11.02 |
| | Chi-sq. | 10.66 | 10.47 | 9.56 | 10.47 | 6.95 | 6.86 | 6.40 | 6.89 | 15.78 | 15.67 | 14.63 | 15.77 | 10.52 | 10.93 | 10.64 | 11.08 |
| | Gain r. | 10.47 | 10.47 | 9.61 | 10.49 | 6.93 | 7.03 | 6.54 | 7.05 | 15.49 | 15.54 | 14.56 | 15.65 | 10.49 | 10.88 | 10.56 | 11.01 |
| | LASSO | 10.67 | 10.50 | 9.56 | 10.47 | 7.08 | 6.89 | 6.46 | 6.90 | 15.72 | 15.67 | 14.57 | 15.74 | 10.62 | 10.94 | 10.66 | 11.13 |
| | StepWise | 10.77 | 10.31 | 9.37 | 10.18 | 7.72 | 6.72 | 6.46 | 6.92 | 15.42 | 15.51 | 14.32 | 15.34 | 11.06 | 11.09 | 10.79 | 11.48 |
| | OneR | 10.13 | 10.42 | 9.57 | 10.53 | 6.60 | 7.23 | 6.61 | 7.39 | 15.07 | 15.34 | 14.46 | 15.46 | 10.28 | 10.75 | 10.46 | 10.88 |
| | RF imp. | 10.82 | 10.40 | 9.51 | 10.34 | 7.23 | 6.73 | 6.35 | 6.61 | 15.93 | 15.70 | 14.63 | 15.78 | 10.68 | 11.01 | 10.69 | 11.14 |
| Outcome | All vars. | 10.77 | 10.35 | 9.16 | 10.17 | 7.73 | 6.78 | 6.24 | 6.93 | 15.42 | 15.54 | 14.06 | 15.30 | 11.07 | 11.04 | 10.49 | 11.41 |
| | Boruta | 10.69 | 10.32 | 9.34 | 10.19 | 7.75 | 6.92 | 6.48 | 7.02 | 14.97 | 14.98 | 13.95 | 14.87 | 10.81 | 10.80 | 10.52 | 11.13 |
| | CFS | 10.64 | 10.25 | 9.40 | 10.18 | 7.84 | 7.02 | 6.53 | 7.02 | 14.90 | 14.51 | 13.59 | 14.63 | 10.48 | 10.43 | 10.14 | 10.46 |
| | Chi-sq. | 10.40 | 10.24 | 9.33 | 10.29 | 7.42 | 7.12 | 6.60 | 7.26 | 14.09 | 14.10 | 13.37 | 14.05 | 10.30 | 10.32 | 10.23 | 10.47 |
| | Gain r. | 10.71 | 10.28 | 9.38 | 10.27 | 7.93 | 6.98 | 6.51 | 7.07 | 13.36 | 13.35 | 12.87 | 13.33 | 9.92 | 9.92 | 9.92 | 9.92 |
| | LASSO | 9.98 | 10.09 | 9.20 | 10.19 | 7.03 | 7.28 | 6.60 | 7.34 | 15.42 | 15.52 | 14.34 | 15.39 | 11.07 | 11.13 | 10.85 | 11.50 |
| | StepWise | 10.25 | 10.21 | 9.25 | 10.28 | 7.24 | 7.06 | 6.42 | 7.05 | 15.41 | 15.49 | 14.26 | 15.35 | 11.06 | 11.09 | 10.76 | 11.47 |
| | OneR | 9.54 | 9.64 | 8.95 | 9.71 | 6.62 | 6.91 | 6.39 | 6.97 | 14.90 | 14.53 | 13.57 | 14.63 | 10.48 | 10.43 | 10.14 | 10.46 |
| | RF imp. | 9.93 | 10.08 | 9.16 | 10.22 | 6.81 | 7.26 | 6.58 | 7.39 | 14.42 | 14.47 | 13.71 | 14.57 | 10.50 | 10.55 | 10.45 | 10.91 |

Table 2: Mean Relative Bias of the estimates of population means for variables $y_5, y_6, y_7, y_8$ in the artificial data simulation for each combination of methods. The closer to zero a value, the less biased the mean estimate obtained.

The efficiency of each variable selection method in comparison to using all variables, if the rest of methods remain equal, is detailed in Tables 3 and 4. These results are in line with those for relative bias in each case, although they reflect some improvement in a much larger set of situations. With the MCAR variables ($y_1$ and $y_2$), MSE reductions of up to 10% were obtained when k-NN and Raking were used to estimate $y_2$, with the Stepwise algorithm. When Raking calibration was not applied, other variable selection methods (Chi-square, Gain Ratio, LASSO and OneR) provided reductions of 7% in the same situation. In the case of $y_1$, variable selection improved the efficiency when Raking calibration was applied, especially if propensities were estimated using neural nets.

Regarding the MAR variables ($y_5$ and $y_6$), very significant improvements in efficiency were obtained in the estimation of $y_6$. When Raking calibration was applied, the use of the Chi-square filter, Gain Ratio or Random Forest reduced MSE by 10% to 50%, when they selected variables using the outcome variable as thetarget, depending on the predictive model used for the propensities. Reductions in MSE with the same variable selection methods were also observed when Raking was not applied. Other methods, too, provided larger efficiency values when $y_5$ was estimated for the cases in which logistic regression was used to estimate the propensities.

Finally, regarding the NMAR variables, the reductions in MSE were around 20% in several cases: in $y_3$ when neural nets were used to estimate propensities and Raking calibration was applied; in $y_7$ when logistic regression was applied(in this case OneR achieved the best results regardless of the target); and in $y_8$ when the Gain Ratio was used,for all situations when the outcome was fixed as the target variable for the algorithm. In the remaining cases in which variable selection algorithms were properly applied, the reductions obtained were around 10-15% of the MSE observed when all available covariates were used.

| | | Raking = No | | | | Raking = Yes | | | | Raking = No | | | | Raking = Yes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
| | | $y_1$ | | | | | | | | $y_2$ | | | | | | | |
| Exposure | Boruta | 1.003 | 1.011 | 1.032 | 0.967 | 1.006 | 1.018 | 1.041 | 0.958 | 0.993 | 0.966 | 1.048 | 0.996 | 0.996 | 0.967 | 1.034 | 0.996 |
| | CFS | 1.015 | 1.030 | 1.075 | 1.010 | 1.007 | 1.037 | 1.111 | 1.000 | 0.989 | 1.001 | 1.002 | 0.988 | 0.990 | 0.993 | 0.991 | 0.980 |
| | Chi-sq. | 1.005 | 1.017 | 1.033 | 1.009 | 1.000 | 1.030 | 1.029 | 1.030 | 0.988 | 0.994 | 0.976 | 0.995 | 0.989 | 1.003 | 0.978 | 1.017 |
| | Gain r. | 1.004 | 1.009 | 1.054 | 0.995 | 1.009 | 1.012 | 1.096 | 0.992 | 0.979 | 0.996 | 1.044 | 1.000 | 0.997 | 1.025 | 1.058 | 1.005 |
| | LASSO | 0.996 | 1.024 | 1.056 | 1.018 | 0.990 | 1.023 | 1.042 | 1.004 | 0.999 | 1.003 | 0.994 | 1.012 | 0.988 | 0.999 | 1.005 | 1.008 |
| | StepWise | 1.001 | 1.007 | 0.983 | 0.990 | 1.004 | 1.005 | 1.002 | 0.997 | 1.000 | 1.001 | 1.002 | 0.957 | 1.004 | 1.013 | 1.014 | 0.945 |
| | OneR | 1.017 | 1.032 | 1.092 | 1.004 | 0.998 | 1.033 | 1.133 | 1.002 | 0.986 | 1.009 | 0.937 | 0.999 | 0.984 | 0.998 | 0.985 | 0.992 |
| | RF imp. | 1.004 | 1.025 | 1.000 | 1.013 | 1.015 | 1.040 | 1.023 | 1.010 | 0.991 | 1.003 | 1.014 | 0.984 | 1.001 | 1.006 | 1.009 | 0.974 |
| Outcome | Boruta | 1.032 | 0.997 | 1.042 | 1.034 | 1.001 | 0.963 | 1.027 | 0.964 | 1.005 | 0.974 | 1.020 | 1.023 | 1.009 | 0.975 | 1.010 | 1.028 |
| | CFS | 0.996 | 1.014 | 0.994 | 1.013 | 0.974 | 0.975 | 0.986 | 0.953 | 1.007 | 1.030 | 0.940 | 1.039 | 0.987 | 0.998 | 0.956 | 1.025 |
| | Chi-sq. | 1.001 | 1.022 | 0.994 | 1.021 | 0.974 | 0.982 | 0.986 | 0.953 | 0.973 | 0.992 | 0.928 | 1.003 | 0.975 | 0.989 | 0.948 | 1.012 |
| | Gain r. | 0.995 | 1.014 | 0.994 | 1.012 | 0.974 | 0.977 | 0.986 | 0.954 | 0.966 | 0.988 | 0.928 | 0.982 | 0.976 | 0.989 | 0.948 | 1.001 |
| | LASSO | 0.981 | 0.994 | 0.996 | 0.996 | 0.978 | 0.971 | 0.982 | 0.953 | 0.939 | 0.962 | 0.928 | 0.960 | 0.976 | 0.991 | 0.948 | 0.998 |
| | StepWise | 0.993 | 0.997 | 1.017 | 0.966 | 0.986 | 0.980 | 1.021 | 0.932 | 0.991 | 1.000 | 0.956 | 1.012 | 0.976 | 0.978 | 0.907 | 0.993 |
| | OneR | 1.024 | 1.059 | 0.989 | 1.056 | 0.966 | 0.983 | 0.983 | 0.950 | 0.969 | 0.988 | 0.928 | 0.995 | 0.975 | 0.988 | 0.948 | 1.002 |
| | RF imp. | 1.001 | 1.027 | 1.022 | 0.997 | 1.001 | 0.984 | 1.022 | 0.935 | 0.948 | 0.983 | 1.020 | 0.976 | 1.005 | 0.990 | 1.021 | 1.011 |
| | | $y_3$ | | | | | | | | $y_4$ | | | | | | | |
| Exposure | Boruta | 1.001 | 1.010 | 1.041 | 1.021 | 1.000 | 1.014 | 1.069 | 1.031 | 1.000 | 1.007 | 1.028 | 1.011 | 1.000 | 1.014 | 1.049 | 1.023 |
| | CFS | 0.983 | 0.986 | 1.019 | 0.989 | 0.971 | 0.960 | 1.018 | 0.947 | 0.981 | 0.978 | 1.019 | 0.976 | 0.971 | 0.951 | 1.002 | 0.926 |
| | Chi-sq. | 0.992 | 0.997 | 1.048 | 1.008 | 0.971 | 0.979 | 1.061 | 0.981 | 0.992 | 0.994 | 1.036 | 0.992 | 0.972 | 0.976 | 1.032 | 0.955 |
| | Gain r. | 0.975 | 0.984 | 1.031 | 0.988 | 0.971 | 0.966 | 1.040 | 0.950 | 0.978 | 0.976 | 1.020 | 0.974 | 0.973 | 0.953 | 1.004 | 0.926 |
| | LASSO | 0.995 | 0.996 | 1.046 | 1.006 | 0.977 | 0.979 | 1.055 | 0.980 | 0.993 | 0.993 | 1.030 | 0.996 | 0.977 | 0.974 | 1.029 | 0.964 |
| | StepWise | 1.000 | 1.005 | 1.038 | 1.019 | 0.999 | 1.006 | 1.058 | 1.032 | 1.000 | 1.008 | 1.034 | 1.013 | 0.999 | 1.012 | 1.056 | 1.021 |
| | OneR | 0.974 | 0.966 | 1.021 | 0.972 | 0.958 | 0.926 | 1.012 | 0.911 | 0.975 | 0.959 | 1.007 | 0.954 | 0.960 | 0.913 | 0.970 | 0.878 |
| | RF imp. | 0.984 | 0.996 | 1.045 | 1.005 | 0.982 | 1.000 | 1.073 | 0.995 | 0.983 | 0.998 | 1.037 | 1.002 | 0.983 | 1.005 | 1.051 | 0.991 |
| Outcome | Boruta | 0.982 | 0.997 | 1.027 | 0.997 | 0.996 | 1.003 | 1.044 | 0.983 | 0.997 | 1.006 | 1.028 | 1.007 | 0.996 | 1.012 | 1.044 | 1.008 |
| | CFS | 0.953 | 0.940 | 0.919 | 0.932 | 0.933 | 0.844 | 0.928 | 0.809 | 0.970 | 1.000 | 1.035 | 0.997 | 0.996 | 1.025 | 1.054 | 0.990 |
| | Chi-sq. | 0.954 | 0.941 | 0.920 | 0.933 | 0.932 | 0.844 | 0.929 | 0.806 | 0.999 | 1.012 | 1.043 | 1.011 | 0.994 | 1.021 | 1.072 | 1.023 |
| | Gain r. | 0.950 | 0.938 | 0.919 | 0.930 | 0.933 | 0.845 | 0.928 | 0.809 | 0.995 | 1.008 | 1.047 | 1.009 | 0.979 | 1.011 | 1.062 | 1.000 |
| | LASSO | 0.829 | 0.819 | 0.909 | 0.814 | 0.922 | 0.834 | 0.920 | 0.798 | 1.001 | 1.012 | 1.050 | 1.014 | 0.976 | 1.004 | 1.060 | 0.996 |
| | StepWise | 0.972 | 0.953 | 0.983 | 0.958 | 0.968 | 0.885 | 0.959 | 0.850 | 1.002 | 1.011 | 1.040 | 1.010 | 0.986 | 1.007 | 1.051 | 0.997 |
| | OneR | 0.915 | 0.908 | 0.960 | 0.912 | 0.961 | 0.917 | 0.965 | 0.897 | 1.000 | 1.014 | 1.042 | 1.012 | 0.999 | 1.027 | 1.071 | 1.024 |
| | RF imp. | 0.952 | 0.958 | 1.001 | 0.967 | 0.987 | 0.954 | 1.007 | 0.930 | 0.999 | 1.011 | 1.050 | 1.013 | 0.982 | 1.016 | 1.071 | 1.007 |

Table 3: Efficiency of the estimates of population means for variables $y_1, y_2, y_3, y_4$ in the artificial data simulation for each combination of methods. Values greater than one indicate inefficiency, while values below one show that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used.

| | | Raking = No | | | | Raking = Yes | | | | Raking = No | | | | Raking = Yes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
| | | $y_5$ | | | | | | | | $y_6$ | | | | | | | |
| Exposure | Boruta | 1.001 | 0.987 | 1.006 | 0.987 | 1.001 | 0.970 | 1.006 | 0.960 | 1.000 | 0.992 | 0.999 | 0.992 | 1.003 | 0.968 | 1.027 | 0.967 |
| | CFS | 0.987 | 1.103 | 1.123 | 1.187 | 0.940 | 1.210 | 1.210 | 1.367 | 0.986 | 1.086 | 1.102 | 1.156 | 0.969 | 1.198 | 1.232 | 1.388 |
| | Chi-sq. | 0.999 | 1.063 | 1.094 | 1.138 | 0.917 | 1.100 | 1.118 | 1.226 | 1.009 | 1.064 | 1.089 | 1.131 | 0.921 | 1.078 | 1.140 | 1.230 |
| | Gain r. | 0.983 | 1.092 | 1.124 | 1.176 | 0.963 | 1.207 | 1.215 | 1.349 | 0.972 | 1.068 | 1.102 | 1.142 | 1.005 | 1.202 | 1.273 | 1.388 |
| | LASSO | 0.998 | 1.059 | 1.079 | 1.121 | 0.932 | 1.085 | 1.106 | 1.178 | 1.007 | 1.063 | 1.070 | 1.121 | 0.933 | 1.070 | 1.125 | 1.183 |
| | StepWise | 1.000 | 0.986 | 1.006 | 1.001 | 0.996 | 0.957 | 1.007 | 0.986 | 1.000 | 0.998 | 1.004 | 1.010 | 0.996 | 0.946 | 1.039 | 0.981 |
| | OneR | 0.979 | 1.151 | 1.171 | 1.268 | 0.886 | 1.316 | 1.292 | 1.592 | 0.998 | 1.136 | 1.158 | 1.235 | 0.895 | 1.287 | 1.324 | 1.602 |
| | RF imp. | 0.993 | 1.006 | 1.039 | 1.048 | 1.020 | 1.040 | 1.065 | 1.054 | 0.961 | 0.988 | 1.022 | 1.034 | 1.080 | 1.055 | 1.132 | 1.117 |
| Outcome | Boruta | 0.990 | 1.019 | 1.042 | 1.046 | 1.012 | 1.079 | 1.097 | 1.139 | 0.983 | 1.057 | 0.939 | 1.080 | 0.859 | 0.874 | 0.998 | 0.925 |
| | CFS | 1.004 | 1.027 | 1.066 | 1.076 | 0.944 | 1.031 | 1.083 | 1.078 | 0.874 | 1.062 | 1.121 | 1.152 | 1.235 | 1.537 | 1.610 | 1.878 |
| | Chi-sq. | 0.975 | 1.077 | 1.112 | 1.155 | 1.039 | 1.263 | 1.278 | 1.421 | 0.943 | 1.201 | 0.813 | 1.287 | 0.490 | 0.639 | 0.895 | 0.755 |
| | Gain r. | 0.999 | 1.004 | 1.038 | 1.058 | 0.934 | 0.979 | 1.023 | 1.020 | 0.943 | 1.204 | 0.813 | 1.287 | 0.490 | 0.639 | 0.895 | 0.755 |
| | LASSO | 0.968 | 1.162 | 1.192 | 1.285 | 0.991 | 1.427 | 1.394 | 1.695 | 0.995 | 1.148 | 1.167 | 1.249 | 0.890 | 1.308 | 1.352 | 1.669 |
| | StepWise | 0.976 | 1.124 | 1.127 | 1.202 | 0.963 | 1.311 | 1.242 | 1.465 | 0.997 | 1.114 | 1.117 | 1.195 | 0.905 | 1.240 | 1.250 | 1.494 |
| | OneR | 0.931 | 1.125 | 1.172 | 1.233 | 1.071 | 1.451 | 1.437 | 1.677 | 0.995 | 1.148 | 1.171 | 1.252 | 0.890 | 1.308 | 1.361 | 1.669 |
| | RF imp. | 0.950 | 1.151 | 1.202 | 1.273 | 1.087 | 1.487 | 1.492 | 1.752 | 0.943 | 1.205 | 0.813 | 1.287 | 0.490 | 0.639 | 0.895 | 0.755 |
| | | $y_7$ | | | | | | | | $y_8$ | | | | | | | |
| Exposure | Boruta | 1.000 | 0.995 | 1.030 | 0.993 | 0.999 | 0.998 | 1.043 | 0.981 | 1.001 | 1.000 | 1.032 | 1.008 | 1.000 | 1.012 | 1.049 | 1.014 |
| | CFS | 0.950 | 1.026 | 1.092 | 1.065 | 0.806 | 1.074 | 1.089 | 1.040 | 1.015 | 1.008 | 1.068 | 1.047 | 0.895 | 0.970 | 1.009 | 0.933 |
| | Chi-sq. | 0.981 | 1.021 | 1.088 | 1.055 | 0.820 | 1.022 | 1.051 | 0.981 | 1.047 | 1.018 | 1.081 | 1.059 | 0.903 | 0.979 | 1.026 | 0.943 |
| | Gain r. | 0.948 | 1.018 | 1.100 | 1.058 | 0.814 | 1.066 | 1.094 | 1.028 | 1.010 | 1.001 | 1.071 | 1.043 | 0.898 | 0.970 | 1.011 | 0.931 |
| | LASSO | 0.985 | 1.025 | 1.089 | 1.055 | 0.852 | 1.025 | 1.071 | 0.981 | 1.041 | 1.019 | 1.073 | 1.057 | 0.921 | 0.982 | 1.031 | 0.952 |
| | StepWise | 1.000 | 0.991 | 1.049 | 1.000 | 0.998 | 0.986 | 1.074 | 0.993 | 1.000 | 0.998 | 1.035 | 1.005 | 0.999 | 1.009 | 1.056 | 1.014 |
| | OneR | 0.885 | 1.008 | 1.090 | 1.064 | 0.740 | 1.118 | 1.110 | 1.113 | 0.953 | 0.975 | 1.054 | 1.018 | 0.863 | 0.948 | 0.991 | 0.911 |
| | RF imp. | 1.009 | 1.007 | 1.080 | 1.030 | 0.880 | 0.985 | 1.038 | 0.906 | 1.067 | 1.023 | 1.080 | 1.063 | 0.931 | 0.994 | 1.036 | 0.954 |
| Outcome | Boruta | 0.986 | 0.994 | 1.041 | 1.003 | 1.011 | 1.041 | 1.078 | 1.023 | 0.945 | 0.933 | 0.986 | 0.947 | 0.955 | 0.959 | 1.008 | 0.954 |
| | CFS | 0.977 | 0.983 | 1.053 | 1.002 | 1.026 | 1.074 | 1.093 | 1.026 | 0.933 | 0.872 | 0.936 | 0.914 | 0.896 | 0.893 | 0.935 | 0.841 |
| | Chi-sq. | 0.936 | 0.981 | 1.040 | 1.023 | 0.930 | 1.101 | 1.112 | 1.087 | 0.840 | 0.829 | 0.907 | 0.847 | 0.867 | 0.875 | 0.953 | 0.845 |
| | Gain r. | 0.989 | 0.986 | 1.049 | 1.019 | 1.052 | 1.055 | 1.082 | 1.037 | 0.752 | 0.740 | 0.837 | 0.761 | 0.802 | 0.807 | 0.894 | 0.756 |
| | LASSO | 0.863 | 0.953 | 1.011 | 1.002 | 0.835 | 1.141 | 1.114 | 1.110 | 1.000 | 0.997 | 1.041 | 1.012 | 0.999 | 1.017 | 1.071 | 1.016 |
| | StepWise | 0.910 | 0.975 | 1.021 | 1.021 | 0.888 | 1.087 | 1.066 | 1.048 | 0.999 | 0.994 | 1.030 | 1.007 | 0.998 | 1.009 | 1.052 | 1.012 |
| | OneR | 0.788 | 0.871 | 0.954 | 0.913 | 0.741 | 1.030 | 1.042 | 0.998 | 0.933 | 0.875 | 0.933 | 0.914 | 0.896 | 0.892 | 0.934 | 0.841 |
| | RF imp. | 0.854 | 0.951 | 1.002 | 1.008 | 0.786 | 1.136 | 1.104 | 1.119 | 0.880 | 0.872 | 0.954 | 0.911 | 0.900 | 0.914 | 0.993 | 0.917 |

Table 4: Efficiency of the estimates of population means for variables $y_5, y_6, y_7, y_8$ in the artificial data simulation for each combination of methods. Values greater than one indicate inefficiency, while values below one show that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used.

## 5.2 Real data

The relative bias results obtained by each combination of methods in the simulation using CIS data are listed in Table 5. Interestingly, the best choice in variable selection differed according to the propensity estimation model considered. For example, PSA using k-NN provided the best results when using all the available covariates, except for the variable measuring central government performance. In the remaining cases, the use of certain variable selection algorithms was associated with a decrease in relative bias. This was especially apparent for the variables measuring the economic situation in Spain, central government management and the preference for a unitary national state without autonomous communities. In these cases, the largest reductions in relative bias(compared to the case in which all variables were used) were obtained when the variable selection algorithms used the outcome (the actual variable to be estimated) as the target variable. Raking calibration had a modest positive effect on the variables measuring the economic situation in Spain, the preference for a unitary national state without autonomous communities and whether the respondent self identified as only Spanish, while its impact on relative bias in the other variables was non-significant or negative. The efficiency of each variable selection algorithm for a given combination of adjustments (propensity model, use of calibration and target variable choice for selection), in comparison with the case in which all variables are used,is shown in Table 6. For all variables, one or more selection algorithms increased the efficiency, in comparison with the case in which all variables were used. MSE reductions of around 20% were measured in some cases; for example, when the OneR algorithm was used to estimate the variable measuring the economic situation in Spain, when CFS was used to estimate ideological self-positioning, and when CFS, Chi-square, Gain ratio or Random Forest were used to estimate the preference for a unitary national state without autonomous communities. In most cases, the efficiency gains produced a 10% reduction in the MSE.

|  |  | Raking = No |  |  |  | Raking = Yes |  |  |  | Raking = No |  |  |  | Raking = Yes |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
|  |  | Econ. situation in Spain ”poor” or ”very poor” |  |  |  |  |  |  |  | Personal econ. situation ”poor” or ”very poor” |  |  |  |  |  |  |  |
| Exposure | All vars. | 7.20 | 6.17 | 5.16 | 5.68 | 7.66 | 6.59 | 5.36 | 6.33 | 11.70 | 11.14 | 10.96 | 10.63 | 24.43 | 23.18 | 20.70 | 22.36 |
|  | Boruta | 7.15 | 6.09 | 5.93 | 5.57 | 7.53 | 6.46 | 6.29 | 6.02 | 11.21 | 10.81 | 11.47 | 10.86 | 23.98 | 23.13 | 21.85 | 22.80 |
|  | Cfs | 7.93 | 6.35 | 5.40 | 6.32 | 7.86 | 6.38 | 5.66 | 6.31 | 13.40 | 12.35 | 11.07 | 12.45 | 24.33 | 23.71 | 20.90 | 23.75 |
|  | Chi | 7.14 | 5.89 | 5.40 | 5.57 | 7.31 | 6.05 | 5.68 | 5.69 | 10.98 | 10.75 | 11.14 | 10.69 | 23.63 | 23.16 | 21.00 | 23.04 |
|  | Gain | 8.23 | 6.55 | 5.37 | 6.70 | 8.12 | 6.52 | 5.62 | 6.63 | 14.66 | 13.50 | 10.90 | 13.75 | 24.63 | 24.12 | 20.72 | 24.32 |
|  | Lasso | 7.15 | 5.98 | 5.77 | 5.65 | 7.37 | 6.19 | 6.09 | 5.86 | 11.09 | 10.83 | 11.59 | 10.69 | 23.95 | 23.36 | 21.68 | 23.21 |
|  | StepWise | 7.34 | 6.09 | 6.27 | 5.71 | 7.51 | 6.30 | 6.62 | 5.83 | 11.24 | 11.00 | 12.23 | 10.75 | 23.98 | 23.44 | 22.37 | 23.17 |
|  | OneR | 6.69 | 5.73 | 5.42 | 5.39 | 7.01 | 6.11 | 5.62 | 5.74 | 10.08 | 10.01 | 11.04 | 9.99 | 22.89 | 22.45 | 20.93 | 22.39 |
|  | RF imp. | 7.34 | 6.05 | 5.45 | 5.80 | 7.48 | 6.21 | 5.77 | 5.96 | 11.61 | 11.11 | 11.11 | 10.83 | 23.69 | 23.18 | 20.98 | 22.84 |
| Outcome | All vars. | 7.20 | 6.14 | 5.14 | 5.77 | 7.66 | 6.50 | 5.35 | 6.35 | 11.70 | 10.97 | 10.95 | 11.04 | 24.43 | 23.02 | 20.70 | 22.62 |
|  | Boruta | 7.06 | 6.03 | 5.57 | 6.03 | 7.31 | 6.29 | 5.69 | 6.33 | 12.20 | 11.49 | 10.98 | 11.49 | 24.59 | 23.41 | 20.81 | 23.02 |
|  | Cfs | 6.87 | 6.13 | 6.06 | 5.92 | 6.90 | 6.26 | 6.06 | 6.13 | 10.68 | 10.13 | 11.48 | 9.95 | 22.64 | 21.74 | 21.38 | 21.40 |
|  | Chi | 6.13 | 5.64 | 5.40 | 5.58 | 6.37 | 5.94 | 5.57 | 5.87 | 10.84 | 10.40 | 11.24 | 10.49 | 22.26 | 21.59 | 21.16 | 21.46 |
|  | Gain | 6.59 | 5.89 | 5.38 | 5.97 | 6.72 | 6.07 | 5.63 | 6.17 | 11.16 | 10.81 | 11.06 | 10.97 | 22.39 | 21.92 | 21.03 | 21.93 |
|  | Lasso | 6.20 | 5.88 | 5.60 | 5.87 | 6.45 | 6.19 | 5.76 | 6.17 | 11.10 | 10.98 | 11.00 | 10.96 | 21.17 | 21.04 | 20.83 | 20.98 |
|  | StepWise | 7.60 | 6.47 | 6.15 | 6.37 | 7.71 | 6.76 | 6.40 | 6.66 | 12.36 | 11.59 | 12.67 | 11.06 | 24.60 | 23.52 | 23.06 | 22.86 |
|  | OneR | 5.78 | 5.56 | 5.53 | 5.54 | 6.08 | 5.86 | 5.77 | 5.87 | 10.76 | 10.63 | 11.05 | 10.59 | 21.61 | 21.34 | 20.92 | 21.13 |
|  | RF imp. | 7.00 | 6.02 | 5.46 | 5.48 | 7.51 | 6.46 | 5.66 | 6.15 | 11.68 | 11.02 | 11.01 | 10.80 | 24.18 | 22.81 | 20.79 | 22.28 |
|  |  | Ideological self-positioning scale (1-10) |  |  |  |  |  |  |  | Central gov. management ”poor” or ”very poor” |  |  |  |  |  |  |  |
| Exposure | All vars. | 2.83 | 2.74 | 2.39 | 2.55 | 2.13 | 2.15 | 1.86 | 2.05 | 0.87 | 1.28 | 1.98 | 2.34 | 5.88 | 5.93 | 4.99 | 6.61 |
|  | Boruta | 2.88 | 2.75 | 2.66 | 2.64 | 2.19 | 2.19 | 2.09 | 2.16 | 0.69 | 1.50 | 1.65 | 2.09 | 5.68 | 6.30 | 4.95 | 6.66 |
|  | Cfs | 2.73 | 2.63 | 2.39 | 2.62 | 2.11 | 2.06 | 1.86 | 2.06 | 1.80 | 2.32 | 1.97 | 2.39 | 5.88 | 6.81 | 5.05 | 6.89 |
|  | Chi | 2.77 | 2.66 | 2.40 | 2.60 | 2.08 | 2.05 | 1.86 | 2.04 | 0.93 | 1.68 | 1.98 | 2.00 | 5.77 | 6.62 | 5.10 | 7.00 |
|  | Gain | 2.71 | 2.62 | 2.40 | 2.63 | 2.14 | 2.08 | 1.86 | 2.08 | 2.23 | 2.63 | 1.95 | 2.59 | 5.87 | 6.76 | 5.01 | 6.70 |
|  | LASSO | 2.80 | 2.68 | 2.45 | 2.60 | 2.10 | 2.08 | 1.89 | 2.07 | 0.80 | 1.59 | 1.97 | 2.01 | 5.69 | 6.58 | 5.26 | 6.96 |
|  | StepWise | 2.78 | 2.67 | 2.51 | 2.61 | 2.08 | 2.07 | 1.94 | 2.07 | 0.91 | 1.70 | 1.76 | 2.11 | 5.79 | 6.70 | 5.05 | 7.08 |
|  | OneR | 2.74 | 2.65 | 2.42 | 2.60 | 2.05 | 2.03 | 1.88 | 2.02 | 0.85 | 1.44 | 2.06 | 1.81 | 5.72 | 6.24 | 5.20 | 6.61 |
|  | RF imp. | 2.76 | 2.65 | 2.41 | 2.58 | 2.09 | 2.06 | 1.87 | 2.03 | 1.23 | 1.80 | 1.97 | 2.22 | 5.81 | 6.57 | 5.08 | 6.95 |
| Outcome | All vars. | 2.83 | 2.73 | 2.39 | 2.57 | 2.13 | 2.16 | 1.86 | 2.08 | 0.87 | 1.29 | 1.98 | 2.24 | 5.88 | 5.90 | 4.98 | 6.29 |
|  | Boruta | 3.31 | 3.31 | 2.40 | 3.28 | 2.80 | 2.81 | 1.87 | 2.76 | 0.65 | 1.03 | 1.73 | 1.52 | 4.91 | 5.19 | 4.75 | 5.39 |
|  | Cfs | 2.53 | 2.53 | 2.39 | 2.53 | 1.95 | 1.96 | 1.86 | 1.95 | 0.45 | 0.01 | 1.29 | 0.27 | 3.52 | 3.76 | 4.40 | 3.96 |
|  | Chi | 3.23 | 3.06 | 2.79 | 2.99 | 2.67 | 2.51 | 2.29 | 2.45 | 0.32 | 0.50 | 1.57 | 0.78 | 4.11 | 4.13 | 4.60 | 4.35 |
|  | Gain | 3.32 | 3.30 | 2.43 | 3.29 | 2.81 | 2.79 | 1.90 | 2.77 | 0.30 | 0.55 | 1.49 | 0.62 | 3.78 | 3.94 | 4.55 | 4.02 |
|  | LASSO | 3.14 | 2.97 | 2.77 | 2.86 | 2.58 | 2.42 | 2.21 | 2.34 | 0.73 | 1.04 | 1.89 | 1.29 | 4.34 | 4.48 | 5.00 | 4.76 |
|  | StepWise | 3.00 | 2.86 | 2.50 | 2.69 | 2.42 | 2.32 | 1.93 | 2.18 | 0.45 | 0.85 | 1.78 | 1.34 | 4.57 | 4.87 | 4.96 | 5.32 |
|  | OneR | 3.27 | 3.13 | 2.83 | 3.06 | 2.74 | 2.60 | 2.35 | 2.53 | 0.50 | 0.75 | 1.96 | 0.94 | 4.21 | 4.40 | 5.03 | 4.53 |
|  | RF imp. | 3.19 | 3.09 | 2.65 | 3.00 | 2.62 | 2.55 | 2.14 | 2.48 | 1.06 | 1.32 | 2.01 | 1.86 | 5.56 | 5.47 | 5.04 | 5.72 |
|  |  | Preference for a state without autonomous comm. |  |  |  |  |  |  |  | Feels only Spanish |  |  |  |  |  |  |  |
| Exposure | All vars. | 12.48 | 11.82 | 10.16 | 11.02 | 8.93 | 8.58 | 7.62 | 8.38 | 10.99 | 11.34 | 10.56 | 10.94 | 9.54 | 10.19 | 9.53 | 9.88 |
|  | Boruta | 12.57 | 11.79 | 10.73 | 11.17 | 8.96 | 8.46 | 8.03 | 8.22 | 11.35 | 11.51 | 10.77 | 11.18 | 9.96 | 10.48 | 9.84 | 10.31 |
|  | Cfs | 12.32 | 11.72 | 10.28 | 11.68 | 9.17 | 8.42 | 7.68 | 8.40 | 11.46 | 11.96 | 10.59 | 11.89 | 10.48 | 11.23 | 9.53 | 11.23 |
|  | Chi | 12.47 | 11.69 | 10.22 | 11.38 | 8.86 | 8.19 | 7.59 | 8.12 | 11.55 | 11.76 | 10.53 | 11.58 | 10.12 | 10.71 | 9.50 | 10.75 |
|  | Gain | 12.30 | 11.77 | 10.24 | 11.88 | 9.37 | 8.57 | 7.66 | 8.70 | 11.25 | 11.98 | 10.57 | 11.95 | 10.58 | 11.38 | 9.53 | 11.35 |
|  | LASSO | 12.52 | 11.73 | 10.47 | 11.37 | 8.99 | 8.29 | 7.81 | 8.27 | 11.64 | 11.75 | 10.80 | 11.56 | 10.19 | 10.75 | 9.70 | 10.84 |
|  | StepWise | 12.53 | 11.80 | 10.66 | 11.41 | 8.97 | 8.36 | 7.98 | 8.27 | 11.48 | 11.75 | 10.85 | 11.57 | 10.08 | 10.76 | 9.82 | 10.82 |
|  | OneR | 12.15 | 11.51 | 10.36 | 11.26 | 8.55 | 8.05 | 7.75 | 8.10 | 11.38 | 11.58 | 10.71 | 11.52 | 9.94 | 10.46 | 9.61 | 10.54 |
|  | RF imp. | 12.35 | 11.67 | 10.30 | 11.37 | 8.86 | 8.23 | 7.66 | 8.18 | 11.41 | 11.78 | 10.57 | 11.69 | 10.20 | 10.87 | 9.54 | 11.05 |
| Outcome | All vars. | 12.48 | 11.81 | 10.18 | 11.01 | 8.93 | 8.52 | 7.65 | 8.32 | 10.99 | 11.26 | 10.55 | 10.80 | 9.54 | 10.13 | 9.53 | 9.74 |
|  | Boruta | 12.24 | 11.73 | 10.49 | 11.36 | 8.93 | 8.50 | 7.91 | 8.56 | 10.82 | 10.97 | 10.84 | 10.92 | 9.75 | 9.98 | 9.81 | 10.06 |
|  | Cfs | 11.02 | 10.93 | 10.54 | 10.94 | 8.29 | 8.19 | 7.93 | 8.21 | 10.58 | 10.53 | 10.64 | 10.50 | 9.64 | 9.60 | 9.59 | 9.54 |
|  | Chi | 10.70 | 10.69 | 10.28 | 10.67 | 8.01 | 8.00 | 7.70 | 7.97 | 10.61 | 10.61 | 10.58 | 10.61 | 9.74 | 9.73 | 9.54 | 9.73 |
|  | Gain | 10.82 | 10.76 | 10.34 | 10.73 | 8.11 | 8.04 | 7.72 | 8.03 | 10.49 | 10.48 | 10.64 | 10.39 | 9.58 | 9.57 | 9.64 | 9.45 |
|  | LASSO | 11.42 | 11.07 | 10.67 | 10.87 | 8.62 | 8.26 | 8.09 | 8.22 | 10.87 | 10.76 | 11.10 | 10.47 | 9.59 | 9.62 | 10.08 | 9.31 |
|  | StepWise | 12.65 | 11.88 | 11.29 | 11.67 | 9.52 | 8.84 | 8.52 | 8.93 | 11.18 | 11.13 | 11.46 | 10.67 | 9.90 | 10.01 | 10.43 | 9.41 |
|  | OneR | 11.43 | 11.26 | 10.26 | 11.15 | 8.63 | 8.51 | 7.67 | 8.53 | 10.94 | 10.97 | 10.72 | 10.83 | 9.82 | 9.91 | 9.72 | 9.83 |
|  | RF imp. | 10.83 | 10.71 | 10.30 | 10.65 | 8.08 | 8.04 | 7.73 | 7.95 | 10.60 | 10.61 | 10.60 | 10.67 | 9.74 | 9.73 | 9.57 | 9.79 |

Table 5: Mean relative bias of the estimates of population means in the real data simulation for each combination of methods. The closer to zero a value, the less biased the mean estimate.

|  |  | Raking = No |  |  |  | Raking = Yes |  |  |  | Raking = No |  |  |  | Raking = Yes |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN | GLM | GBM | kNN | NN |
|  |  | Econ. situation in Spain ”poor” or ”very poor” |  |  |  |  |  |  |  | Personal econ. situation ”poor” or ”very poor” |  |  |  |  |  |  |  |
| Exposure | Boruta | 0.983 | 0.995 | 1.100 | 0.993 | 0.970 | 0.984 | 1.112 | 0.981 | 0.962 | 0.990 | 1.041 | 0.989 | 0.966 | 0.995 | 1.070 | 1.006 |
|  | Cfs | 1.049 | 0.994 | 0.959 | 1.058 | 1.010 | 0.965 | 0.959 | 1.002 | 1.038 | 1.030 | 0.939 | 1.051 | 0.972 | 1.018 | 0.974 | 1.053 |
|  | Chi | 0.953 | 0.939 | 0.977 | 0.969 | 0.938 | 0.927 | 0.978 | 0.941 | 0.934 | 0.958 | 0.970 | 0.985 | 0.935 | 0.981 | 0.989 | 1.015 |
|  | Gain | 1.083 | 1.010 | 0.956 | 1.104 | 1.051 | 0.982 | 0.959 | 1.040 | 1.088 | 1.067 | 0.928 | 1.101 | 0.986 | 1.036 | 0.962 | 1.081 |
|  | Lasso | 0.975 | 0.964 | 1.032 | 0.992 | 0.951 | 0.941 | 1.033 | 0.948 | 0.953 | 0.983 | 0.994 | 1.027 | 0.961 | 1.002 | 1.034 | 1.039 |
|  | StepWise | 1.000 | 0.984 | 1.112 | 0.993 | 0.973 | 0.957 | 1.137 | 0.965 | 0.951 | 0.988 | 1.059 | 0.984 | 0.957 | 1.006 | 1.098 | 1.018 |
|  | OneR | 0.908 | 0.919 | 0.972 | 0.972 | 0.891 | 0.900 | 0.963 | 0.922 | 0.884 | 0.917 | 0.983 | 0.940 | 0.893 | 0.939 | 0.998 | 0.969 |
|  | RF imp. | 0.973 | 0.951 | 0.989 | 1.010 | 0.958 | 0.936 | 0.998 | 0.958 | 0.963 | 0.984 | 0.955 | 1.005 | 0.940 | 0.991 | 0.988 | 1.010 |
| Outcome | Boruta | 0.984 | 0.990 | 1.023 | 1.004 | 0.980 | 0.973 | 1.017 | 0.988 | 0.988 | 1.011 | 1.045 | 0.998 | 1.002 | 1.017 | 1.029 | 1.015 |
|  | Cfs | 0.953 | 1.015 | 1.095 | 1.034 | 0.942 | 0.999 | 1.080 | 0.990 | 0.947 | 0.956 | 1.063 | 0.996 | 0.895 | 0.918 | 1.054 | 0.938 |
|  | Chi | 0.879 | 0.958 | 0.984 | 0.947 | 0.873 | 0.949 | 0.982 | 0.921 | 0.936 | 0.960 | 1.039 | 0.989 | 0.878 | 0.918 | 1.035 | 0.934 |
|  | Gain | 0.897 | 0.964 | 0.959 | 0.989 | 0.900 | 0.956 | 0.965 | 0.953 | 0.942 | 0.985 | 0.959 | 0.989 | 0.873 | 0.929 | 0.991 | 0.944 |
|  | Lasso | 0.844 | 0.936 | 0.999 | 0.951 | 0.879 | 0.943 | 0.985 | 0.934 | 0.889 | 0.928 | 0.940 | 0.926 | 0.796 | 0.858 | 0.975 | 0.865 |
|  | StepWise | 1.048 | 1.040 | 1.180 | 1.059 | 1.032 | 1.058 | 1.182 | 1.029 | 1.021 | 1.018 | 1.110 | 1.038 | 0.995 | 1.017 | 1.150 | 1.011 |
|  | OneR | 0.808 | 0.914 | 0.993 | 0.945 | 0.809 | 0.894 | 0.997 | 0.886 | 0.891 | 0.931 | 0.953 | 0.918 | 0.826 | 0.884 | 0.983 | 0.881 |
|  | RF imp. | 0.979 | 0.991 | 1.062 | 1.007 | 0.982 | 0.992 | 1.069 | 1.006 | 0.994 | 0.991 | 1.051 | 1.012 | 0.987 | 0.986 | 1.026 | 0.993 |
|  |  | Ideological self-positioning scale (1-10) |  |  |  |  |  |  |  | Central gov. management ”poor” or ”very poor” |  |  |  |  |  |  |  |
| Exposure | Boruta | 1.018 | 1.006 | 1.164 | 1.042 | 1.011 | 1.026 | 1.162 | 1.059 | 0.982 | 0.989 | 0.932 | 1.028 | 0.962 | 1.014 | 0.941 | 1.021 |
|  | Cfs | 0.890 | 0.901 | 0.966 | 1.006 | 0.898 | 0.887 | 0.939 | 0.940 | 1.016 | 1.005 | 0.910 | 1.016 | 0.938 | 1.023 | 0.894 | 0.992 |
|  | Chi | 0.927 | 0.929 | 0.981 | 1.003 | 0.890 | 0.885 | 0.961 | 0.932 | 0.948 | 0.965 | 0.905 | 1.003 | 0.927 | 1.021 | 0.895 | 1.033 |
|  | Gain | 0.882 | 0.897 | 0.968 | 1.011 | 0.915 | 0.895 | 0.945 | 0.949 | 1.038 | 1.019 | 0.911 | 1.024 | 0.940 | 1.016 | 0.895 | 0.963 |
|  | Lasso | 0.951 | 0.944 | 1.013 | 1.010 | 0.924 | 0.923 | 0.984 | 0.970 | 0.950 | 0.963 | 0.944 | 0.973 | 0.923 | 1.013 | 0.945 | 1.008 |
|  | StepWise | 0.939 | 0.941 | 1.068 | 1.014 | 0.906 | 0.916 | 1.025 | 0.964 | 0.948 | 0.959 | 0.922 | 0.985 | 0.927 | 1.014 | 0.917 | 1.009 |
|  | OneR | 0.915 | 0.920 | 0.997 | 0.993 | 0.884 | 0.886 | 0.975 | 0.919 | 0.951 | 0.953 | 0.926 | 0.991 | 0.925 | 0.976 | 0.928 | 0.986 |
|  | RF imp. | 0.917 | 0.920 | 0.985 | 0.990 | 0.895 | 0.891 | 0.965 | 0.924 | 0.983 | 0.993 | 0.908 | 1.000 | 0.933 | 1.015 | 0.913 | 1.012 |
| Outcome | Boruta | 1.265 | 1.366 | 0.967 | 1.486 | 1.414 | 1.451 | 0.944 | 1.488 | 0.969 | 1.006 | 0.951 | 0.948 | 0.880 | 0.936 | 0.946 | 0.904 |
|  | Cfs | 0.796 | 0.855 | 0.965 | 0.947 | 0.817 | 0.832 | 0.941 | 0.870 | 0.968 | 0.968 | 0.924 | 0.931 | 0.806 | 0.830 | 0.919 | 0.784 |
|  | Chi | 1.225 | 1.211 | 1.233 | 1.300 | 1.348 | 1.251 | 1.260 | 1.281 | 0.971 | 0.953 | 0.916 | 0.919 | 0.818 | 0.829 | 0.888 | 0.831 |
|  | Gain | 1.272 | 1.358 | 0.992 | 1.487 | 1.427 | 1.441 | 0.975 | 1.486 | 0.977 | 0.956 | 0.886 | 0.935 | 0.806 | 0.816 | 0.866 | 0.790 |
|  | Lasso | 1.182 | 1.157 | 1.236 | 1.228 | 1.295 | 1.196 | 1.229 | 1.211 | 0.964 | 0.954 | 0.945 | 0.922 | 0.840 | 0.862 | 0.935 | 0.849 |
|  | StepWise | 1.100 | 1.085 | 1.095 | 1.104 | 1.181 | 1.113 | 1.074 | 1.109 | 0.980 | 0.974 | 1.027 | 0.955 | 0.880 | 0.925 | 0.984 | 0.896 |
|  | OneR | 1.247 | 1.253 | 1.258 | 1.342 | 1.384 | 1.314 | 1.300 | 1.332 | 0.994 | 0.975 | 0.943 | 0.933 | 0.827 | 0.847 | 0.926 | 0.812 |
|  | RF imp. | 1.204 | 1.228 | 1.152 | 1.307 | 1.319 | 1.283 | 1.165 | 1.286 | 1.017 | 0.994 | 0.988 | 0.957 | 0.970 | 0.965 | 0.976 | 0.912 |
|  |  | Preference for a state without autonomous comm. |  |  |  |  |  |  |  | Feels only Spanish |  |  |  |  |  |  |  |
| Exposure | Boruta | 0.997 | 0.981 | 1.074 | 1.015 | 0.978 | 0.955 | 1.053 | 0.982 | 1.012 | 0.990 | 1.026 | 1.025 | 1.009 | 1.003 | 1.020 | 1.073 |
|  | Cfs | 0.950 | 0.951 | 0.970 | 1.048 | 0.976 | 0.922 | 0.929 | 0.940 | 1.011 | 1.023 | 0.964 | 1.078 | 1.046 | 1.063 | 0.950 | 1.103 |
|  | Chi | 0.977 | 0.959 | 0.978 | 1.024 | 0.952 | 0.914 | 0.940 | 0.944 | 1.033 | 1.013 | 0.960 | 1.055 | 1.005 | 1.000 | 0.948 | 1.067 |
|  | Gain | 0.947 | 0.955 | 0.964 | 1.069 | 0.999 | 0.934 | 0.925 | 0.966 | 0.982 | 1.020 | 0.961 | 1.081 | 1.054 | 1.077 | 0.950 | 1.111 |
|  | Lasso | 0.980 | 0.964 | 1.008 | 1.017 | 0.964 | 0.923 | 0.968 | 0.939 | 1.039 | 1.010 | 0.990 | 1.058 | 1.015 | 1.020 | 0.972 | 1.092 |
|  | StepWise | 0.979 | 0.970 | 1.041 | 1.024 | 0.951 | 0.918 | 0.996 | 0.941 | 1.020 | 1.014 | 0.995 | 1.053 | 1.002 | 1.018 | 0.989 | 1.090 |
|  | OneR | 0.945 | 0.939 | 1.004 | 1.002 | 0.922 | 0.893 | 0.974 | 0.916 | 1.020 | 0.995 | 0.989 | 1.045 | 0.996 | 0.988 | 0.967 | 1.054 |
|  | RF imp. | 0.960 | 0.953 | 0.981 | 1.015 | 0.957 | 0.915 | 0.942 | 0.926 | 1.016 | 1.014 | 0.978 | 1.057 | 1.025 | 1.034 | 0.967 | 1.092 |
| Outcome | Boruta | 0.972 | 0.990 | 1.018 | 1.029 | 0.982 | 0.988 | 0.990 | 0.984 | 0.980 | 0.985 | 1.029 | 1.023 | 1.012 | 1.006 | 1.022 | 1.033 |
|  | Cfs | 0.830 | 0.894 | 1.014 | 0.986 | 0.885 | 0.932 | 0.976 | 0.937 | 0.945 | 0.926 | 0.971 | 0.955 | 0.996 | 0.943 | 0.958 | 0.959 |
|  | Chi | 0.790 | 0.858 | 0.971 | 0.938 | 0.849 | 0.901 | 0.934 | 0.894 | 0.949 | 0.931 | 0.962 | 0.968 | 1.005 | 0.952 | 0.951 | 0.971 |
|  | Gain | 0.798 | 0.863 | 0.980 | 0.947 | 0.851 | 0.898 | 0.934 | 0.902 | 0.938 | 0.923 | 0.971 | 0.950 | 0.993 | 0.947 | 0.964 | 0.954 |
|  | Lasso | 0.861 | 0.899 | 1.036 | 0.962 | 0.919 | 0.941 | 0.996 | 0.920 | 0.974 | 0.945 | 1.064 | 0.958 | 0.995 | 0.944 | 1.056 | 0.948 |
|  | StepWise | 1.015 | 1.009 | 1.151 | 1.087 | 1.056 | 1.038 | 1.101 | 1.054 | 1.014 | 0.985 | 1.126 | 0.989 | 1.031 | 0.985 | 1.129 | 0.973 |
|  | OneR | 0.871 | 0.921 | 0.990 | 0.995 | 0.939 | 0.969 | 0.952 | 0.975 | 0.964 | 0.952 | 1.020 | 0.970 | 0.981 | 0.947 | 1.018 | 0.965 |
|  | RF imp. | 0.805 | 0.862 | 0.969 | 0.933 | 0.859 | 0.909 | 0.932 | 0.890 | 0.948 | 0.931 | 0.966 | 0.973 | 1.005 | 0.953 | 0.954 | 0.976 |

Table 6: Efficiency of the estimates of population means in the real data simulation for each combination of methods. Values greater than one indicate inefficiency, while values below one suggest that the use of a given variable selection method provides more efficient estimates than the case in which all variables are used.

# 6 Application study

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

The probability sample used as the reference sample was obtained from a survey conducted in 2015,targeting students at the University of Granada (UGR), Spain. The sample was composed of $n_r = 856$ respondents, recruited in face-to-face interviews under a three-stage cluster sampling design, which produced an estimated sampling error of $\pm 3.3\%$ in the case of $p = q = 0.5$ with a confidence level of 95%. The survey questionnaire included screening instruments for abuse and dependence, namely the Spanish Mobile Phone Abuse Questionnaire (ATeMo) (Olivencia-Carrión et al. 2018), the Cannabis Abuse Screening Test (CAST)(Legleye et al. 2007) and the Severity of Dependence Scale (SDS) (Gossop et al. 1995), together with subscales regarding internet and videogames addiction from the MULTICAGE-CAD4 instrument (Pedrero Pérez et al. 2007). The survey also recorded the age, gender and university faculty of each participant.

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

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

- Mean score on the total ATeMo scale, which was 30.066 in the reference sample and 32.558 in the unweighted convenience sample.

- Mean score on the item ”I have tried to spend less time using my mobile phone but I cannot do it” (number 16 in the ATeMo instrument). The mean score of this item in the reference sample was 0.776 while in the unweighted convenience sample it was 1.217, this being the greatest difference observed in in any ATeMo item between the reference sample and the convenience sample.

Table 7 shows the distributions of the covariates available for PSA in both samples. Except for gender, the values differ greatly in the distribution of the covariates between the two samples. Overall, respondents to the online sample were younger and more prone to cannabis consumption. In addition, their score for the MULTICAGE subscales of internet and videogames addiction tended to be higher than those of the reference sample members. Finally, the Science Faculty at the UGR was clearly overrepresented in the online sample, as to a lesser extent was the Medicine Faculty, while the other faculties were underrepresented. Given that the variability between samples can be identified in the covariates, it seems likely that PSA balanced the online sample more efficiently.

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

[a] Standard deviation of the age
[b] Two sample test for equality of proportions with continuity correction
[c] Welch two sample t-test
[d] Pearson's Chi-squared test

Table 7: Distributions of covariates in online and reference samples

Estimation of the population means followed the same procedure as described in Section 4.3: each algorithm for variable selection was applied before PSA (with the same predictive models -and hyperparameter optimisation- as in the simulations: logistic regression, GBM, k-NN and neural networks) using the described reference and convenience samples, and the resulting weights were used directly in the estimators or as initial weights for Raking calibration. The estimated population means for each combination of methods and the estimated Leave-One-Out jackknife variance (Quenouille 1956) are shown in Tables 8 and 9 respectively.

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

| | | Estimation of pop. mean for Item number 16 | | | | Estimation of pop. mean for ATeMo score | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Raking | | Raking | | No Raking | | Raking | |
| | | Exposure | Outcome | Exposure | Outcome | Exposure | Outcome | Exposure | Outcome |
| Convenience sample | | 1.217 | | | | 32.62 | | | |
| Reference sample | | 0.776 | | | | 30.07 | | | |
| Log. reg. | All vars. | 1.424 | 1.424 | 0.816 | 0.816 | 30.32 | 30.32 | 30.64 | 30.64 |
| | Stepwise | 1.411 | 1.105 | 0.819 | 0.812 | 30.43 | 30.10 | 30.58 | 30.20 |
| | CFS | 1.329 | 1.105 | 0.810 | 0.812 | 31.65 | 30.10 | 31.86 | 30.20 |
| | Chi-sq. | 1.329 | 1.329 | 0.810 | 0.810 | 31.65 | 32.00 | 31.86 | 31.83 |
| | Gain r. | 1.271 | 0.972 | 0.810 | 0.862 | 33.62 | 31.18 | 31.96 | 31.66 |
| | OneR | 1.262 | 1.329 | 0.827 | 0.810 | 33.59 | 32.00 | 32.59 | 31.83 |
| | RF imp. | 0.972 | 1.105 | 0.862 | 0.812 | 31.18 | 30.10 | 31.66 | 30.20 |
| | Boruta | 1.328 | 1.222 | 0.770 | 0.760 | 30.23 | 30.22 | 30.86 | 30.34 |
| | LASSO | 0.939 | 1.217 | 0.805 | 0.862 | 28.93 | 30.10 | 29.94 | 30.20 |
| GBM | All vars. | 1.093 | 1.242 | 0.768 | 0.797 | 29.98 | 31.19 | 30.84 | 31.12 |
| | Stepwise | 1.190 | 1.143 | 0.776 | 0.858 | 30.37 | 30.42 | 30.70 | 30.05 |
| | CFS | 1.129 | 1.151 | 0.772 | 0.864 | 31.95 | 30.05 | 31.68 | 29.97 |
| | Chi-sq. | 1.243 | 1.222 | 0.777 | 0.801 | 31.36 | 31.32 | 31.42 | 31.15 |
| | Gain r. | 1.267 | 1.000 | 0.821 | 0.862 | 33.58 | 31.26 | 32.08 | 31.66 |
| | OneR | 1.259 | 1.168 | 0.832 | 0.797 | 33.54 | 32.31 | 32.61 | 31.46 |
| | RF imp. | 0.995 | 1.144 | 0.862 | 0.854 | 31.32 | 30.25 | 31.66 | 29.97 |
| | Boruta | 1.300 | 1.295 | 0.831 | 0.797 | 29.62 | 29.89 | 30.55 | 30.36 |
| | LASSO | 1.017 | 1.217 | 0.836 | 0.862 | 29.09 | 30.37 | 30.11 | 30.11 |
| k-NN | All vars. | 1.192 | 1.192 | 0.860 | 0.860 | 32.50 | 32.50 | 31.72 | 31.72 |
| | Stepwise | 0.860 | 1.217 | 0.891 | 0.862 | 27.76 | 32.62 | 31.60 | 31.66 |
| | CFS | 1.140 | 1.217 | 0.823 | 0.862 | 31.55 | 32.62 | 30.91 | 31.66 |
| | Chi-sq. | 1.140 | 1.140 | 0.823 | 0.823 | 31.55 | 32.40 | 30.91 | 31.54 |
| | Gain r. | 1.201 | 1.217 | 0.837 | 0.862 | 32.46 | 32.62 | 31.51 | 31.66 |
| | OneR | 1.217 | 1.140 | 0.862 | 0.823 | 32.62 | 32.40 | 31.66 | 31.54 |
| | RF imp. | 1.217 | 1.217 | 0.862 | 0.862 | 32.62 | 32.62 | 31.66 | 31.66 |
| | Boruta | 1.191 | 1.217 | 0.855 | 0.862 | 32.53 | 32.63 | 31.67 | 31.51 |
| | LASSO | 1.276 | 1.217 | 0.899 | 0.862 | 33.51 | 32.62 | 32.44 | 31.66 |
| Neural nets | All vars. | 1.200 | 1.249 | 0.808 | 0.787 | 30.81 | 35.99 | 31.09 | 33.36 |
| | Stepwise | 1.452 | 1.142 | 0.838 | 0.841 | 32.33 | 30.54 | 32.78 | 30.16 |
| | CFS | 1.237 | 1.142 | 0.778 | 0.841 | 31.59 | 30.54 | 31.37 | 30.16 |
| | Chi-sq. | 1.216 | 1.249 | 0.825 | 0.811 | 31.40 | 32.08 | 31.58 | 32.12 |
| | Gain r. | 1.263 | 0.992 | 0.835 | 0.862 | 33.55 | 31.32 | 32.33 | 31.66 |
| | OneR | 1.257 | 1.216 | 0.837 | 0.788 | 33.51 | 32.06 | 32.61 | 32.15 |
| | RF imp. | 0.992 | 1.142 | 0.862 | 0.841 | 31.32 | 30.54 | 31.66 | 30.16 |
| | Boruta | 1.269 | 1.315 | 0.765 | 0.837 | 31.19 | 30.18 | 31.19 | 30.48 |
| | LASSO | 0.945 | 1.217 | 0.809 | 0.862 | 29.54 | 30.54 | 30.39 | 30.15 |

Table 8: Mean estimates of the population mean for both target variables after applying of each combination of adjustments.

| | | Estimation of pop. mean for Item number 16 | | | | Estimation of pop. mean for ATeMo score | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Raking | | Raking | | No Raking | | Raking | |
| | | Exposure | Outcome | Exposure | Outcome | Exposure | Outcome | Exposure | Outcome |
| Convenience sample | | 0.010 | | | | 1.51 | | | |
| Log. regr. | All vars. | 0.171 | 0.171 | 0.036 | 0.036 | 14.12 | 14.12 | 6.71 | 6.71 |
| | Stepwise | 0.179 | 0.009 | 0.036 | 0.022 | 14.03 | 1.24 | 6.43 | 6.14 |
| | CFS | 0.125 | 0.009 | 0.034 | 0.022 | 11.76 | 1.24 | 5.38 | 6.14 |
| | Chi-sq. | 0.125 | 0.125 | 0.034 | 0.034 | 11.76 | 10.07 | 5.38 | 5.51 |
| | Gain r. | 0.026 | 0.022 | 0.034 | 0.024 | 4.26 | 3.28 | 5.40 | 5.08 |
| | OneR | 0.025 | 0.125 | 0.036 | 0.034 | 4.16 | 10.07 | 5.17 | 5.51 |
| | RF imp. | 0.022 | 0.009 | 0.024 | 0.022 | 3.28 | 1.24 | 5.08 | 6.14 |
| | Boruta | 0.116 | 0.026 | 0.030 | 0.031 | 14.45 | 1.73 | 6.69 | 6.73 |
| | LASSO | 0.021 | 0.010 | 0.022 | 0.024 | 4.00 | 1.24 | 6.48 | 6.14 |
| GBM | All vars. | 0.077 | 0.060 | 0.027 | 0.027 | 8.21 | 9.88 | 6.57 | 6.08 |
| | Stepwise | 0.032 | 0.009 | 0.025 | 0.024 | 3.71 | 1.62 | 5.98 | 7.46 |
| | CFS | 0.086 | 0.009 | 0.029 | 0.024 | 8.11 | 1.72 | 5.86 | 7.69 |
| | Chi-sq. | 0.051 | 0.032 | 0.028 | 0.026 | 4.81 | 3.88 | 5.45 | 4.87 |
| | Gain r. | 0.024 | 0.021 | 0.034 | 0.024 | 4.02 | 3.08 | 5.22 | 5.08 |
| | OneR | 0.024 | 0.033 | 0.037 | 0.025 | 4.02 | 3.64 | 5.26 | 5.32 |
| | RF imp. | 0.021 | 0.009 | 0.024 | 0.024 | 3.13 | 1.71 | 5.08 | 7.65 |
| | Boruta | 0.050 | 0.027 | 0.026 | 0.031 | 6.00 | 1.78 | 6.26 | 6.45 |
| | LASSO | 0.019 | 0.010 | 0.023 | 0.024 | 3.84 | 1.56 | 6.81 | 7.24 |
| k-NN | All vars. | 0.012 | 0.012 | 0.024 | 0.024 | 1.93 | 1.93 | 5.62 | 5.62 |
| | Stepwise | 0.012 | 0.010 | 0.028 | 0.024 | 2.48 | 1.51 | 6.17 | 5.08 |
| | CFS | 0.010 | 0.010 | 0.022 | 0.024 | 1.60 | 1.51 | 5.39 | 5.08 |
| | Chi-sq. | 0.010 | 0.099 | 0.022 | 0.022 | 1.60 | 1.59 | 5.39 | 5.47 |
| | Gain r. | 0.011 | 0.010 | 0.024 | 0.024 | 1.49 | 1.51 | 4.80 | 5.08 |
| | OneR | 0.010 | 0.010 | 0.026 | 0.022 | 1.56 | 1.59 | 5.28 | 5.47 |
| | RF imp. | 0.010 | 0.010 | 0.024 | 0.024 | 1.51 | 1.51 | 5.08 | 5.08 |
| | Boruta | 0.012 | 0.011 | 0.024 | 0.025 | 1.81 | 3.41 | 5.61 | 6.11 |
| | LASSO | 0.016 | 0.010 | 0.027 | 0.024 | 3.29 | 1.51 | 5.32 | 5.08 |
| Neural nets | All vars. | 0.101 | 0.129 | 0.029 | 0.030 | 8.89 | 15.04 | 5.15 | 5.69 |
| | Stepwise | 0.061 | 0.009 | 0.030 | 0.023 | 11.61 | 1.51 | 6.50 | 6.97 |
| | CFS | 0.125 | 0.009 | 0.029 | 0.023 | 10.16 | 1.51 | 5.16 | 6.97 |
| | Chi-sq. | 0.091 | 0.067 | 0.028 | 0.031 | 7.85 | 7.61 | 6.28 | 5.37 |
| | Gain r. | 0.023 | 0.021 | 0.035 | 0.024 | 3.93 | 3.13 | 5.29 | 5.08 |
| | OneR | 0.023 | 0.091 | 0.035 | 0.033 | 3.81 | 8.03 | 5.14 | 5.62 |
| | RF imp. | 0.021 | 0.009 | 0.024 | 0.023 | 3.13 | 1.51 | 5.08 | 6.97 |
| | Boruta | 0.121 | 0.025 | 0.030 | 0.038 | 10.32 | 1.54 | 5.72 | 7.22 |
| | LASSO | 0.019 | 0.010 | 0.022 | 0.024 | 3.02 | 1.51 | 5.35 | 6.97 |

Table 9: Estimated variance of the estimators, obtained via Leave-One-Out jackknife, after applying each combination of methods.

# 7 Discussion and conclusions

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

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

Our analysis shows that feature selection makes a significant contribution to reducing relative bias. However, the best feature selection algorithm, in this respect and regarding efficiency, varies according to the dataset considered and the adjustment choices made. As observed by Bommert et al. (2020), the best variable selection method depends on the dataset, meaning there is no one-size-fits-all solution. However, the reduction of model complexity associated with variable selection consistently produced more efficient estimators. As expected, selecting variables according to their impact on the outcome variable provided the best results overall. In line with Austin and Stuart (2015), we find that the propensity score balances the covariates included in the model, so it is preferable to include prognostically important variables (related to the outcome) as the probability to balance the target variable will also be higher. In view of these results, in practice the combination of several variable selection approaches, rather than just one, might be useful to identify the best subset in each situation.

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

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

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

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

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

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

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

[4] Bethlehem J (2010) Selection Bias in Web Surveys. Int Stat Rev 78(2):161-188.

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

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

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

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

[9] Breiman L (2001) Random forests. Mach Learn 45(1):5-32.

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

[11] Buelens B, Burger J, van den Brakel, JA (2018) Comparing Inference Methods for Non-probability Samples. Int Stat Rev, 86(2):322–343.

[12] Buskirk TD, Kolenikov S (2015) Finding respondents in the forest: a comparison of logistic regression and random forest models for response propensity weighting and stratification. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach.

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

[14] Castro-Martín L, Rueda MM, Ferri-García R (2020b) Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques. Mathematics 8(6):879.

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

[16] Cochran WG (1968) The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. Biometrics 24(2):295-313.

[17] Couper M (2000) Web Surveys: A Review of Issues and Approaches. Public Opin Quart 64(4):464-494.

[18] Couper M, Kapteyn A, Schonlau M, Winter J (2007) Noncoverage and Nonresponse in an Internet Survey. Soc Sci Res, 36:131-148.

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

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

[21] Elliott MR, Valliant R (2017) Inference for Nonprobability Samples. Stat Sci, 32(2):249-264.

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

[23] Ferri-García R, Castro-Martín L, Rueda MM (2020) Evaluating machine learning methods for estimation in online surveys with superpopulation modeling. Math Comput Simulat (in press).

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

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

[26] Hall MA (1999) Correlation-based Feature Selection for Machine Learning. Dissertation, University of Waikato, Department of Computer Science.

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

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

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

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

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

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

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

[34] National Institute of Statistics (2018) Survey on Equipment and Use of Information and Communication Technologies in Households. http://www.ine.es/prensa/tich_2018.pdf. Accessed 19 January 2020.

[35] Marken S (2018) Still Listening: The State of Telephone Surveys. https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx. Accessed 21 January 2020.

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

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

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

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

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

[41] Quenouille MH (1956) Notes on bias in estimation. Biometrika 43(3/4):353-360.

[42] Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81-106.

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

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

[45] Rosenbaum PR, Rubin DB (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70(1):41-55.

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

[47] Rueda MM (2019) Comments on: Deville and Särndal's calibration: Revisiting a 25 years old successful optimization problem. Test 28(4):1077-1081.

[48] Rueda M, Martínez S, Martínez H, Arcos A (2006) Mean estimation with calibration techniques in presence of missing data. Comput Stat Data An, 50(11):3263-3277.

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

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

[51] Schonlau M, Couper M (2017) Options for Conducting Web Surveys. Stat Sci 32(2):279-292.

[52] Spanish Center for Sociological Research (2019) January Barometer (study number 3238). http://www.cis.es/cis/opencm/EN/1_encuestas/estudios/ver.jsp?estudio=14442. Accessed 18 January 2020.

[53] Taylor H (2000) Does Internet research work? Int J Market Res 42(1):51-63.

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

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

[56] Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58(1):267-288.

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

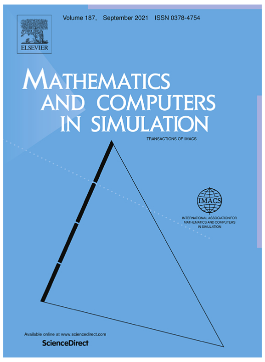
[58] Valliant R (2020) Comparing Alternatives for Estimation from Nonprobability Samples. J Surv Stat Methodol 8(2):231-263.

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

# Appendix A5

## Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling

Ferri-García, Ramón; Castro-Martín, Luis; Rueda, María del Mar (2021)
Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling
*Mathematics and Computers in Simulation*, vol. 186, pp. 19-28
DOI: 10.1016/j.matcom.2020.03.005


Volume 187, September 2021 ISSN 0378-4754
ELSEVIER
MATHEMATICS AND COMPUTERS IN SIMULATION
TRANSACTIONS OF IMACS
IMACS
INTERNATIONAL ASSOCIATION FOR MATHEMATICS AND COMPUTERS IN SIMULATION
Available online at www.sciencedirect.com
ScienceDirect

| MATHEMATICS, APPLIED | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 1.620 | 68/261 | Q2 |

**Abstract**

Online surveys, despite their cost and effort advantages, are particularly prone to selection bias due to the differences between target population and potentially covered population (online population). This leads to the unreliability of estimates coming from online samples unless further adjustments are applied. Some techniques have arisen in the last years regarding this issue, among which superpopulation modeling can be useful in Big Data context where censuses are accesible. This technique 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. The modeling step has been usually done with linear regression or LASSO models, but machine learning (ML) algorithms has been pointed out as promising alternatives. In this study we examine the use of these algorithms in the online survey context, in order to evaluate and compare their performance and adequacy to the problem. A simulation study shows that ML algorithms can effectively volunteering bias to a greater extent than traditional methods in several scenarios.

# 1 Introduction

Online surveys have become one of the most used modes of survey administration worldwide. They are a powerful tool for recruiting respondents fast and effortlessly with small costs in comparison to traditional survey administration modes. However, samples from online surveys are usually collected using a nonprobabilistic scheme, given that access to all members of the target population is not guaranteed in most cases and the inclusion probability cannot be obtained because of the absence of a sampling frame. As a result, selection bias derived from this procedure, defined by [8] as the presence of a substantial difference between observed and unobserved population, makes survey estimates not valid for inference [24].

Different inference procedures are proposed in the literature to correct for selection bias induced by non-random selection mechanisms. There are three important approaches: the pseudo-design based inference (or pseudo-randomization [6]), 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 consists of estimating response probabilities and using them in Horvitz-Thompson type estimators to account for unequal selection probabilities. The most used method to estimate the response probabilities is propensity scoring proposed by [21] (see e.g. [14]). This method uses a probability reference sample to construct a propensity model for the non-probability sample. Sample matching is another approach also applied to reduce selection bias in non-probability samples by combining them with a probability sample.

In this paper, we consider the situation where there is only a non-probability sample available for measuring the target information, in addition to some auxiliary information of the full population of interest, and we consider several predictive inference methods. Predictive methods are based on superpopulation models. In this approach, a statistical model is fitted for the analysis variable $y$ 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) let us use auxiliary information about covariates on different methods for predicting the unknown values. The objective of this study is to examine the use of Machine Learning algorithms in the online survey context, to evaluate and compare their performance and adequacy to the problem. A simulation study is performed for that matter.

## 2 Predictive inference for non-probability samples

Let $s$ be the online sample, $\bar{s}$ the population not included in the sample, and $U$ the complete target population so $s \cup \bar{s} = U$. The goal is to estimate the population parameter of a target variable, $y$, which has been measured in $s$ but it is not available in data from $\bar{s}$.

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

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

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

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

$$\hat{y}_i = E_m(y_i|\mathbf{x}_i) \tag{2}$$

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

If we want to estimate the total of $y$, $\overline{Y}$, we can use the auxiliary information in several ways and we can define several estimators:

- the model-based estimator:

$$\hat{\overline{Y}}_m = \frac{1}{N}\left(\sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i\right) \tag{3}$$

- the model-assisted estimator:

$$\hat{\overline{Y}}_{ma} = \frac{1}{N}\left(\sum_{i \in U} \hat{y}_i + \sum_{i \in s} (y_i - \hat{y}_i) w_i\right) \tag{4}$$

being $w_i$ a weight of the unit $i$ (set by the researcher to adjust the lack of response, lack of coverage, voluntariness, ... usually employing post-stratification).

- the model-calibrated estimator:

$$\hat{\bar{Y}}_{cal} = \frac{1}{N} \sum_{k \in s} y_k w_k^{CAL} \tag{5}$$

where $w_k^{CAL}$ are such that they minimize $\sum_{k \in s} G\left(w_k^{CAL}, w_k\right)$, where $G(\cdot, \cdot)$ is a particular distance function, subject to $\sum_{k \in s} w_k^{CAL} \hat{y}_i = \sum_{k \in U} \hat{y}_i$.

# 3 Machine learning techniques in superpopulation modelling

Usually the linear regression model is considered for estimation, $E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta$, and the above estimators can be rewritten as a type of regression estimators. Alternatively to the linear regression model, Machine Learning (ML) methods have been proposed for the estimation of the nonsampled population values. In situations where additivity and/or linearity do not hold, ML algorithms are more suitable for regression and classification. Some of these algorithms, such as decision trees and related (Random Forests, Gradient Boosting Machines) can also take interactions into account without the need of specifying the terms. The use of some ML algorithms for probabilistic samples has been studied in the last few years for deriving model-assisted estimators ([15]; [1]; [23]; [25]; [4]). In this section, we consider some of the most important ML algorithms that can be used to define model-assisted, model-based and model-calibrated estimators for a non-probability sample.

## 3.1 Advanced linear regression models

$\beta$ coefficients of a linear regression estimated by ordinary least squares are estimated as $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. However, as [10] stated, this estimation becomes sensitive when $\mathbf{X}'\mathbf{X}$ is far from being a unit matrix (i. e. multicollinearity is present in covariates). In such a case, ridge regression can be an alternative. It estimates regression coefficients adding an identity term to control instability, $\beta = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$, where $k \geq 0$ is a coefficient which depends on (unknown) real regression parameters and therefore has to be chosen arbitrarily or via hyperparameter tuning. From a Bayesian point of view, the resulting $\beta$ can be considered the posterior mean of a prior Normal distribution with zero mean and a variance of $\mathbf{I}\sigma^2/k$ as described in [11]. Gibbs sampling can provide Bayesian estimates for $\beta$ in such a case.

An alternative to ridge regression is the Least Absolute Shrinkage and Selection Operator (LASSO) regression, described in [22], where coefficients are estimated through minimizing the least-squares with a penalty parameter, $\alpha$, subject to a restriction on a tuning parameter:

$$\begin{array}{l} argmin \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \\ subject\ to \sum_j |\beta_j| \leq t \end{array} \tag{6}$$

The restriction $t$ is fixed to allow shrinkage of the solutions towards zero, allowing some coefficients to be equal to zero. As a consequence, this approach performs variable selection, in contrast to ridge regression where coefficients are always different from zero. LASSO estimates can be seen as posterior estimates of the distribution mode of prior Laplace independent and identical distributions. Therefore, Bayesian procedures can be used for LASSO estimation as described in [19].

### 3.2 Bagged Trees

Estimating the expectance $E_m(y_i|\mathbf{x}_i)$ under decision tree modeling results in the following expression:

$$E_m(y_i|\mathbf{x}_i) = \begin{cases} \overline{y(s^{J_1})} & \{i \in s/\mathbf{x}_i \in J_1\} \\ ... & ... \\ \overline{y(s^{J_k})} & \{i \in s/\mathbf{x}_i \in J_k\} \end{cases} \tag{7}$$

where $\overline{y(s^{J_i})}$ is the mean of $y$ for the members of the sampled population, $s$, which meet the criteria of the $i$th terminal node. If considering the Bagged Trees method, predictions are made by averaging results from a range of $m$ unpruned trees known as *weak classifiers*, each one trained in a bootstrapped subsample of the complete dataset:

$$E_m(y_i|\mathbf{x}_i) = \frac{\sum_{j=1}^{m} \phi_j(\mathbf{x}_i)}{m}, \ \phi_j(\mathbf{x}_i) = \begin{cases} \overline{y(s^{J_1^j})} & \{i \in s/\mathbf{x}_i \in J_1^j\} \\ ... & ... \\ \overline{y(s^{J_k^j})} & \{i \in s/\mathbf{x}_i \in J_k^j\} \end{cases} \tag{8}$$

where $\overline{y(s^{J_i^j})}$ is the mean of $y$ for the members of the sampled population, $s$, which meet the criteria of the $i$th terminal node of the $j$th tree.

### 3.3 Gradient Boosting Machine

Gradient Boosting Machine (GBM) algorithm can be used for prediction in superpopulation modeling. The new formula of the estimates of $y$ would be:

$$E_m(y_i|\mathbf{x}_i) = v^T J(\mathbf{x}_i) \tag{9}$$

where $J(\mathbf{x}_i)$ stands for a matrix of terminal nodes of $m$ decision trees used for boosting, which is obtained through an iterative process that aims to minimize a given loss function, and $v$ is a vector representing the weight of each tree.

### 3.4 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) can also be used for prediction, although they constitute a much simpler algorithm. The expectance of $y_i$ is calculated by averaging the value of $y$ for its $k$ nearest neighbors, this is, the $k$ individuals closer to the $i$th individual according to the covariates $\mathbf{x}_i$:

$$E_m(y_i|\mathbf{x}_i) = \frac{\sum_{j \in s/d(\mathbf{x}_i,\mathbf{x}_j) \leq d(\mathbf{x}_i,\mathbf{x}_{(k)})} y_j}{k} \tag{10}$$

where $x_{(1)},...,x_{(n-1)}$ denote, respectively, the closest and the furthest individual to $x_i$ according to the distance $d$.

### 3.5 Neural networks with Bayesian Regularization

Approaches based on neural networks have been considered in the literature for superpopulation modeling [4]. In that class of models, expectance of $y_i$ is calculated through an iterative process as defined in [18]:

$$E_m(y_i|\mathbf{x}_i) = g\left(\sum_{k=1}^{L} v_k f_k(\cdot) + b\right) \tag{11}$$

where $g$ and $f_k$ stand for activation functions which can have the same image, $v_k$ are the weights of the $k$th neuron of the hidden layer and $b$ is the activation threshold. The inputs are noted as $f_k(\cdot)$ given that several hidden layers can be fixed and, as a result, the inputs would go through an iterative process before reaching the last layer where the outputs are calculated. Alternatively, and as a regularization method to avoid overfitting, prior distributions can be imposed in $v_k$ weights so they can be estimated by calculating those who maximize the posterior density or via maximum likelihood. Further details are described in [18].

## 4 The simulation study

### 4.1 Data

We have selected 3 different populations to experiment with. Also, for each one, we have tried different sampling strategies.

The first population, P1, consists of the 2012 edition of the Spanish Life Conditions Survey microdata [17]. The dataset contains information on economic and life conditions variables for 28,610 adult individuals. We pretend to predict the mean self-reported health on a scale from 1 to 5. For training, we used the 56 related variables. In this population, we tested two sampling strategies. The first one, P1S1, was a simple random sampling (SRS) among the population with internet access. For the second one, P1S2, a propensity to participate was considered according to the formula $Pr(yr) = \frac{yr^2 - 1900^2}{1996^2 - 1900^2}$, where $yr$ is the year the individual was born.

The second population, P2, is BigLucy [9]. It corresponds to some financial variables of 85,396 industrial companies of a city in a particular fiscal year. We used the annual income as the target variable. For training, we took into account the level of the company (small, medium or big), the number of employees, whether it is ISO certified and the company's income tax. In this population, we tried two different sampling methods. The first one, P2S1, was SRS excluding the companies without SPAM options and the small companies. In this scenario, we tested if the algorithms could accurately predict data without any training sample (since any small company can be sampled). The second one, P2S2, only filtered by SPAM availability but it included a propensity to participate with the formula $Pr(taxes) = min(taxes^2/30, 1)$ where *taxes* is the company's income tax.

The third population, P3, is the Bank Marketing Data Set [16], related to direct marketing campaigns (phone calls) of a Portuguese banking institution. We aimed to estimate the mean contact duration. We trained the algorithms with 18 variables. For sampling, we filtered by the number of contacts performed for each client and tested two possibilities. In the first one, P3S1, we performed SRS among those contacted more than 3 times. In the second one, P3S2, we tested another SRS among those contacted more than twice.

## 4.2 Procedure

For each population and sampling strategy described, we ran an experiment with 3 different sample sizes: 1000, 2000 and 5000. For each sample size, 500 simulations were executed. In each simulation, model-based, model-assisted and model-calibrated estimates were obtained using the following predictive algorithms: 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*). Default parameters were used for every algorithm except for k-Nearest Neighbors since its results were especially sensitive to parameter optimization. The proper $k$ is chosen via bootstrap. The optimization, training of models and prediction steps were performed in R ([20]) using *caret* package ([12]).

The relative mean bias, relative standard deviation and the relative Root Mean Square Error in each scenario are measured as follows:

$$RBias\ (\%) = \left( \frac{\sum_{i=1}^{500} \hat{p}_{yi}}{500} - p_y \right) \cdot \frac{100}{p_y};\ RSD\ (\%) = \sqrt{\frac{\sum_{i=1}^{500} (\hat{p}_{yi} - \hat{\bar{p}}_y)^2}{499}} \cdot \frac{100}{p_y}$$

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

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

To compare each estimator, we consider three metrics: its mean efficiency, its median efficiency and the number of times it has been among the best. An estimator has performed as the best when its RMSE differs from the minimum RMSE by less than 1%. The efficiency is defined as follows:

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

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

Additionally, the results were analyzed using linear mixed-effects regression, to obtain estimates of the effect sizes of each algorithm on the final Root Mean Square Error (RMSE). All the analyses were performed in R using packages *lme4* ([3]), *lmerTest* ([13]) and *MuMIn* ([2]).

### 4.3 Results

RMSEs of each estimator for each population, sampling method and sample size can be observed in Table 1. Some algorithms achieve good results consistently, like Ridge regression. Others can greatly outperform the rest for some cases while getting poor results for the rest, like k-Nearest Neighbors. Bayesian-regularized Neural Networks are a special case since they produce promising estimations but can suffer due to a lack of data. Finally, there is a group of algorithms that never seem to be the right choice, like Bagged Trees. In any case, there is not much difference between model-based, model-assisted and model-calibrated estimates.

In order to confirm those impressions, the ranking can be seen in Table 2. Model-assisted Ridge regression has the best mean efficiency, median efficiency and the number of times it has been among the best. This is not a surprise since it is a technique for analyzing data that suffer from multicollinearity, which is expected to be the case for most biased samples.

Results of the linear mixed-effects regression (see Appendix) confirm these conclusions: there is no evidence in the simulations' results that the effect is different between Ridge regression, GLM, LASSO maximum-likelihood regression (both bayesian and non-bayesian), k-Nearest Neighbors or Bayesian-regularized Neural Networks. Nonetheless, there is evidence of a smaller RMSE reduction effect for Gradient Boosting Machines and Bagged Trees in comparison to the algorithms aforementioned, except for k-Nearest Neighbors.

## 5 Conclusions

This paper describes some options for estimation in non-probability samples using ML techniques in three approaches: model-based, model-assisted and model-calibrated. The paper clarifies which assumptions are required and illustrates how these proposed estimators perform empirically. The main conclusion in our simulation study is that the selection of the ML algorithm used in the process is more important than the approach used in the estimation. There is a group of ML techniques that are similar in their good performance, highlighting the Ridge regression method.

Table 1: RMSE (%) of each estimator for each population, sampling method and sample size

| | 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 | 8.4 | 8.5 | 12.9 | 12.8 | 12.8 | 70.6 | 70.5 | 70.5 | 32.9 | 32.8 | 32.8 | 13.5 | 13.4 | 13.2 | 6.7 | 6.4 | 6 |
| model assisted blasso | 3.5 | 3.1 | 2.7 | 6.9 | 6 | 5.1 | **24.8** | **24.8** | **24.7** | 12.6 | 12.7 | 12.8 | 9.7 | 5.6 | 2.3 | 4.8 | 2.8 | 2.1 |
| model based blasso | 3.6 | 3.1 | 2.7 | 6.9 | 6 | 5.3 | **24.5** | **24.6** | **24.5** | 12.5 | 12.6 | 12.7 | 9.9 | 6 | 2.8 | 4.7 | 2.9 | 2.4 |
| model calibrated blasso | 3.6 | 3.1 | 2.7 | 6.8 | 5.9 | 5.1 | **24.6** | **24.6** | **24.6** | 12.5 | 12.7 | 12.7 | 9.7 | 5.6 | 2.1 | 4.9 | 2.6 | 2 |
| model assisted bridge | 3.7 | 3.2 | 2.7 | 7.1 | 6.1 | 5.2 | **24.8** | **24.7** | **24.7** | 12.6 | 12.7 | 12.8 | 10.1 | 5.5 | 3.2 | 5.4 | 3.2 | 2.3 |
| model based bridge | 3.8 | 3.3 | 2.9 | 7.3 | 6.3 | 5.4 | **24.6** | **24.5** | **24.5** | 12.5 | 12.6 | 12.6 | 10.4 | 5.7 | 3.1 | 5.5 | 3.2 | 2.3 |
| model calibrated bridge | 3.6 | 3.2 | 2.7 | 7 | 6.1 | 5.2 | **24.5** | **24.5** | **24.6** | 12.6 | 12.7 | 12.7 | 10 | 5.5 | 3 | 5.5 | 3.3 | 2.4 |
| model assisted brnn | **2.9** | **2.5** | **2.2** | **5.7** | **4.9** | **4.5** | 25.5 | 25.2 | **24.8** | 12.8 | 12.9 | 12.9 | 7.4 | 5.5 | 2.3 | 5.2 | 4.3 | 4.3 |
| model based brnn | **2.7** | **2.4** | **2.3** | **5.3** | **4.7** | **4.3** | 25.8 | 25.3 | **24.9** | 13.3 | 13.3 | 13.3 | 7.4 | 5.3 | 2.4 | 5.1 | 4.4 | 4.4 |
| model calibrated brnn | **2.9** | **2.5** | **2.3** | **5.4** | **4.9** | **4.5** | 25.6 | 25.1 | **24.6** | 12.9 | 13 | 12.9 | 7.2 | 10.6 | 2.1 | 5.4 | 4.1 | 4.3 |
| model assisted gbm | 5.5 | 5.4 | 5.4 | 9.1 | 8.9 | 8.9 | 46.2 | 46.7 | 47.1 | 17.1 | 17.4 | 17.6 | 7 | 5.3 | 5.2 | 9 | 7.2 | 4.6 |
| model based gbm | 5.5 | 5.4 | 5.5 | 9.3 | 8.9 | 8.9 | 46.2 | 46.6 | 47.1 | 17 | 17.4 | 17.5 | 7.7 | 4.8 | 5.2 | 8.9 | 7.4 | 4.6 |
| model calibrated gbm | 5.5 | 5.4 | 5.4 | 9 | 8.9 | 8.9 | 46 | 46.5 | 46.9 | 16.9 | 17.2 | 17.3 | 8 | 5.3 | 5.1 | 9 | 7 | 4.8 |
| model assisted glm | **2.8** | **2.7** | **2.6** | **5.1** | **4.9** | **4.7** | **24.5** | **24.6** | **24.6** | 12.6 | 12.7 | 12.8 | **4.1** | **2.8** | **1.5** | 5 | 4 | 3.5 |
| model based glm | **2.9** | **2.6** | **2.5** | **5** | **4.8** | **4.7** | **24.4** | **24.6** | **24.6** | 12.6 | 12.7 | 12.7 | **4.1** | **2.9** | **1.4** | 5.1 | 4.1 | 3.5 |
| model calibrated glm | **2.9** | **2.7** | **2.5** | **5.1** | **4.8** | **4.7** | **24.4** | **24.5** | **24.5** | 12.6 | 12.7 | 12.7 | **4.2** | **2.7** | **1.4** | 5 | 4.2 | 3.5 |
| model assisted glmnet | **2.9** | **2.7** | **2.6** | **5.1** | **4.9** | **4.9** | 25.4 | 25.5 | 25.6 | 12.7 | 12.8 | 12.8 | **3.9** | **2.7** | **1.4** | 4.8 | 4.1 | 3.4 |
| model based glmnet | **2.9** | **2.8** | **2.6** | **5.2** | **5** | **4.9** | 25.5 | 25.5 | 25.5 | 12.7 | 12.8 | 12.8 | **4.2** | **2.8** | **1.4** | 5 | 4 | 3.4 |
| model calibrated glmnet | **2.9** | **2.8** | **2.6** | **5.1** | **5** | **4.9** | 25.3 | 25.4 | 25.5 | 12.7 | 12.7 | 12.7 | **4.2** | **2.5** | **1.3** | 5 | 4 | 3.5 |
| model assisted knn | 4 | 3.6 | 3.1 | 6.7 | 6.2 | 5.5 | 34.5 | 34.3 | 34.1 | **7.4** | **5.8** | **4.3** | 7.5 | 6.8 | 6.3 | **3.9** | **2.6** | **1.5** |
| model based knn | 5.1 | 4.5 | 3.5 | 7.8 | 7.1 | 6.1 | 34.1 | 34.2 | 34.1 | **7.1** | **5.7** | **4.3** | 7.6 | 6.6 | 5.8 | **4.2** | **2.7** | **1.6** |
| model calibrated knn | 3.9 | 3.4 | 2.9 | 6.5 | 6.1 | 5.4 | 34.3 | 34.2 | 34.1 | **7.2** | **5.8** | **4.2** | 7.4 | 6.6 | 6.4 | **3.9** | **2.7** | **1.5** |
| model assisted lasso | 7.1 | 7.2 | 7.3 | 10.9 | 11 | 11.1 | 65.5 | 65.6 | 65.6 | 29.8 | 29.8 | 29.8 | 7.9 | 6.9 | 6.6 | **4** | **2.6** | **1.8** |
| model based lasso | 7.1 | 7.2 | 7.3 | 10.9 | 11 | 11.1 | 65.5 | 65.6 | 65.6 | 29.7 | 29.8 | 29.7 | 7.9 | 6.9 | 6.6 | **3.9** | **2.8** | **1.9** |
| model calibrated lasso | 7.2 | 7.2 | 7.2 | 10.8 | 11.1 | 11.1 | 65.4 | 65.4 | 65.5 | 29.8 | 29.7 | 29.7 | 7.9 | 6.7 | 6.5 | **3.9** | **2.7** | **1.8** |
| model assisted ridge | **2.8** | **2.7** | **2.5** | **5** | **4.8** | **4.7** | **24.5** | **24.6** | **24.6** | 12.6 | 12.7 | 12.7 | **4.1** | **2.8** | **1.3** | 4.8 | 4 | 3.4 |
| model based ridge | **2.8** | **2.6** | **2.5** | **5** | **4.7** | **4.7** | **24.5** | **24.6** | **24.7** | 12.6 | 12.7 | 12.8 | **4.2** | **2.8** | **1.4** | 5 | 4.2 | 3.5 |
| model calibrated ridge | **2.8** | **2.7** | **2.5** | **5** | **4.9** | **4.7** | **24.4** | **24.5** | **24.5** | 12.6 | 12.7 | 12.7 | **4.3** | **2.9** | **1.3** | 5.1 | 4.3 | 3.5 |
| model assisted treebag | 3.8 | 3.7 | 4.2 | 6.7 | 6 | 6.4 | 45.6 | 45.8 | 46 | 16.6 | 16.6 | 16.6 | 13 | 5.7 | 8.9 | 13.7 | 9.3 | 2.6 |
| model based treebag | 3.9 | 3.8 | 4.2 | 6.7 | 6.4 | 6.4 | 45.6 | 45.7 | 46 | 16.5 | 16.5 | 16.5 | 11.9 | 6.7 | 8.8 | 13.3 | 8.4 | 2.8 |
| model calibrated treebag | 3.7 | 3.6 | 4.2 | 6.7 | 6.1 | 6.3 | 45.6 | 45.8 | 46 | 16.4 | 16.6 | 16.6 | 14.5 | 6.3 | 8.7 | 15.9 | 9.1 | 2.7 |

Table 2: Mean and median efficiency (%) of each estimator and times it has been among the best. ma = Model-assisted, mb = Model-based, mc = Model-calibrated

|  | Mean | Median | Best |  | Mean | Median | Best |
|---|---|---|---|---|---|---|---|
| ma ridge | 62,2 | 64,3 | 13 | mb blasso | 57,4 | 60,8 | 10 |
| mb ridge | 61,9 | 64,1 | 12 | mc bridge | 56,3 | 60,9 | 9 |
| ma glm | 61,7 | 64,3 | 12 | ma bridge | 56,2 | 61,2 | 9 |
| mb glm | 61,7 | 64,1 | 12 | mc brnn | 55,8 | 61,4 | 9 |
| mc glm | 61,7 | 64,3 | 12 | mb knn | 55,7 | 51,6 | 6 |
| mc ridge | 61,6 | 64,3 | 12 | mb bridge | 55,7 | 61,3 | 7 |
| ma glmnet | 61,6 | 62,8 | 11 | ma gbm | 32,6 | 34,7 | 0 |
| mc glmnet | 61,5 | 63 | 12 | mc gbm | 32,4 | 35,1 | 0 |
| mb glmnet | 61,3 | 63 | 9 | mb gbm | 32,4 | 35 | 0 |
| mc knn | 59,1 | 53,1 | 7 | mb treebag | 32,3 | 49,6 | 0 |
| ma knn | 58,5 | 52,7 | 7 | ma treebag | 31,5 | 49,3 | 0 |
| mc blasso | 58,5 | 61,3 | 10 | mc treebag | 29,1 | 49,5 | 1 |
| ma blasso | 58,2 | 61,2 | 11 | mc lasso | 25,1 | 14,9 | 3 |
| mb brnn | 57,9 | 61,8 | 8 | ma lasso | 24,8 | 14,5 | 3 |
| ma brnn | 57,8 | 61,4 | 9 | mb lasso | 24,7 | 14,3 | 3 |

[6] also evaluates the behavior of various ML methods for model-based estimators. We have conducted a study with a broader class of estimators and more ML methods. The results obtained in our study agree on those obtained in the study by [6] in the sense that Machine Learning methods are more powerful at removing selection bias in non-probability samples than traditional estimators. However, their performance is strongly dependent on the dataset characteristics, meaning that there could not be a unique algorithm for maximizing the estimates' accuracy. Further research should consider algorithm-specific data preprocessing steps in the analysis.

## References

## References

[1] Baffetta, F., Fattorini, L., Franceschi, S. & Corona P. (2009). 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.

[2] Barton, K.: MuMIn (2018). Multi-Model Inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

[3] Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[4] Breidt, F. J. & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

[5] Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Classification and regression trees. Wadsworth, Belmont, California.

[6] Buelens, B., Burger, J. & van den Brakel, JA. (2018) Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, 86(2), 322–343.

[7] Chen, J. K. T., Valliant, R. L., & Elliott, M. R. (2019). 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.

[8] Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

[9] Gutiérrez, H. A. (2009). Estrategias de muestreo. Diseño de encuestas y estimación de parámetros. Universidad Santo Tomás, Bogotá (Colombia).

[10] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[11] Hsiang, T. C. (1975). A Bayesian View on Ridge Regression. *The Statistician*, 24(4), 267. doi:10.2307/2987923

[12] Kuhn, M. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret.

[13] Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. doi: 10.18637/jss.v082.i13

[14] Lee, S. & Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods and Research*, 37(3), 319-343.

[15] Montanari, G. E. & Ranalli, M. G. (2007). Multiple and ridge model calibration. In: *Proceedings of Workshop on Calibration and Estimation in Surveys*, Ottawa, Canada, October 31-November 1 2007.

[16] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31.

[17] National Institute of Statistics: Life Conditions Survey. Microdata (2012). Retrieved from https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=ultiDatos&idp=1254735976608 (accessed 7 november 2019).

[18] Okut, H. (2016). Bayesian regularized neural networks for small n big p data. In Rosa, J. L. G. (Ed.). (2016). Artificial Neural Networks: Models and Applications. BoD–Books on Demand.

[19] Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

[20] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[21] Rosenbaum, P. R. & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41-55.

[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

[23] Tipton, J., Opsomer, J. & Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sensing of Environment*, 139, 130-137.

[24] Valliant, R. (2019). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, smz003, https://doi.org/10.1093/jssam/smz003

[25] Wang, J. C., Opsomer., J. D. & Wang, H. (2014). Bagging non-differentiable estimators in complex surveys. *Survey Methodology*, 40, 189-209.

# Appendix

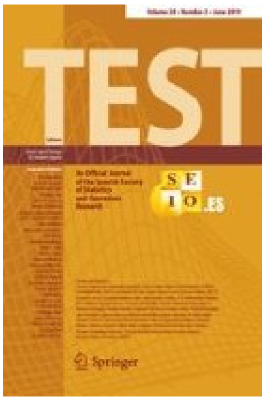
| Coefficient | Estimate | Std. Error | D. f. | t value | IC 95% | p-value |
|---|---|---|---|---|---|---|
| (Intercept) | 24.073 | 5.118 | 5.641 | 4.704 | [11.354; 36.792] | 0.0039 |
| blasso | -14.805 | 1.454 | 542.000 | -10.183 | [-17.661; -11.949] | 2.10e-16 |
| bridge | -14.637 | 1.454 | 542.000 | -10.068 | [-17.493; -11.781] | 5.69e-16 |
| brnn | -14.767 | 1.454 | 542.000 | -10.157 | [-17.623; -11.911] | 2.63e-16 |
| gbm | -8.880 | 1.454 | 542.000 | -6.108 | [-11.736; -6.024] | 1.93e-03 |
| glm | -15.444 | 1.454 | 542.000 | -10.623 | [-18.299; -12.588] | 4.51e-18 |
| glmnet | -15.268 | 1.454 | 542.000 | -10.502 | [-18.124; -12.412] | 1.31e-17 |
| knn | -14.132 | 1.454 | 542.000 | -9.720 | [-16.988; -11.276] | 1.08e-14 |
| lasso | -3.509 | 1.454 | 542.000 | -2.413 | [-6.364; -0.653] | 0.0161 |
| ridge | -15.457 | 1.454 | 542.000 | -10.632 | [-18.313; -12.602] | 4.15e-18 |
| treebag | -8.964 | 1.454 | 542.000 | -6.166 | [-11.820; -6.108] | 1.37e-03 |
| Group | Variance | Std. Dev. | | | | |
| Dataset | 147.66 | 12.151 | | | | |
| Residual | 28.53 | 5.342 | | | | |
| Dataset | Sampling | Intercept | | | | |
| P1 | P1S1 | 16.082 | | | | |
| P1 | P1S2 | 18.852 | | | | |
| P2 | P2S1 | 47.410 | | | | |
| P2 | P2S2 | 27.340 | | | | |
| P3 | P3S1 | 17.981 | | | | |
| P3 | P3S2 | 16.776 | | | | |

Table 3: Linear mixed-effects model considering algorithms as a fixed effect and datasets as random effects.

# Appendix A6

## Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest

Ferri-García, Ramón; Beaumont, Jean-François; Bosa, Keven; Charlebois, Joanne; Chu, Kenneth (2021)
Weight smoothing in adjustments for nonprobability surveys with multiple variables of interest
*TEST*



| STATISTICS & PROBABILITY | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 1.205 | 58/124 | Q2 |

**Abstract**

Adjustment techniques to mitigate selection bias in nonprobability samples often involve weighting. Procedures for estimating weights can be successful if the covariates selected for the adjustments are related to the variable of interest and the propensity to participate in the nonprobability sample. In many situations, especially in large-scale official surveys, the number of variables of interest can be large, making adjustments difficult to determine as they could be suitable for some variables but unsuitable for other ones in terms of variability. The standard compromise is to include a large number of explanatory variables in the weighting model but this may increase variability of the estimates. Weight smoothing techniques, developed for weighting in probability surveys, could be helpful in these situations. They aim to remove the variablity caused by overfit weighting models, using prediction models to substitute the fitted values from the aforementioned models for the original weights. In this study, we apply weight smoothing in the nonprobability survey context to understand how it can be helpful in multipurpose surveys under several adjustment methods for the selection bias.

# 1 Introduction

Probability sampling has been the gold standard for empirical research since its development in the XX$^{\text{th}}$ century based on the work of Neyman (1934) and Horvitz and Thompson (1952) among others. For a sample to be considered probabilistic and therefore valid for population inferences, it must be drawn under the assumption that all the individuals in the target population have a known and non-null inclusion probability. If any of these conditions do not apply, we have a nonprobability sample instead. The use of such samples in empirical sciences is widespread nowadays thanks to technological development and social media, which allows pollsters and vendors to use new questionnaire administration methods such as online and smartphone surveys. These surveys are usually administered via opt-in panels or by recruiting volunteers via snowball sampling (see Schonlau and Couper (2017) for an extensive review of methods).

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

It is possible to apply several methods to reduce selection bias when a probability sample from the same target population is available. Here we mention Propensity Score Adjustment (PSA), the Tree-based Inverse-Propensity-Weighted estimator (TrIPW) and Statistical Matching (also referred to as Sample Matching), as well as doubly robust estimators that combine Statistical Matching and PSA. Both PSA and TrIPW aim to estimate propensities of participation in a nonprobability sample by comparing to a probability sample drawn from the same population. However, while PSA estimates the propensities via predictive modeling, TrIPW aims to apply the Classification And Regression Trees (CART) methodology (Breiman et al. 1984), based on learning decision rules (that optimize an homogeneity measure) from data to build trees that allow to give a prediction of the response indicator given a set of covariates. TrIPW can be viewed as an extension of Chen et al. (2019) to CART. On the other hand, Statistical Matching focuses on another model-based approach, where the objective is to predict the values of the variable of interest for the probability sample, where the variable has not been measured. The predictive model is fitted using data from the nonprobability sample.

When the propensity model is properly specified, PSA is able to reduce bias in the estimation from nonprobability samples at the cost of increasing the variability of the estimates (Lee 2006; Lee and Valliant 2009; Valliant and Dever 2011; Ferri-García and Rueda2018). TrIPW shows itself as a more robust adjustment under complex relationships between variables, such as nonlinearities (Chu and Beaumont 2019), although PSA using Machine Learning algorithms to model propensities could also be useful in those situations (Ferri-García and Rueda 2020). Statistical Matching has also been proven to mitigate selection bias in nonprobability samples (Castro-Martín et al. 2020). The combination of both strategies via doubly robust estimators is able to outperform both approaches on their own (Chen et al. 2019).

Despite the statistical advantage of Matching techniques, they can be disadvantageous in multipurpose surveys where the variables of interest are multiple. In those situations, which are common in official statistics surveys, each variable of interest would need a specific model to predict its values in the probability sample. This could lead to impractical situations and a higher probability of model misspecifications. The use of reweighting strategies such as PSA and TrIPW would be a reasonable solution, as the same vector of weights could be used to estimate all of the variables of interest. However, research has shown that propensity techniques are more efficient when the covariates used for the estimation are related to the outcome variables, that is, the variables of interest (Hirano and Imbens 2001; Brookhart et al. 2006). In a multipurpose survey, the adequacy of the covariates may vary between variables, leading to model misspecifications. The standard compromise is to include a large number of covariates in the weighting model. This may increase the variability of the resulting estimates due to overfitting.

In multipurpose probability surveys, weight smoothing (Beaumont 2008) has been shown to be effective at minimizing the problem posed by the existence of multiple target variables. Weight smoothing aims to reduce the variability of the estimates, which can be large if the design variables are unrelated to the target variables, using predictive modeling. To the best of our knowledge, this technique has not been evaluated in a nonprobability survey context, where the inclusion probabilities have to be estimated. The objective of this study is to examine the adequacy of weight smoothing for multipurpose nonprobability surveys, and explore the situations that could enhance the efficiency of this technique.

## 2 Weighting in nonprobability surveys

Let $U$ be a target population of size $N$ from which we want to estimate a population parameter for a given variable of interest, $Y$. To this end, we obtain a nonprobability sample, $s_v$, from the population, $U$. The participation mechanism may depend on features such as self-selection or device availability (computer, internet access, etc.). In this case, the probability of being included in $s_v$ for each individual in $U$, $\pi$, cannot be known a priori. Let $R$ be the indicator variable which measures whether an individual from $U$ has participated. If we assume that a vector of covariates, $\mathbf{X}$, is available and related to $\pi$ such that:

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

we can estimate the inclusion probability if a probability sample, $s_r$, drawn from the full population $U$ is available or if $\mathbf{X}_i$ is known for every $i \in U$. For all the individuals in $s_r$, $\mathbf{X}$ must have been measured but $Y$ is not measured. In that case, we can obtain an estimate of $\pi$ by modelling the response indicator.

PSA was originally developed to mitigate selection bias in nonrandomized clinical trials (Rosenbaum and Rubin 1983), but it was adapted to the survey nonresponse field shortly after (Little 1986). PSA adapted to the nonprobability survey context as a method to mitigate selection bias was developed by Lee (2006) and Lee and Valliant (2009). Research on PSA has focused on using logistic regression to estimate propensities as

$$\hat{\pi}_i^{LR} = \frac{exp(\hat{\beta}\mathbf{x}_i)}{1+exp(\hat{\beta}\mathbf{x}_i)}, i \in U \tag{2}$$

where $\hat{\beta}$ is an estimator of the unknown vector of model parameters $\beta$ obtained by pooling $s_r$ and $s_v$. Recent literature has also considered some nonparametric methods, such as Machine Learning classification algorithms, to estimate propensities (Ferri-García and Rueda 2020). This PSA approach is valid provided the participation rate is small (see Beaumont 2020). Chen et al. (2019) developed a method that does not require this assumption.

The TrIPW estimator, developed in Chu and Beaumont (2019), uses one of the mentioned Machine Learning classification algorithms: Classification And Regression Trees (CART) (Breiman et al. 1984), and does not require the participation rate to be small. Although PSA and TrIPW use estimated participation propensities, the methodology of the latter is slighty different and takes into account design weights of the probability sample. The propensity for each individual $i \in s_v$ is estimated as:

$$\tilde{\pi}_i^{CART} = \frac{\#(l(i) \cap s_v)}{\#(l(i))} \tag{3}$$

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 participants among the population individuals that would be classified in the terminal node $l$. However, as data from $U - s_v$ is not available, the propensity has to be estimated using a modified CART algorithm and estimating proportions by taking design weights into account as follows:

$$\hat{\pi}_i^{CART} = \frac{\#(l(i) \cap s_v)}{\hat{\#}(l(i))} = \frac{\#(l(i) \cap s_v)}{\sum_{j \in l(i) \cap s_r} \frac{1}{p_j}} \tag{4}$$

where $p_j$ is the inclusion probability for individual $j$ in $s_r$. The equation above is equivalent to dividing the number of individuals from $s_v$ that belong to $l(i)$ by the sum of the sampling weights of those individuals from $s_r$ that belong to $l(i)$. This non-parametric approach shows acceptable results under non-linearity conditions (Chu and Beaumont 2019).

Propensities are often transformed into weights with the inverse probability formula, as noted in Valliant (2019):

$$w_i^{IPW1} = \frac{1}{\hat{\pi}_i}, i \in s_v \tag{5}$$

A variant of this formula was mentioned in Schonlau and Couper (2017), which assumes a lower bound of 0 for the vector of weights:

$$w_i^{IPW2} = \frac{1 - \hat{\pi}_i}{\hat{\pi}_i}, i \in s_v \tag{6}$$

The original literature of PSA for nonprobability sampling (Lee 2006; Lee and Valliant 2009) considered the stratification of propensities into $g$ partitions, usually $g = 5$ following the criteria of Cochran (1968), and the calculation of weights using a correction factor that takes into account the original design weights of those individuals belonging to a given propensity stratum:

$$w_i^{STR1} = d_i \cdot f_c = d_i \frac{\sum_{j \in s_r(g_i)} d_j / \sum_{j \in s_r} d_j}{\sum_{j \in s_v(g_i)} d_j^v / \sum_{j \in s_v} d_j^v}, i \in s_v \tag{7}$$

where $d_i^v$ is some weight for individual $i$ of $s_v$, $d_j$ is the design weight of individual $j$ of $s_r$, and $s_v(g_i)$ and $s_r(g_i)$ represent the set of individuals from the probability and the nonprobability sample respectively that belong to the $g$-th propensity stratum, which contains individual $i$. Valliant and Dever (2011) used a similar approach that also involved stratification of propensities, but the strata are used to construct inverse probability weights instead, using the mean propensity of each strata:

$$w_i^{STR2} = \frac{\#(s_v(g_i))}{\sum_{j \in s_v(g_i)} \pi_j}, i \in s_v \tag{8}$$

An alternative approach is to compute weights using 4 and 5, with the terminal nodes being taken as the propensity strata. This usually brings a certain robustness to model failure.

# 3 Weight smoothing

Previous studies show that the application of adjustment techniques in nonprobability samples contributes to reducing the bias by different degrees depending on the selection mechanism, but at the cost of increasing the variance of the estimates (Lee 2006; Lee and Valliant 2009; Ferri-García and Rueda 2018). This variance is directly tied to the weights' variance; therefore it seems reasonable to focus on strategies that reduce the variability of the weights, especially in case of model misspecification where no gains would be expected in terms of bias reduction, which is often the case for multipurpose surveys where a single vector of weights is used for the estimation of all population parameters. One of these strategies could be weight smoothing developed by Beaumont (2008) for the probability sampling context.

This method assumes that, for a given probability sample $s$, the vector of design weights $d$ is linked to the set of variables of interest, $\mathbf{Y}$, such that:

$$d_i = f(\mathbf{Y}_i;\gamma) + \varepsilon_i, i \in s \tag{9}$$

where $\gamma$ is a vector of unknown coefficients and $\varepsilon$ is a random variable with $E[\varepsilon] = 0$ and $Var(\varepsilon) < \infty$ which represents the noise. Weight smoothing substitutes the original weights, $w$, with smoothed weights, $\tilde{w}$, which are equivalent to their expected value given the set of variables of interest $\mathbf{Y}$ and a model $M$:

$$\tilde{w}_i = E_M\left[w_i|I,\mathbf{Y}\right], i \in s, \tag{10}$$

where $I$ is a binary variable indicating whether an individual from $U$ belongs to $s$ or not. This approach has desirable properties, such as unbiasedness, as long as the model $M$ is properly specified. The smoothed estimator is also expected to have a smaller variance than the non-smoothed estimator. As described in Haziza and Beaumont (2017), this method aims to eliminate the random noise part of the weights, hence reducing variance while maintaining unbiasedness.

It is reasonable to assume that nonprobability sample weights are associated with the target variables, especially if the covariates of the propensity model are related to $\mathbf{Y}$. We propose the following weight smoothing approach for obtaining $\tilde{w}$ in the nonprobability sample:

$$\tilde{w}_i = E_M\left[w_i^k|R,\mathbf{Y}\right], i \in s_v \tag{11}$$

where $k = \{$ IPW1, IPW2, STR1, STR2 $\}$ represents the weighting approach used to convert propensities to weights. Note that the indicator variable for the nonprobability sample, $R$, is substituted for $I$ from Eq. 10 meaning that $s_v$ is used to fit the model $M$. In order to evaluate weight smoothing, we applied this methodology in two simulations, one with artificial data and the other with real data, considering different scenarios using the weighting approach $k =$IPW1.

# 4 Data

## 4.1 Artificial data

We created a population of size $N = 500000$ with 10 covariates $(x_1, ..., x_{10})$, 10 variables of interest $(y_1, ..., y_{10})$ and a variable indicating the propensity of each individual to take part in a volunteer sample, $\pi_i$. The covariates followed Bernoulli and Normal distributions:

$$x_1, x_4, x_7 \sim Be(0.5) \tag{12}$$

$$x_3, x_6, x_9 \sim Be(0.2) \tag{13}$$

$$x_2, x_5, x_8, x_{10} \sim N(0,1) \tag{14}$$

The propensities depended on the values of $x_7, x_8$ and $x_9$ according to the following logistic formula:

$$\pi_i = \frac{exp(-0.5 + 2.5x_7 + \sqrt{2\pi}x_8 - (11/3)x_9)}{1 + exp(-0.5 + 2.5x_7 + \sqrt{2\pi}x_8 - (11/3)x_9)}, \quad i = 1, 2, ..., 500000 \tag{15}$$

This formula was intended to create weights with high variability, a situation where the advantages of weight smoothing might be more visible. The histogram of propensities can be observed in Figure 1; the mean value of $\pi$ is 0.5089, with a standard deviation of 0.3741 which means a coefficient of variation of 0.7351. First and 3rd quartile are 0.1111 and 0.9002 respectively.

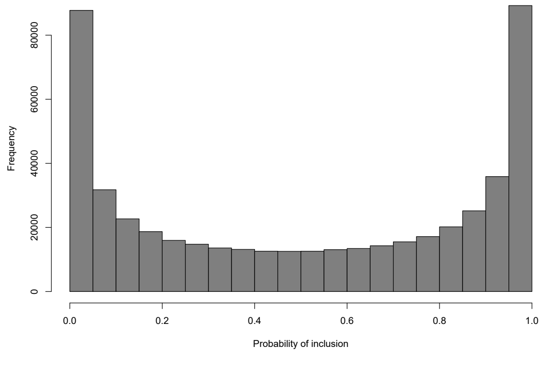


Figure 1: Histogram of the population propensities

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

Sc. 1. No relationship between any variable in $(y_1, ..., y_{10})$ and $\pi$

Sc. 2. Relationship between every variable in $(y_1,...,y_{10})$ and $\pi$

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

$$y_1 \sim B\left(\frac{exp(-1+3x_1+x_2+x_3+\mathbf{1}_{\text{Scenario 2}}5\pi)}{1+exp(-1+3x_1+x_2+x_3+\mathbf{1}_{\text{Sc. 2}}5\pi)}\right) \tag{16}$$

$$y_2 \sim N(0.1)-1+3x_1+x_2+x_3+\mathbf{1}_{\text{Sc. 2}}5\pi \tag{17}$$

$$y_3 \sim B\left(\frac{exp(-1+x_1+x_2+3x_3-\mathbf{1}_{\text{Sc. 2}}5\pi)}{1+exp(-1+x_1+x_2+3x_3-\mathbf{1}_{\text{Sc. 2}}5\pi)}\right) \tag{18}$$

$$y_4 \sim B\left(\frac{exp(\mathbf{1}_{\text{Sc. 2}}\pi)}{1+exp(\mathbf{1}_{\text{Sc. 2}}\pi)}\right) \tag{19}$$

$$y_5 \sim N(0.1)+2+\mathbf{1}_{\text{Sc. 2}}2\pi \tag{20}$$

$$y_6 \sim B\left(\frac{exp(-1.5+\mathbf{1}_{\text{Sc. 2}}\pi)}{1+exp(-1.5+\mathbf{1}_{\text{Sc. 2}}\pi)}\right) \tag{21}$$

$$y_7 \sim B\left(\frac{exp(-\mathbf{1}_{\text{Sc. 2}}\pi)}{1+exp(-\mathbf{1}_{\text{Sc. 2}}\pi)}\right) \tag{22}$$

$$y_8 \sim N(0.1)-2-\mathbf{1}_{\text{Sc. 2}}2\pi \tag{23}$$

$$y_9 \sim B\left(\frac{exp(-1.5-\mathbf{1}_{\text{Sc. 2}}\pi)}{1+exp(-1.5-\mathbf{1}_{\text{Sc. 2}}\pi)}\right) \tag{24}$$

$$y_{10} \sim N(0.1)+\mathbf{1}_{\text{Sc. 2}}2\pi \tag{25}$$

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

Table 1: Population Pearson's correlations between $\pi$ and $(y_1,...,y_{10})$ in Scenarios 1 and 2.

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

Table 2: Results of t-tests for mean difference in $\pi$ between binary classes of $y_1$, $y_3$, $y_4$, $y_6$, $y_7$ and $y_9$ in the simulated population in Scenario 2.

| k | Mean of $\pi$ $(y_k = 0)$ | Mean of $\pi$ $(y_k = 1)$ | t | p-value |
|---|---|---|---|---|
| 1 | 0.1761 | 0.5728 | -401.35 | ¡2.2e-16 |
| 3 | 0.5922 | 0.2043 | 397.20 | ¡2.2e-16 |
| 4 | 0.4234 | 0.5612 | -128.61 | ¡2.2e-16 |
| 6 | 0.4717 | 0.6061 | -116.84 | ¡2.2e-16 |
| 7 | 0.5613 | 0.4234 | 128.70 | ¡2.2e-16 |
| 9 | 0.5259 | 0.3891 | 88.17 | ¡2.2e-16 |

### 4.2 Real data

The dataset used to experiment in a real life situation came from the 2012 edition of the Spanish Life Conditions Survey (National Institute of Statistics 2012). This is an annual survey measuring several aspects of life conditions, such as health status, degree of deprivation and employment conditions, in the Spanish adult population. The survey includes specific modules in each edition; in 2012, the module consisted of a battery of questions regarding household conditions. The survey sampling follows a stratified cluster scheme, where the primary units are the households and the secondary units are their members. The total sample size in 2012 was $n = 33579$.

For its use as a pseudopopulation, the sample dataset was filtered to rule out those individuals and variables with high amounts of missing data. This reduced the dataset to $n = 28210$ and 146 variables, from which 61 were selected for the simulations. The sample was subsequently bootstrapped in order to increase its size to $n = 1000000$, and finally all the individuals who selected any of the responses related to refusal options (”Does not know” or ”Does not answer”) were also ruled out of the analysis to avoid further problems with rare classes in the simulations. The final pseudopopulation size for the experiments was $N = 990838$.

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

- SRSWOR from the population who have a computer at home.
- Unequal probability sampling without replacement from the population who have a computer at home, where the probabilities are calculated through the formula:

$$\pi_i = \frac{(\text{Year of birth} - 1925)^4}{(1996 - 1925)^4} \tag{26}$$

Regarding covariates, two complementary sets were defined:

- A set of nine demographic variables, more precisely: region, urbanization level, number of members of the household and consumption units (weighted mean of the number of members of the household following OECD criteria, where adults have more weight than teenagers and teenagers have more weight than children), sex, marital status, country of birth, nationality, and whether the individual is currently a student or not.

- A set of eight variables related to economic and material deprivation, more precisely: capacity of the household to make ends meet, minimum income required by the household to make ends meet, whether the household has the capacity to go on holiday, have a meat or fish meal at least every two days, and deal with unforeseen expenses, household under the poverty threshold, person under the poverty threshold, and household in a situation of severe material deprivation.

Ten variables were defined as variables of interest:

- $y_1$ = Household expenses are a heavy burden (ordinal scale, 1-3)

- $y_2$ = Household has a car (dichotomous)

- $y_3$ = Self-reported health (5-point Likert scale)

- $y_4$ = Disability in the previous 6 months (ordinal scale, 1-3)

- $y_5$ = Number of months working part-time (integer, 0-12)

- $y_6$ = Household expenses in EUR (continuous)

- $y_7$ = Household with noise problems (dichotomous)

- $y_8$ = Household with heating system (dichotomous)

- $y_9$ = Simulated random Be(0.5) variable.

- $y_{10}$ = Simulated random N(0, 1) variable.

## 5 Experimental design and metrics

The settings of the experiment were kept as equal as possible for all simulation scenarios. Each simulation was run 500 times, drawing probability and nonprobability samples of equal sizes ($n_r = n_v = 1000$) using the sampling designs described in the previous section. Three approaches were applied to estimate nonprobability sample propensities: logistic regression, k-Nearest Neighbors, and CART with fixed parameters of minimum cell size (50) and minimum impurity of terminal nodes (0.0001). Logistic regression modeled the propensities with the usual logit link given the set of covariates $\mathbf{x}$:

$$\pi_i^* = \frac{exp(\beta^T \mathbf{x_i})}{1+exp(\beta^T \mathbf{x_i})}, \quad i \in s_r \cup s_v \tag{27}$$

where $\beta$ is the vector of coefficients associated to each predictor, estimated through Iterative Reweighted Least Squares (IWLS) with $R^*$ as the target variable, which is defined as $R_i^* = 1$ if $i \in s_v$ and $R_i^* = 0$ if $i \in s_r$. The proximity of $R^*$ to $R$ depends on the overlap between $s_r$ and $s_v$, which tends to be smaller if the sampling fraction for $s_v$ is small. k-Nearest Neighbors used the individuals in $s_r \cup s_v$ to estimate the propensity of the $i$-th individual, $i \in s_r \cup s_v$, to be selected in $s_v$ using the proportion of neighbors that belong to $s_v$, that is, those for which $R^* = 1$. The formula for propensities can therefore be expressed as follows:

$$\pi_i^* = \frac{\sum_{j \in s_r \cup s_v / d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k)})} R_j^*}{k}, \quad i \in s_r \cup s_v \tag{28}$$

where $d$ is the function (in this case Euclidean) that measures the distance between two individuals given a set of covariates, and $\mathbf{x}_{(k)}$ represents the covariates of the $k$-th closest individual. This means that $d(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between the $i$-th individual for which propensity is calculated and an individual $j \in s_r \cup s_v$, and $d(\mathbf{x}_i, \mathbf{x}_{(k)})$ represents the distance between the $i$-th individual and the $k$-th closest individual according to function $d$. Any individual that provides a value of $d$ below $d(\mathbf{x}_i, \mathbf{x}_{(k)})$ is considered a neighbor and subsequently used to compute the propensities. This is the same k-NN algorithm used in Ferri-García and Rueda (2020), where a more detailed explanation can be found. The optimal number of neighbors was chosen by 10-fold cross-validation in each simulation run using the R package *caret* (Kuhn 2018). CART was applied as defined in Section 2, where all of the sampling probabilities for $s_r$ were $p_j = 1000/N, j \in s_r$ given that the reference sample was always drawn by SRSWOR. This means that the formula for propensity estimation given in Equation 4 could be simplified to:

$$\hat{\pi}_i^{CART} = \frac{\#(l(i) \cap s_v)}{\hat{\#}(l(i))} = \frac{\#(l(i) \cap s_v)}{\frac{\#(l(i) \cap s_r) \cdot N}{1000}} = \frac{\#(l(i) \cap s_v) \cdot 1000}{\#(l(i) \cap s_r) \cdot N} \tag{29}$$

In all cases, inverse probability weighting was used to build the weights from the propensities using the $w_i = 1/\hat{\pi}_i^*$ formula from Valliant (2019). The variance and coefficient of variation of the vector of weights was calculated in each run and weighting approach to evaluate how the inherent variability of weights could further explain the performance of weight smoothing strategies.

Two different predictive models were used in the weight smoothing step, which used the different sets of variables of interest described in the previous section as input variables (with the weights as the output variable):

- XGBoost algorithm (XGB) using the *xgboost* package in R (Chen 2020).

- Least Absolute Shrinkage and Selection Operator (LASSO) regression using the *glmnet* package in R (Friedman 2010).

XGBoost algorithm was trained with default hyperparameters: a L1 regularization term of $\alpha = 0.1$, a L2 regularization term of $\lambda = 0.0001$ and a learning rate of $\eta = 0.3$. The number of rounds was fixed at 50. In the case of LASSO, the optimal shrinkage parameter, $\lambda$, was obtained with a 10-fold cross-validation procedure in each run of the simulation.

Relative measures of bias, standard deviation and MSE were also calculated to allow the comparison between simulation scenarios, as well as the comparison with the baseline cases where no adjustments are applied. More precisely, bias was normalized by the true population parameter, $\overline{Y}$, while standard deviation and MSE were compared to the unadjusted case where no reweighting is applied.

$$\text{RelBias}_k = \left|\frac{\text{Bias}_k}{\overline{Y}}\right| = \left|\frac{\hat{\overline{Y}} - \overline{Y}}{\overline{Y}}\right| = \left|\frac{\frac{\sum_{i=1}^{500}\hat{\bar{y}}_i}{500} - \overline{Y}}{\overline{Y}}\right| \tag{30}$$

$$\text{StdDev}_k = \sqrt{\frac{\left(\sum_{i=1}^{500}\hat{\bar{y}}_i - \frac{\sum_{i=1}^{500}\hat{\bar{y}}_i}{500}\right)^2}{499}} \tag{31}$$

$$\text{RelMSE}_k = \frac{\text{MSE}_k}{\text{MSE}_{\text{Unw}}} = \frac{\text{Bias}_k^2 + \text{StdDev}_k^2}{\text{Bias}_{\text{Unw}}^2 + \text{StdDev}_{\text{Unw}}^2} \tag{32}$$

where $k \in \{\text{Unw, NS, XGB, LASSO}\}$ represents the algorithm used for weight smoothing, the case where no smoothing is applied (NS), or the case where no adjustment at all is applied (Unw) and $\hat{\bar{y}}_i$ is the estimated mean from the $i$-th simulation run.

The calculation of $\hat{\bar{y}}_i$ for the unweighted case was the arithmetic mean of $y$ in the nonprobability sample:

$$\hat{\bar{y}}_i^{\text{Unw}} = \frac{\sum_{j \in s_v^i} y_j}{1000}, i = 1, ..., 500 \tag{33}$$

where $s_v^i$ represents the nonprobability sample drawn in the $i$-th simulation run. In the case where PSA or TrIPW were applied but no smoothing was done, $\hat{\bar{y}}_i$ was calculated as follows:

$$\hat{\bar{y}}_i^{\text{NS}} = \frac{\sum_{j \in s_v^i} w_j y_j}{\sum_{j \in s_v^i} w_j}, i = 1, ..., 500 \tag{34}$$

where $w = (w_1, ..., w_{1000})$ is the vector of inverse propensity weights for the individuals of $s_v^i$ described before. Finally, when any weight smoothing method was applied on PSA or TrIPW weights, the formula used to calculate $\hat{\bar{y}}_i$ was:

$$\hat{\bar{y}}_i^{\text{k}} = \frac{\sum_{j \in s_v^i} \tilde{w}_j y_j}{\sum_{j \in s_v^i} \tilde{w}_j}, i = 1, ..., 500, k = \{\text{XGB, LASSO}\} \tag{35}$$

where $\tilde{w}$ represents the vector of smoothed weights.

# 6 Results

## 6.1 Artificial data simulation

Relative bias of adjustments in each scenario can be consulted in Table 3. As expected, all of the adjustments in Scenario 1 (where there is no relationship between any variable in $(y_1,...,y_{10})$ and $\pi$), as well as the case where no adjustment is done, present a very low amount of bias because of the independence between the variables and the inclusion probabilities. The exception observed for $y_{10}$ can be explained by the nature of the variable itself: as it is centered on 0, the relative bias tends to overstate otherwise negligible differences.

This is not the case of Scenario 2 (where each $y$ variable is related to $\pi$), where the bias induced by the sampling mechanism gets reduced with PSA and TrIPW. When comparing the three propensity estimation methods, TrIPW provides the smallest bias among them, achieving reductions of more than a half of the original bias for almost every variable, although PSA is also able to reduce bias to a lesser extent.

In both scenarios, the application of weight smoothing did not produce important changes in bias.

Table 3: Relative bias (*RelBias*) for each variable, adjustment method and artificial data scenario.

| | | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sc. | Obj. | Unw | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| 1 | $y_1$ | 0.0001 | 0.0008 | 0.0007 | 0.0016 | 0.0017 | 0.0006 | 0.0012 | 0.0001 | 0.0013 | 0.0002 |
| | $y_2$ | 0.0016 | 0.0026 | 0.0005 | 0.0035 | 0.0007 | 0.0007 | 0.0002 | 0.0054 | 0.0008 | 0.0001 |
| | $y_3$ | 0.0007 | 0.0003 | 0.0009 | 0.0001 | 0.0017 | 0.0009 | 0.0009 | 0.0007 | 0.0011 | 0.0020 |
| | $y_4$ | 0.0001 | 0.0006 | 0.0021 | 0.0004 | 0.0012 | 0.0002 | 0.0017 | 0.0035 | 0.0046 | 0.0015 |
| | $y_5$ | 0.0007 | 0.0010 | 0.0014 | 0.0011 | 0.0011 | 0.0016 | 0.0008 | 0.0010 | 0.0005 | 0.0014 |
| | $y_6$ | 0.0008 | 0.0026 | 0.0021 | 0.0011 | 0.0013 | 0.0038 | 0.0019 | 0.0113 | 0.0019 | 0.0047 |
| | $y_7$ | 0.0002 | 0.0004 | 0.0021 | 0.0004 | 0.0007 | 0.0026 | 0.0007 | 0.0012 | 0.0011 | 0.0017 |
| | $y_8$ | 0.0007 | 0.0009 | 0.0015 | 0.0012 | 0.0007 | 0.0024 | 0.0009 | 0.0004 | 0.0012 | 0.0010 |
| | $y_9$ | 0.0006 | 0.0066 | 0.0028 | 0.0063 | 0.0058 | 0.0020 | 0.0065 | 0.0034 | 0.0045 | 0.0036 |
| | $y_{10}$ | 0.6252 | 1.7615 | 1.4013 | 1.7762 | 0.2383 | 1.4420 | 0.6990 | 0.5801 | 0.5235 | 1.1523 |
| 2 | $y_1$ | 0.1256 | 0.0989 | 0.0993 | 0.0993 | 0.1104 | 0.1099 | 0.1117 | 0.0700 | 0.0709 | 0.0720 |
| | $y_2$ | 0.4233 | 0.3088 | 0.3073 | 0.3098 | 0.3595 | 0.3605 | 0.3643 | 0.1778 | 0.1820 | 0.1848 |
| | $y_3$ | 0.5961 | 0.4692 | 0.4642 | 0.4709 | 0.5264 | 0.5264 | 0.5347 | 0.3179 | 0.3296 | 0.3290 |
| | $y_4$ | 0.1030 | 0.0771 | 0.0747 | 0.0780 | 0.0880 | 0.0883 | 0.0899 | 0.0464 | 0.0467 | 0.0507 |
| | $y_5$ | 0.1826 | 0.1331 | 0.1320 | 0.1336 | 0.1554 | 0.1561 | 0.1578 | 0.0771 | 0.0772 | 0.0806 |
| | $y_6$ | 0.1947 | 0.1385 | 0.1352 | 0.1405 | 0.1641 | 0.1612 | 0.1683 | 0.0749 | 0.0796 | 0.0853 |
| | $y_7$ | 0.1682 | 0.1258 | 0.1271 | 0.1272 | 0.1422 | 0.1437 | 0.1460 | 0.0730 | 0.0749 | 0.0809 |
| | $y_8$ | 0.1826 | 0.1331 | 0.1331 | 0.1336 | 0.1560 | 0.1553 | 0.1584 | 0.0752 | 0.0786 | 0.0786 |
| | $y_9$ | 0.2374 | 0.1747 | 0.1799 | 0.1768 | 0.1997 | 0.2104 | 0.2047 | 0.1092 | 0.1203 | 0.1189 |
| | $y_{10}$ | 0.5363 | 0.3923 | 0.3884 | 0.3937 | 0.4571 | 0.4546 | 0.4642 | 0.2234 | 0.2268 | 0.2335 |

Relative MSE or efficiency of adjustments in each scenario in comparison to the unweighted case can be seen in Table 4. Values below 1 indicate that the adjustment performed better than the non-adjusted case. In Scenario 1, weight smoothing is able to slightly increase the efficiency of PSA. The application of weight smoothing is highly beneficial when adjusting with TrIPW, given that this method induces a large variability that decreases efficiency. When smoothing TrIPW weights with LASSO, the MSE of the estimates were comparable to the MSE of the unadjusted case. In the non-smoothed case, the MSE of the estimates was around twice the

MSE of the unadjusted one. In Scenario 2, the application of weight smoothing strategies did not provide any noticeable improvement in terms of MSE.

Table 4: Relative MSE (*RelMSE*) for each variable, adjustment method and artificial data scenario.

| | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sc. | Obj. | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| 1 | $y_1$ | 1.0650 | 0.9965 | 1.0373 | 1.1940 | 0.9698 | 1.1072 | 2.1226 | 2.0063 | 1.2088 |
| | $y_2$ | 0.7589 | 0.7271 | 0.8387 | 0.8758 | 0.8472 | 0.8643 | 2.0490 | 2.1111 | 1.1808 |
| | $y_3$ | 0.9735 | 0.8442 | 0.9613 | 0.9254 | 0.8901 | 0.8556 | 1.9767 | 1.8238 | 1.1513 |
| | $y_4$ | 1.2027 | 1.1451 | 1.0477 | 1.3222 | 1.3840 | 1.1168 | 2.2965 | 1.8026 | 1.3997 |
| | $y_5$ | 1.0973 | 1.0343 | 1.0034 | 1.0782 | 1.1426 | 0.9731 | 1.5602 | 1.8057 | 0.9491 |
| | $y_6$ | 0.9953 | 1.1502 | 0.9141 | 1.1862 | 1.3163 | 0.9953 | 1.8365 | 1.9509 | 1.0120 |
| | $y_7$ | 1.2386 | 1.2535 | 1.1363 | 1.2386 | 1.2386 | 1.0660 | 2.2152 | 1.8522 | 1.2837 |
| | $y_8$ | 1.1755 | 0.9830 | 1.0738 | 1.2964 | 1.3110 | 1.0938 | 1.9624 | 2.1363 | 1.1343 |
| | $y_9$ | 1.0034 | 0.9876 | 0.9254 | 1.1180 | 1.0518 | 0.9562 | 1.9357 | 1.8052 | 1.1690 |
| | $y_{10}$ | 0.9037 | 0.9981 | 0.8359 | 1.0533 | 1.0409 | 0.9327 | 1.8232 | 1.8560 | 1.1036 |
| 2 | $y_1$ | 0.6279 | 0.6339 | 0.6324 | 0.7787 | 0.7721 | 0.7989 | 0.3398 | 0.3454 | 0.3577 |
| | $y_2$ | 0.5335 | 0.5282 | 0.5369 | 0.7223 | 0.7263 | 0.7418 | 0.1871 | 0.1949 | 0.2013 |
| | $y_3$ | 0.6198 | 0.6064 | 0.6247 | 0.7765 | 0.7765 | 0.8000 | 0.3082 | 0.3239 | 0.3283 |
| | $y_4$ | 0.5897 | 0.5574 | 0.6039 | 0.7520 | 0.7600 | 0.7787 | 0.3021 | 0.3038 | 0.3348 |
| | $y_5$ | 0.5337 | 0.5247 | 0.5377 | 0.7257 | 0.7322 | 0.7493 | 0.1903 | 0.1895 | 0.2064 |
| | $y_6$ | 0.5632 | 0.5496 | 0.5769 | 0.7772 | 0.7360 | 0.8097 | 0.2676 | 0.2887 | 0.3066 |
| | $y_7$ | 0.5944 | 0.6015 | 0.6063 | 0.7388 | 0.7573 | 0.7733 | 0.2823 | 0.2888 | 0.3208 |
| | $y_8$ | 0.5353 | 0.5356 | 0.5393 | 0.7333 | 0.7268 | 0.7558 | 0.1816 | 0.1967 | 0.1974 |
| | $y_9$ | 0.6203 | 0.6463 | 0.6306 | 0.7728 | 0.8553 | 0.8018 | 0.4272 | 0.4620 | 0.4488 |
| | $y_{10}$ | 0.5377 | 0.5268 | 0.5417 | 0.7282 | 0.7208 | 0.7508 | 0.1852 | 0.1910 | 0.2010 |

### 6.2 Real data simulation

Relative bias of the estimates in the real data simulation with the different covariates for the adjustments and SRSWOR can be observed in Table 5. It is noticeable that the performance of the adjustments depends on the set of covariates used; when using the demographics set, the bias is mostly reduced in variables $y_3$ (self-reported health, 5-point Likert), $y_5$ (number of months working part time, integer 0-12) and $y_7$ (household with noise problems, dichotomous) after the application of the methods for propensity estimation. The comparison between algorithms shows that TrIPW has the largest reduction in bias, but the differences are scarce. When using deprivation covariates, which are assumed to be more related to the target variables, the bias is reduced in most of the cases, with TrIPW showing an advantage again. As in the previous example, weight smoothing did not produce substantial changes in bias, except for $y_{10}$ (independent Gaussian random variable) where the application of weight smoothing with LASSO regression modifies the distribution of the estimator and therefore the relative bias differs from the non-smoothed and smoothed with XGBoost cases.

Table 5: Relative bias (*RelBias*) for each variable, adjustment method and covariates, in the real data experiment when drawing $s_v$ with SRSWOR from the subpopulation having a computer at home.

| | | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cov. | Obj. | Unw | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| Dem. | $y_1$ | 0.011 | 0.014 | 0.014 | 0.013 | 0.016 | 0.016 | 0.013 | 0.021 | 0.021 | 0.018 |
| | $y_2$ | 0.119 | 0.113 | 0.113 | 0.114 | 0.108 | 0.108 | 0.112 | 0.110 | 0.111 | 0.112 |
| | $y_3$ | 0.083 | 0.072 | 0.073 | 0.074 | 0.073 | 0.073 | 0.077 | 0.069 | 0.069 | 0.072 |
| | $y_4$ | 0.036 | 0.031 | 0.031 | 0.032 | 0.032 | 0.032 | 0.033 | 0.030 | 0.031 | 0.031 |
| | $y_5$ | 0.197 | 0.192 | 0.192 | 0.197 | 0.175 | 0.176 | 0.196 | 0.155 | 0.155 | 0.168 |
| | $y_6$ | 0.106 | 0.094 | 0.094 | 0.096 | 0.091 | 0.091 | 0.098 | 0.090 | 0.090 | 0.094 |
| | $y_7$ | 0.047 | 0.043 | 0.043 | 0.048 | 0.032 | 0.032 | 0.046 | 0.027 | 0.027 | 0.036 |
| | $y_8$ | 0.083 | 0.081 | 0.080 | 0.080 | 0.088 | 0.088 | 0.084 | 0.089 | 0.089 | 0.087 |
| | $y_9$ | 0.000 | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.001 |
| | $y_{10}$ | 2.379 | 6.151 | 6.191 | 5.400 | 11.816 | 11.898 | 9.837 | 3.239 | 3.355 | 0.819 |
| Dep. | $y_1$ | 0.011 | 0.004 | 0.004 | 0.005 | 0.006 | 0.006 | 0.008 | 0.002 | 0.002 | 0.003 |
| | $y_2$ | 0.119 | 0.110 | 0.110 | 0.110 | 0.106 | 0.106 | 0.109 | 0.097 | 0.097 | 0.100 |
| | $y_3$ | 0.083 | 0.080 | 0.080 | 0.081 | 0.078 | 0.078 | 0.080 | 0.075 | 0.076 | 0.077 |
| | $y_4$ | 0.036 | 0.034 | 0.034 | 0.035 | 0.034 | 0.034 | 0.035 | 0.033 | 0.033 | 0.033 |
| | $y_5$ | 0.197 | 0.211 | 0.210 | 0.207 | 0.216 | 0.216 | 0.209 | 0.218 | 0.217 | 0.209 |
| | $y_6$ | 0.106 | 0.084 | 0.085 | 0.086 | 0.085 | 0.086 | 0.092 | 0.066 | 0.067 | 0.072 |
| | $y_7$ | 0.047 | 0.061 | 0.061 | 0.062 | 0.068 | 0.068 | 0.064 | 0.076 | 0.076 | 0.071 |
| | $y_8$ | 0.083 | 0.051 | 0.051 | 0.053 | 0.055 | 0.055 | 0.063 | 0.026 | 0.026 | 0.033 |
| | $y_9$ | 0.000 | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 |
| | $y_{10}$ | 2.379 | 2.238 | 2.193 | 2.306 | 13.331 | 13.499 | 19.138 | 6.116 | 6.124 | 0.200 |

Relative bias of the estimates in the real data simulation with the different covariates for the adjustments and unequal probability sampling proportional to age can be seen in Table 6. Adjustments perform generally poorly at reducing bias, except for $y_8$ (household with heating system, dichotomous) where TrIPW with deprivation covariates achieves a large reduction making it almost zero.

Table 6: Relative bias (*RelBias*) for each variable, adjustment method and covariates, in the real data experiment when drawing $s_v$ with unequal probability sampling (proportional to age) from the subpopulation having a computer at home.

| | | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cov. | Obj. | Unw | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| Dem. | $y_1$ | 0.011 | 0.013 | 0.014 | 0.018 | 0.013 | 0.013 | 0.018 | 0.012 | 0.012 | 0.018 |
| | $y_2$ | 0.119 | 0.124 | 0.123 | 0.122 | 0.124 | 0.124 | 0.122 | 0.131 | 0.130 | 0.126 |
| | $y_3$ | 0.083 | 0.204 | 0.204 | 0.215 | 0.204 | 0.204 | 0.213 | 0.173 | 0.173 | 0.182 |
| | $y_4$ | 0.036 | 0.074 | 0.074 | 0.076 | 0.075 | 0.075 | 0.075 | 0.072 | 0.072 | 0.072 |
| | $y_5$ | 0.197 | 0.426 | 0.423 | 0.351 | 0.365 | 0.362 | 0.320 | 0.603 | 0.594 | 0.493 |
| | $y_6$ | 0.106 | 0.184 | 0.184 | 0.180 | 0.193 | 0.193 | 0.183 | 0.255 | 0.255 | 0.233 |
| | $y_7$ | 0.047 | 0.076 | 0.077 | 0.090 | 0.078 | 0.079 | 0.097 | 0.090 | 0.092 | 0.096 |
| | $y_8$ | 0.083 | 0.041 | 0.042 | 0.039 | 0.047 | 0.048 | 0.043 | 0.062 | 0.063 | 0.060 |
| | $y_9$ | 0.000 | 0.002 | 0.002 | 0.003 | 0.001 | 0.001 | 0.002 | 0.000 | 0.000 | 0.002 |
| | $y_{10}$ | 2.379 | 6.943 | 6.755 | 12.126 | 29.566 | 29.440 | 3.778 | 50.925 | 50.706 | 38.733 |
| Dep. | $y_1$ | 0.011 | 0.012 | 0.012 | 0.013 | 0.012 | 0.012 | 0.015 | 0.002 | 0.002 | 0.005 |
| | $y_2$ | 0.119 | 0.118 | 0.118 | 0.118 | 0.113 | 0.113 | 0.116 | 0.106 | 0.106 | 0.109 |
| | $y_3$ | 0.083 | 0.243 | 0.243 | 0.243 | 0.241 | 0.241 | 0.243 | 0.239 | 0.239 | 0.241 |
| | $y_4$ | 0.036 | 0.080 | 0.080 | 0.080 | 0.080 | 0.080 | 0.080 | 0.080 | 0.080 | 0.080 |
| | $y_5$ | 0.197 | 0.310 | 0.309 | 0.300 | 0.319 | 0.318 | 0.299 | 0.360 | 0.359 | 0.332 |
| | $y_6$ | 0.106 | 0.153 | 0.153 | 0.155 | 0.155 | 0.155 | 0.163 | 0.131 | 0.131 | 0.138 |
| | $y_7$ | 0.047 | 0.091 | 0.091 | 0.092 | 0.094 | 0.095 | 0.093 | 0.110 | 0.111 | 0.105 |
| | $y_8$ | 0.083 | 0.012 | 0.012 | 0.014 | 0.015 | 0.015 | 0.021 | 0.004 | 0.004 | 0.002 |
| | $y_9$ | 0.000 | 0.003 | 0.003 | 0.003 | 0.005 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 |
| | $y_{10}$ | 2.379 | 8.167 | 8.128 | 10.613 | 17.879 | 17.923 | 21.377 | 12.155 | 12.150 | 18.735 |

Relative MSE of the estimates in the real data simulation with the different covariates for the adjustments and SRSWOR can be seen in Table 7. It is noticeable that, when using demographic covariates, there is always an efficient adjustment for each variable but the best one differs across variables. TrIPW works better for variables $y_2$ to $y_7$ but PSA with logistic regression for propensity estimation is the best option for $y_1$ (household expenses are a heavy burden, ordinal from 1 to 3) and $y_8$ to $y_{10}$ (household having a heating system and the two simulated independent random variables). The situation is similar when using deprivation covariates: TrIPW provides the largest efficiency for $y_2$ to $y_4$ (household having a car -dichotomous-, self-reporting health -5-point Likert scale- and presence of a disability -ordinal from 1 to 3-), $y_6$ (household expenses in EUR, numeric) and $y_8$, while PSA with logistic regression works better for the rest of variables. In both cases, it is important to note the impact of weight smoothing: when the efficiency is already below 1 (that is, the proposed method is more efficient than the unadjusted case), applying weight smoothing does not improve it, and can even contribute to increase the quotient. On the other hand, when the efficiency is above 1, weight smoothing helps to reduce the quotient and move it towards 1, with LASSO regression revealing itself as the best option in this case.

Table 7: Relative MSE (*RelMSE*) for each variable, adjustment method and covariates, in the real data experiment when drawing $s_v$ with SRSWOR from the subpopulation having a computer at home.

| | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cov. | Obj. | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| Dem. | $y_1$ | 1.229 | 1.187 | 1.123 | 1.624 | 1.591 | 1.243 | 2.252 | 2.101 | 1.803 |
| | $y_2$ | 0.906 | 0.911 | 0.918 | 0.832 | 0.837 | 0.890 | 0.868 | 0.873 | 0.900 |
| | $y_3$ | 0.772 | 0.779 | 0.797 | 0.785 | 0.780 | 0.876 | 0.706 | 0.715 | 0.764 |
| | $y_4$ | 0.783 | 0.786 | 0.810 | 0.799 | 0.783 | 0.879 | 0.750 | 0.745 | 0.794 |
| | $y_5$ | 0.988 | 1.018 | 1.021 | 0.900 | 0.970 | 1.030 | 0.754 | 0.804 | 0.805 |
| | $y_6$ | 0.793 | 0.816 | 0.825 | 0.752 | 0.795 | 0.859 | 0.731 | 0.778 | 0.803 |
| | $y_7$ | 0.894 | 0.930 | 0.947 | 0.930 | 0.986 | 0.986 | 0.883 | 0.916 | 0.921 |
| | $y_8$ | 0.953 | 0.936 | 0.950 | 1.100 | 1.034 | 1.008 | 1.154 | 1.098 | 1.095 |
| | $y_9$ | 0.915 | 0.907 | 0.901 | 1.189 | 1.142 | 1.097 | 1.185 | 1.139 | 1.095 |
| | $y_{10}$ | 0.917 | 0.912 | 0.889 | 1.097 | 1.079 | 0.997 | 1.154 | 1.136 | 1.060 |
| Dep. | $y_1$ | 0.497 | 0.513 | 0.517 | 0.620 | 0.651 | 0.693 | 0.694 | 0.665 | 0.659 |
| | $y_2$ | 0.852 | 0.865 | 0.862 | 0.800 | 0.806 | 0.848 | 0.673 | 0.694 | 0.707 |
| | $y_3$ | 0.945 | 0.949 | 0.959 | 0.899 | 0.890 | 0.933 | 0.812 | 0.812 | 0.850 |
| | $y_4$ | 0.943 | 0.939 | 0.949 | 0.937 | 0.906 | 0.951 | 0.840 | 0.814 | 0.854 |
| | $y_5$ | 1.168 | 1.162 | 1.135 | 1.234 | 1.283 | 1.167 | 1.241 | 1.265 | 1.141 |
| | $y_6$ | 0.646 | 0.681 | 0.669 | 0.669 | 0.715 | 0.770 | 0.417 | 0.472 | 0.486 |
| | $y_7$ | 1.191 | 1.191 | 1.185 | 1.464 | 1.487 | 1.319 | 1.719 | 1.706 | 1.549 |
| | $y_8$ | 0.457 | 0.475 | 0.484 | 0.526 | 0.526 | 0.639 | 0.232 | 0.244 | 0.286 |
| | $y_9$ | 1.030 | 1.022 | 1.013 | 1.240 | 1.186 | 1.140 | 1.154 | 1.105 | 1.046 |
| | $y_{10}$ | 1.030 | 1.028 | 1.030 | 1.166 | 1.144 | 1.120 | 1.093 | 1.080 | 1.035 |

Relative MSE of the estimates in the real data simulation with the different covariates for the adjustments and unequal probability sampling proportional to age can be seen in Table 8. Adjustments for nonprobability samples only work for $y_8$ (with PSA showing the best results) when using demographic covariates, and $y_1$ (household expenses are a heavy burden, ordinal from 1 to 3), $y_2$ (household has a car, dichotomous), $y_8$ (household has a heating system, dichotomous) and $y_9$ (independent Bernoulli random variable) with TrIPW showing the best results when using deprivation covariates. It is important to note that in this case, and in contrast to the SRSWOR scenario, weight smoothing worked in some of those favourable situations as well as many of the unfavourable ones, although its performance was limited when the estimator was largely inefficient. In those cases where weight smoothing worked well, LASSO provided the largest efficiency gains (except for $y_1$ and $y_2$ when using demographic covariates); however, in cases where weight smoothing did not help, LASSO sometimes performed worse than XGBoost.

Table 8: Relative MSE (*RelMSE*) for each variable, adjustment method and covariates, in the real data experiment when drawing $s_v$ with unequal probability sampling (proportional to age) from the subpopulation having a computer at home.

| | | PSA (Log. reg.) | | | PSA (k-NN) | | | TrIPW | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cov. | Obj. | NS | XGB | LASSO | NS | XGB | LASSO | NS | XGB | LASSO |
| Dem. | $y_1$ | 1.501 | 1.396 | 1.776 | 1.471 | 1.420 | 1.930 | 1.705 | 1.695 | 2.072 |
| | $y_2$ | 1.092 | 1.046 | 1.059 | 1.100 | 1.042 | 1.053 | 1.218 | 1.113 | 1.130 |
| | $y_3$ | 5.982 | 6.108 | 6.656 | 6.026 | 6.127 | 6.568 | 4.336 | 4.519 | 4.780 |
| | $y_4$ | 4.311 | 4.234 | 4.435 | 4.314 | 4.213 | 4.408 | 4.046 | 3.935 | 3.991 |
| | $y_5$ | 3.924 | 4.044 | 2.692 | 2.955 | 3.195 | 2.238 | 7.484 | 7.309 | 5.308 |
| | $y_6$ | 2.973 | 3.189 | 2.813 | 3.242 | 3.528 | 2.920 | 5.679 | 5.971 | 4.767 |
| | $y_7$ | 2.040 | 2.026 | 1.722 | 1.930 | 2.149 | 1.963 | 2.817 | 2.887 | 2.264 |
| | $y_8$ | 0.432 | 0.373 | 0.342 | 0.466 | 0.402 | 0.359 | 0.786 | 0.686 | 0.654 |
| | $y_9$ | 1.714 | 1.331 | 1.100 | 1.603 | 1.355 | 1.012 | 2.083 | 1.737 | 1.340 |
| | $y_{10}$ | 1.794 | 1.636 | 1.077 | 1.818 | 1.727 | 1.207 | 2.143 | 1.989 | 1.277 |
| Dep. | $y_1$ | 0.993 | 1.043 | 1.064 | 1.060 | 1.065 | 1.335 | 0.712 | 0.709 | 0.768 |
| | $y_2$ | 0.982 | 0.992 | 0.993 | 0.905 | 0.911 | 0.953 | 0.804 | 0.822 | 0.843 |
| | $y_3$ | 8.463 | 8.474 | 8.493 | 8.381 | 8.362 | 8.463 | 8.235 | 8.234 | 8.334 |
| | $y_4$ | 4.948 | 4.947 | 4.953 | 4.945 | 4.905 | 4.958 | 4.922 | 4.879 | 4.932 |
| | $y_5$ | 2.017 | 1.980 | 1.912 | 2.208 | 2.273 | 1.947 | 2.782 | 2.763 | 2.410 |
| | $y_6$ | 2.033 | 2.099 | 2.085 | 2.089 | 2.179 | 2.316 | 1.522 | 1.646 | 1.686 |
| | $y_7$ | 1.727 | 1.733 | 1.746 | 1.946 | 1.981 | 1.807 | 2.352 | 2.337 | 2.133 |
| | $y_8$ | 0.134 | 0.137 | 0.141 | 0.152 | 0.146 | 0.171 | 0.141 | 0.135 | 0.134 |
| | $y_9$ | 0.998 | 0.994 | 0.990 | 1.145 | 1.097 | 1.030 | 1.098 | 1.044 | 0.993 |
| | $y_{10}$ | 1.015 | 1.010 | 0.998 | 1.142 | 1.119 | 1.032 | 1.287 | 1.263 | 1.181 |

# 7 Discussion

This paper introduces the use of weight smoothing methods in the nonprobability survey context. Given its advantages in probability sampling, which were proven by Beaumont (2008), in terms of minimizing the variance induced by sampling design variables, it was expected that such methods could be used to address the issue of modeling propensities with prognostically irrelevant covariates. We designed two simulation studies which considered this approach in nonprobability sampling for both artificial and real data, using PSA and TrIPW to estimate propensities. The results show that weight smoothing contributes to reduce MSE in those situations where the efficiency of non-smoothed estimates is poor. These are the situations where the propensity models are not correctly specified or their covariates are weakly related to the variables of interest. When the estimates are already efficient, weight smoothing does not add much, with some exceptions where the set of covariates is good but not optimal.

In the real data simulation, there were some remarkable exceptions to the behavior described above. In some cases, the adjusted estimates (with PSA or TrIPW) were largely inefficient, but weight smoothing could not improve the results. The explanation could be that propensity adjustments can sometimes contribute to increasing the bias, rather than reducing it, when the models do not capture the actual relationships between covariates (Lee 2006; Ferri-García and Rueda 2020). In those cases, the MSE increases but not because of an increase in variance, meaning that weight smoothing is unable to make an impact as its focus is on reducing the noise.

Regarding weight smoothing, LASSO regression presented better results overall than XGBoost in terms of MSE reduction. LASSO regression involves variable selection, which can be particularly relevant in those cases where some target variables are more related to the weights than others. Finally, and despite the fact that the suitability of propensity estimation adjustments varies across situations, TrIPW provided better results than PSA in a majority of cases. It is also important to note that in those cases where TrIPW could not perform as well as PSA, weight smoothing succeeded at reducing MSE and getting TrIPW estimates to a similar level as PSA estimates. Therefore, a reasonable strategy could be to use TrIPW with further weight smoothing as a default choice when adjusting for selection bias in nonprobability surveys.

The present study has some limitations that should be noted. First, transformation of propensities in weights was performed only using one approach: the $w_i = 1/\pi_i$ formula from Valliant (2020). Other formulas might be useful in those contexts where the weights are less concentrated or have larger variability, and therefore propensity stratification could lead to smaller variances and larger efficiency values. Second, only two predictive algorithms were proposed in weight smoothing. There is currently a wide range of possibilities in Machine Learning literature regarding regression techniques that involve different approaches and paradigms, whose performance depends largely on the available dataset. Further studies could explore new applications of weight smoothing in nonprobability sampling where other regression techniques might provide better results. Finally, the data used for simulations covers a limited range of situations; for instance, the artificial data simulation only considered the most complicated scenario for weight smoothing (an U-shaped distribution for the vector of weights), and the real data simulation presented a situation where the selection bias was not extremely large. More realistic situation, such as right-skewed vectors of weights or more biased nonprobability samples, should be considered in future research on the matter.

# References

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

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

[3] Breiman L, Friedman J, Stone CJ, Olshen, RA (1984) Classification and regression trees. CRC press.

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

[5] Castro-Martín L, Rueda M, Ferri-García R (2020) Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques. Mathematics 8(6):879. doi:10.3390/math8060879

[6] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y (2020). xgboost: Extreme Gradient Boosting. R package version 1.0.0.2. Accessed March 2021. https://CRAN.R-project.org/package=xgboost

[7] Chen Y, Li P, Wu C (2020) Doubly Robust Inference With Non-probability Survey Samples. J Am Stat Assoc 115(532):2011-2021. doi:10.1080/01621459.2019.1677241

[8] Chu K, Beaumont J-F (2019) The Use Of Classification Trees To Reduce Selection Bias For A Non-Probability Sample With Help From A Probability Sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting, May 26-29, 2019. Calgary, Canada: Statistical Society of Canada. Available at: https://ssc.ca/sites/default/files/imce/survey_methods_4_-_the_use_of_classification_trees_to_reduce_selection_bias_for_a_non-probability_sample_with_help_from_a_probability_sample_chu_beaucmont-2019.pdf (accessed December 2020).

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

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

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

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

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

[14] Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw, 33(1):1-22.

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

[16] Haziza D, Beaumont JF (2017) Construction of weights in surveys: A review. Stat Sci 32(2):206-226.

[17] Hirano K, Imbens GW (2001) Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes research methodology 2(3-4):259-278.

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

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

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

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

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

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

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

[25] Neyman J (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. J R Stat Soc A 97:558-606.

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

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

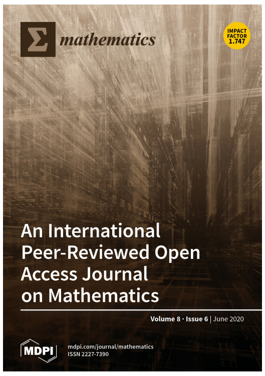
[28] Valliant R, Dever JA (2011) Estimating propensity adjustments for volunteer web surveys. Sociol Method Res 40(1):105-137.

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

# Appendix A7

## Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment

Ferri-García, Ramón; Rueda, María del Mar; Cabrera-León, Andrés (2021)
Self-Perceived Health, Life Satisfaction and Related Factors among Healthcare Professionals and the General Population: Analysis of an Online Survey, with Propensity Score Adjustment
*Mathematics*, vol. 9, number 7, p. 791
DOI: 10.3390/math9070791


mathematics
IMPACT FACTOR 1.747
An International Peer-Reviewed Open Access Journal on Mathematics
Volume 8 · Issue 6 | June 2020
MDPI
mdpi.com/journal/mathematics
ISSN 2227-7390

| MATHEMATICS | | | |
|---|---|---|---|
| JCR Year | Impact factor | Rank | Quartile |
| 2019 | 1.747 | 28/325 | Q1 |

**Abstract**

Healthcare professionals (HCPs) often suffer high levels of depression, stress, anxiety and burnout. Our main study aimswereto estimate the prevalences of poor self-perceived health, life dissatisfaction, chronic disease and unhealthy habits among HCPs and to explore the use of machine learning classification algorithms to remove selection bias. A sample of Spanish HCPs was asked to complete a web survey. Risk factors were identified by multivariate ordinal regression models. To counteract the absence of probabilistic sampling and representation, the sample was weighted by propensity score adjustment algorithms. The logistic regression algorithm was considered the most appropriate for dealing with misestimations. Male HCPs had significantly worse lifestyle habits than their female counterparts, together with a higher prevalence of chronic disease and of health problems. Members of the general population reported significantly poorer health and less satisfaction with life than the HCPs. Among HCPs, the prior existence of health problems was most strongly associated with worsening self-perceived health and decreased life satisfaction, while obesity had an important negative impact on female practitioners' self-perception of health. Finally, the HCPs who worked as nurses had poorer self-perceptions of health than other HCPs, and the men who worked in primary care had less satisfaction with their lives than those who worked in other levels of healthcare.

# 1 Introduction

One of the elements of the physician's pledge in the 2017 revision of the Declaration of Geneva, adopted by the World Medical Association (WMA), states: 'I will attend to my own health, well-being, and abilities in order to provide care of the highest standard [1]'. This addition to the previous Declaration of Geneva acknowledges that patients suffer when the well-being of healthcare professionals (HCPs) is compromised [2] and was adopted in response to the growing awareness that physicians and nurses present high levels of depression, stress, anxiety and burnout [3]. In fact, suicide is the only cause of death that has a higher prevalence among physicians than in the general population [4], and the situation among nurses is likely to be similar [5]. Moreover, the prevalence of substance abuse and/or addiction among physicians is likely to be similar to that found among the general public, or even higher [6].

The WMA recommends that more research be conducted into physicians' health and well-being and into the impact of these parameters on the patient care provided [7]. In view of these considerations, the main objectives of this research were to estimate the prevalence among HCPs of ill health, dissatisfaction, chronic disease and unhealthy lifestyle habits and to identify and analyse factors associated with life satisfaction and perceived health status.

We addressed these study goals by means of an online survey, an approach that offers substantial advantages over traditional survey techniques in terms of financial and time savings.Health surveys have traditionally used probability sampling of addresses and data collection facilitated by an interviewer who visits each address, but this traditional approach has some limitations, such as the great economic and temporal cost and the susceptibility to nonresponse bias. The main motivation for using nonprobability samples (as volunteer web surveys) is their low cost, lowrespondent burden and quick turnaround since they allow for producing estimates shortly after the information needs have been identified.

Although the validity of internet research for subjective surveys of personal well-being is well established [8] and online questionnaires are recognised as an important tool for epidemiological research [9], many surveys of this type are subject to self-selection [10,11]. Ref. [12] found in a health study that the bias in web surveys is too important, even when additional quotas are set. Statistical adjustments are the key to obtaining reliable estimates from online survey data. 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-randomised clinical trials [13], was adapted to nonprobability surveys in the work of [14,15]. PSA aims to estimate the propensity of each individual's participation in a survey by using logistic regression. [16] assessed the ability of PSA to remove bias in the context of sensitive sexual health research and the potential of web panel surveys to replace or supplement probability surveys.

Another goal of this research was to explore the use of machine learning (ML) classification algorithms to remove selection bias by reweighting the study variables via PSA. ML techniques are commonly employed in epidemiology [17,18,19], and statistical algorithms have been used to weight variables in recent health surveys [20,21,22].These techniques have also shown good properties in simulated data in terms of bias reduction [23,24] but at the cost of increasing the variance of the estimates. However, the mean square error (MSE), which combines bias and variance, is reduced with PSA in some situations, meaning that its application can be recommended in nonprobability sampling contexts. The objective of this study was to compare the performance and applicability of ML algorithms for PSA using several transformations to convert the probabilities provided by PSA into weights in a real-world context. This work pioneers the use of ML techniques to adjust the voluntary response bias in a real health survey and shows the capabilities of the different methods compared with the usual non-adjustment methodology.

# 2 Materials and methods

## 2.1 Target population

In 2014, according to census data, the Public Health System of Andalusia (SAS) employed 137,882 HCPs. However, for the purposes of this study, only those with a university degree were considered for inclusion, and so the target population was composed of the 73,465 HCPs who had this academic qualification.

## 2.2 Sample

In 2014, the participants in an online course on holistic care for patients with chronic diseases were asked to complete a web survey. These participants (n =

1797) were all university graduates working in the SAS as HCPs.

## 2.3 Variables

The following variables were present in both datasets (web survey and census): sex, age, degree and type of medical care provided (Table 1).

Table 1: Variables present in both datasets.

| Variable | Web Survey (%) | Census (%) |
|---|---|---|
| Sex | | |
| Male | 31.66 | 33.12 |
| Female | 68.34 | 66.88 |
| Age[1] | | |
| <25 | 1.11 | 2.16 |
| 26–35 | 21.09 | 26.96 |
| 36–45 | 34.78 | 25.74 |
| 46–55 | 32.78 | 24.50 |
| >55 | 10.24 | 20.65 |
| Healthcare area | | |
| Specialised | 50.97 | 66.74 |
| Primary | 49.03 | 33.26 |
| Degree subject area | | |
| Medicine | 43.80 | 40.44 |
| Nursing | 44.46 | 52.86 |
| Other | 11.74 | 6.69 |
| Valid sample | n = 1797 | n = 73,465 |

[1] Age data were not available for 383 individuals (0.52%) in the census data.

In addition to the variables presented in the table, the following variables were also addressed in the web survey:

- Self-perceived health status (scored on a 5-point Likert scale, ranging from 1 = very bad to 5 = very good)
- Satisfaction with life (scored on a 10-point Likert scale, ranging from 1 = completely unsatisfied to 10 = completely satisfied)
- Alcohol intake (once a day/once a week/once a month/less than once a month/never)
- Tobacco use (never/ex-smoker/occasional smoker/regular smoker)
- Physical activity (none/occasional/regular/intensive)
- Body mass index (BMI), obtained from dividing the weight (in kilograms) by the square of the height (in centimetres) and categorised as low or normal

weight (<25 kg/m2), overweight (25–29 kg/m2) and obesity (≥30 kg/m2) [25]

- Hours of sleep per night (numeric)
- Physical, mental or sensorial disability (presence/absence)
- Chronic disease (presence/absence)
- Health problems (none/one/two or more)

In order to make the prevalences of the healthcare professional survey comparable with those of the general population, the same categorisation and cut-off points of the Andalusian Health Survey [26] were applied for those study variables considered in both surveys, as follows: poor health ≤3 (i.e., fair, bad or very bad); dissatisfaction with life ≤6; ≥1 alcoholic drink per month; and insufficient sleep <7 h of sleep per night.

## 2.4 Sampling Weights

As shown in Table 1, HCPs aged 36–55 years were over-represented in the web survey sample with respect to the target population as well as to primary care HCPs. On the other hand, there was an under-representation of HCPs with a degree in nursing. Given a volunteer survey $s_v$, the usual estimator of the population proportion is the Horvitz–Thompson estimator given by

$$p_{ht} = \frac{1}{N} \sum_{i \in s_v} A_i w_i \tag{1}$$

where $A_i = 1$ if the unit $i$ in the sample $s_v$ has the desired characteristics and 0 else, and $w_i$ is the weight (the inverse of the sampling rate).

To adjust for the lack of probability sampling and the resulting non-representativeness, the sample was weighted, using the standard procedure of propensity score adjustment (PSA) for web surveys [14,15].

This approach aims to estimate the propensity of an individual to be included in the nonprobability sample by combining the data from the sample $s_v$ with a reference probability sample $s_r$ and training a predictive model on the variable $\delta$, 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:

$$P(\delta_i = 1 | x_i) = \pi(x_i, \gamma) \tag{2}$$

for some function $\pi$ of the observed covariates $x_i$ and a parameter $\gamma$. The usual procedure is to estimate the parameter $\gamma$ by using logistic regression and to transform the estimated propensities to weights by inverting them:

$$p_{PSA1} = \frac{1}{\sum_{i \in s_v} 1/\pi(\hat{x_i})} \sum_{i \in s_v} A_i * 1/\pi(\hat{x_i}) \tag{3}$$

where $\hat{\pi(x_i)}$ denotes the estimated propensity for the individual $i \in s_v$. This transformation is equivalent to the Hajek estimator of the population proportion. An alternative that takes into account the fact that individuals of $s_v$ must be excluded from the target population of $s_r$ is the formula presented in [27]:

$$p_{PSA2} = \frac{1}{\sum_{i \in s_v} (1 - \hat{\pi(x_i)})/\hat{\pi(x_i)}} \sum_{i \in s_v} A_i * (1 - \hat{\pi(x_i)})/\hat{\pi(x_i)} \tag{4}$$

We considered the following algorithms for estimating the aforementioned propensities:

- Logistic regression
- Decision trees (C5.0 algorithm [28])
- The k-nearest neighbours algorithm, with k = 5 (5-NN)
- Naïve Bayes with no Laplace smoothing
- Random forest with 500 trees
- Gradient boosting machine (GBM) with 100 trees, interaction depth of 1 and learning rate of 0.1
- Feed-forward neural networks with one hidden layer, initialising weights to 0 and considering three cases with 1, 3 and 5 units in the hidden layer

In all cases, the probabilities calculated in PSA were transformed into weights for Hajek estimators, following the formula for $p_{PSA2}$ stated in [27]. Weights for Horvitz–Thompson estimators were also calculated, in accordance with [15]. PSA was performed in R 3.1.5 [29] using the packages sampling [30], survey [31], C50 [32], randomForest [33], gbm [34], e1071 [35], caret [36] and nnet [37]. The weights for the Horvitz–Thompson estimators were discarded, as they were unstable and produced unacceptably high variances. In general, the Horvitz–Thompson weights, although they correlated with the Hajek weights obtained by the same methods, presented higher levels of skewness, probably caused by the grouping features of the weighting method (see Appendix A). Moreover, the weights obtained by PSA using decision trees and neural networks with five units were also discarded, as they were found to be equal to the design weights and so provided the same outputs as in the unadjusted case.

## 2.5 Statistical Analysis

Several weights were applied in estimating the prevalence of each of the variables considered. To reflect potential differences between male and female HCPs in these prevalence values, sex was taken as a stratification variable. The variances of the proportion estimators were calculated using the leave-one-out jackknifealgorithm [38], implemented in the bootstrap package in R [39]. Prevalence values for the study population were compared with those for the general population [26] in the same age range (22–67 years).

Multivariate ordinal logistic regression models were run to characterise the ordinal variables of life satisfaction and self-perceived health status. Sampling weights were applied in the models, which were constructed independently for male and female HCPs. In the statistical analysis, the scales for life satisfaction and self-perceived health status were inverted; thus, odds ratios (OR) $>$ 1 mean that the explanatory variable increases the probability of dissatisfaction with life or of poor self-perceived health. In addition, those reference categories of the explanatory variables which obtained a better interpretation of odds ratios (i.e., OR $>$ 1) were chosen. The following explanatory variables were included in the models:

- Health problems (none/one/two or more)
- Tobacco use (never/ex-smoker/occasional smoker/regular smoker)
- Hours of sleep per night ($<$7 h/$\geq$7 h)
- Physical activity (none/occasional/regular/intensive)
- Body mass index (BMI), categorised as low or normal weight ($<$25 kg/m2), overweight (25–29 kg/m2) and obesity ($\geq$30 kg/m2) [25]
- Level of healthcare (Primary/ Other)
- Age in years (numeric)
- Degree (Medicine/Nursing/Other)

Multicollinearity of the independent variables was assessed using the variance inflation factor (VIF) [40], which indicates collinearity if the factor takes large values. The factor was discarded for VIF $>$ 3 [41]. Therefore,‘chronic diseases’ and ‘physical, mental or sensorial disability’ were not included in the final model. Alcohol consumption was also excluded because of its low association with the dependent variables of the models, which was assessed with a preliminary regression analysis where the alcohol variable was not significant and had a beta coefficient around zero. The rest of the coefficients and test statistics remained almost unchanged with respect to the case without the alcohol consumption variable.To observe the range of values in which the coefficients would be applicable to the entire population, 95% confidence intervals were calculated. Hypothesis testing of the beta coefficients was performed with the Wald test. Statistical and graphical analyses were performed in R 3.5.1 using the packages poliscidata [42] and ggplot2 [43], respectively, in addition to those mentioned above.

# 3 Results

## 3.1 Prevalence Estimations

According to results provided by PSA with logistic regression,10.3% of male HCPs (Table 2) and 12.6% of female HCPs (Table 3) were dissatisfied with their life and 8.4% of male and 7.8% of female professionals perceived their own health as poor. Regarding lifestyle habits, 62.3% of the men and 42.8% of the women drank alcohol at least once a week, while 31.1% of the men and 26.7% of the women slept for less than seven hours a day. Finally, 31.8% of the men and 22.3% of the women reported havingat least one chronic disease. Moreover, 26.3% of the men and 20.6% of the women had one health problem, 10.4% and 6%, respectively, had two or more health problems, and 7% of men and 6% of women had a disability (Table 2 and Table 3).

Table 2: Point estimate, variance and difference from the non-adjusted case of estimators of prevalence in male healthcare professionals (HCPs) for each propensity score adjustment (PSA) (algorithms are sorted from the least to the most complex).

| Algorithm Used in PSA | Poor Self-Perceived Health | | | | Dissatisfied with Life (Score of 6 or Less) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.088 | 0.00014 | Estimate | Variance | 0.1002 | 0.00016 | Estimate | Variance |
| Logistic regression | 0.084 | 0.00016 | -4.34% | 17% | 0.1031 | 0.00023 | 2.93% | 45% |
| Neural net (1 unit) | 0.087 | 0.00023 | -0.58% | 62% | 0.1090 | 0.00043 | 8.84% | 174% |
| Algorithm Used in PSA | Alcohol once a week | | | | <7 h of sleep | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.6232 | 0.00041 | Estimate | Variance | 0.3093 | 0.00038 | Estimate | Variance |
| Logistic regression | 0.6234 | 0.00053 | 0.02% | 29% | 0.3118 | 0.00049 | 0.82% | 30% |
| Neural net (1 unit) | 0.6004 | 0.00085 | -3.66% | 106% | 0.3395 | 0.00085 | 9.76% | 126% |
| Algorithm Used in PSA | Disability (physical. mental or sensorial) | | | | Chronic disease | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.0645 | 0.00011 | Estimate | Variance | 0.3369 | 0.00040 | Estimate | Variance |
| Logistic regression | 0.0695 | 0.00016 | 7.74% | 46% | 0.3179 | 0.00048 | -5.63% | 19% |
| Neural net (1 unit) | 0.0584 | 0.00015 | -9.42% | 43% | 0.3065 | 0.00065 | -9.03% | 63% |
| Algorithm Used in PSA | One health problem | | | | Two or more health problems | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.2742 | 0.00036 | Estimate | Variance | 0.1072 | 0.00017 | Estimate | Variance |
| Logistic regression | 0.2630 | 0.00044 | -4.09% | 22% | 0.1037 | 0.00019 | -3.23% | 13% |
| Neural net (1 unit) | 0.2361 | 0.00054 | -13.90% | 51% | 0.1048 | 0.00024 | -2.22% | 41% |

Table 3: Point estimate, variance and difference from the non-adjusted case of estimators of prevalence in female HCPs for each propensity score adjustment (PSA) (algorithms are sorted from the least to the most complex).

| Algorithm Used in PSA | Poor Self-Perceived Health | | | | Dissatisfied with Life (Score of 6 or Less) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.0839 | 0.00006 | Estimate | Variance | 0.1205 | 0.00009 | Estimate | Variance |
| Logistic regression | 0.0784 | 0.00007 | -6.49% | 15% | 0.1261 | 0.00012 | 4.61% | 39% |
| Neural net (1 unit) | 0.0720 | 0.00008 | -14.21% | 29% | 0.1270 | 0.00019 | 5.36% | 114% |
| Algorithm Used in PSA | Alcohol once a week | | | | <7 h of sleep | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.4223 | 0.00020 | Estimate | Variance | 0.2671 | 0.00016 | Estimate | Variance |
| Logistic regression | 0.4275 | 0.00026 | 1.23% | 30% | 0.2670 | 0.00021 | -0.03% | 29% |
| Neural net (1 unit) | 0.4281 | 0.00039 | 1.39% | 95% | 0.2547 | 0.00028 | -4.64% | 79% |
| Algorithm Used in PSA | Disability (physical. mental or sensorial) | | | | Chronic disease | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.0628 | 0.00005 | Estimate | Variance | 0.2230 | 0.00014 | Estimate | Variance |
| Logistic regression | 0.0602 | 0.00006 | -4.14% | 23% | 0.2228 | 0.00019 | -0.05% | 29% |
| Neural net (1 unit) | 0.0581 | 0.00008 | -7.46% | 74% | 0.2253 | 0.00029 | 1.05% | 99% |
| Algorithm Used in PSA | One health problem | | | | Two or more health problems | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.2122 | 0.00014 | Estimate | Variance | 0.0562 | 0.00004 | Estimate | Variance |
| Logistic regression | 0.2056 | 0.00017 | -3.13% | 22% | 0.0601 | 0.00006 | 6.95% | 50% |
| Neural net (1 unit) | 0.2095 | 0.00025 | -1.26% | 83% | 0.0568 | 0.00010 | 1.17% | 123% |

Figure A8 and Figure A9 of Appendix B show the 95% confidence intervals for the prevalence of each of the variables considered. All of the estimations were very similar, whichever method was applied, although some point estimates varied slightly due to the influence of certain algorithms on the propensity estimation step. In consequence, there were no statistical differences between the prevalences estimated among any of the weighting methods applied. The logistic regression algorithm obtained the best results in terms of both prevalence and variance deviations compared with no weighting adjustment (see Table A3 and Table A4 of Appendix B). As stated before, PSA contributed to increasing the variance of the estimators but reduced their bias, meaning that the estimates based in PSA might be more valuable as they mitigated the effect of non-sampling errors in the final estimates. Given that the estimates provided by PSA with different algorithms were very similar (and therefore might reduce the bias in the same amount), the choice that reduced MSE to the minimum extent might be the estimate with the lowest variance.

Table 4 shows the prevalences of the study variables for the general population [26] and the HCPs. The latter group self-reported significantly better health and greater satisfaction with life than the general population. In addition, while women in the general population reporteda significantly worse perception of their health than men (17.5% and 12.1%, respectively, reported poor health), female HCPs had a better, although non-significant, perception in this respect, compared with their male counterparts (7.8% and 8.5%, respectively). On the contrary, women reported significantly less satisfaction with their life than men, both those in the general population (19.2% vs. 16.3%, respectively) and among the HCPs (12.6% vs. 10.3%, respectively).

Table 4: Prevalence of the study variables in the general population [26] and in healthcare professionals according to survey data (Andalusia).

| Study Variables | | General Population | | Healthcare Professionals (Weighted with Propensity Score Adjustment Using Logistic Regression) | |
|---|---|---|---|---|---|
| | | % | 95% CI | % | 95% CI |
| Poor self-perceived health (fair/bad/very bad) in the last 12 months | Total | 14.8 | (13.5; 16) | 8.1 | (6.7; 9.5) |
| | Men | 12.1 | (10.6; 14) | 8.4 | (5.9; 1.,9) |
| | Women | 17.5 | (15.6; 19) | 7.8 | (6.2; 9.5) |
| Dissatisfied with life (6 or less on a scale from 1 to 10) | Total | 17.8 | (16.2; 20) | 10.7 | (9.2; 12.3) |
| | Men | 16.3 | (14.6; 18) | 10.3 | (7.3; 13.3) |
| | Women | 19.2 | (17.1; 21) | 12.6 | (10.5; 14.8) |
| Alcohol consumption (at least once in a month) | Total | 49.5 | (47; 52) | 66.4 | (63.9; 68.8) |
| | Men | 62.5 | (59.9; 65) | 79.8 | (76.1; 83.5) |
| | Women | 37.1 | (33.7; 41) | 60.0 | (56.9; 63.1) |
| Less than 7 h of sleep | Total | 20 | (17.8; 22) | 27.9 | (25.6; 30.3) |
| | Men | 17.7 | (15.3; 20) | 31.2 | (26.8; 35.5) |
| | Women | 22.1 | (19.7; 25) | 26.7 | (23.9; 29.5) |
| Presence of a chronic disease | Total | 40.7 | (38.6; 43) | 26.6 | (24.2; 28.9) |
| | Men | 35.9 | (33.6; 38) | 31.8 | (27.5; 36.1) |
| | Women | 45.3 | (42.7; 48) | 22.3 | (19.6; 25) |
| Physical, mental or sensorial disability | Total | 3.54 | (2.94; 4) | 6.0 | (4.8; 7.2) |
| | Men | 3.95 | (3.16; 5) | 7.0 | (4.5; 9.4) |
| | Women | 3.16 | (2.45; 4) | 6.0 | (4.5; 7.5) |

With respect to alcohol consumption (at least once in a month), the men in the general population and among HCPs reportedsignificantly higher prevalencesthan women. In addition, alcohol consumption was significantly more prevalent among male and female HCPs than among men and women in the general population (79.8% and 60%, 62.5% and 37.1%, respectively). Regarding hours of sleep per day, significantly more HCPs than persons in the general population slept for less than 7 h. This difference was especially marked among men (31.2% vs. 17.7%, respectively). In addition, significantly more male than female HCPs slept for less than 7 h per day (31.2% vs. 26.7%, respectively), which is contrary to the pattern observed in the general population.

The presence of chronic disease was much more prevalent among women in the general population than among female HCPs (45.3% vs. 22.3%, respectively), but no such difference was observed between the two groups of men (35.9% vs. 31.8%, respectively). The prevalence of disability was almost twice as high among HCPs as in the general population (6% vs. 3.5%, respectively). In this respect, there were no differences between men and women.

## 3.2 Regression Modelling

As described above, the regression modelling was performed using three types of weighting: no adjustment, PSA using logistic regression for prevalence estimation and PSA using a neural net with one unit for prevalence estimation. These weighting methods were selected taking into account the low degree of variability among them, which means that one or more could be discarded if necessary to avoid redundancy (see Appendix A for further information on the similarity among weights).

In almost every case, the strength of evidence against the explanatory variable having a null effect weakened with reweighting, not only because the variance increased (for example, with larger confidence intervals) but also when the beta coefficient shifted towards zero (or towards one; see Table 5, Table 6, Table 7 and Table 8). In other words, when reweighting was performed, it merely addressed misestimation of the association between explanatory variables, caused by the nonprobabilistic sampling method applied in the survey.

Table 5 and Table 6 depict the results for the models assessing self-perceived health, and Table 7 and Table 8 depict those concerning satisfaction with life. Figure 1 and Figure 2 illustrate the OR for self-perceived health and satisfaction with life, respectively, for male and female participants.

Table 5: Regression models for poorer self-perceived health among men according to each weighting adjustment method. Reference classes for categorical variables: no health problems, never smoked, $\geq$7 h of sleep, physical exercise several days a week, normal weight or underweight, working in a specialised field of healthcare and degree in medicine ($n = 558$ observations, Nagelkerke $R^2$=0.281).

| | No PSA Adjustment | | PSA with Logistic Regression | | PSA with Neural Net (1 Unit) | |
|---|---|---|---|---|---|---|
| Predictors | Odds ratio | 95% CI | Odds ratio | 95% CI | Odds ratio | 95% CI |
| 1—2 intercept | 7.44 | 2.71–20.4 | 9.43 | 3.26–27.3 | 10.15 | 2.90–35.5 |
| 2—3 intercept | 279.11 | 249–313 | 331.25 | 289–380 | 314.56 | 264–375 |
| 3—4 intercept | 2411.0 | 1733–3354 | 2979.3 | 2078–4273 | 3151.0 | 2100–4728 |
| 4—5 intercept | 5792.5 | 2086–16,085 | 7636.5 | 2635–22,132 | 6489.2 | 1890–22,276 |
| One health problem | 3.23 | 2.14–4.86 | 2.82 | 1.78–4.45 | 2.59 | 1.53–4.38 |
| Two or more health problems | 8.31 | 4.11–16.8 | 7.24 | 2.99–17.6 | 7.18 | 1.79–28.9 |
| Daily smoker | 1.30 | 0.66–2.58 | 1.46 | 0.72–2.96 | 1.43 | 0.67–3.05 |
| Non-daily smoker | 0.45 | 0.18–1.12 | 0.39 | 0.16–1.00 | 0.19 | 0.07–0.50 |
| Ex-smoker | 0.88 | 0.59–1.31 | 0.85 | 0.54–1.33 | 0.60 | 0.34–1.05 |
| <7 h of sleep | 1.78 | 1.23–2.59 | 1.83 | 1.19–2.81 | 1.95 | 1.17–3.26 |
| No physical activity at all | 2.94 | 1.36–6.35 | 2.76 | 1.02–7.43 | 1.98 | 0.32–12.3 |
| Occasional physical activity | 1.60 | 1.03–2.46 | 1.65 | 1.01–2.69 | 1.78 | 1–3.17 |
| Regular physical activity | 1.36 | 0.84–2.19 | 1.36 | 0.81–2.28 | 1.45 | 0.78–2.71 |
| Obesity | 1.39 | 0.78–2.49 | 1.40 | 0.71–2.77 | 1.78 | 0.81–3.95 |
| Overweight | 1.50 | 1.01–2.22 | 1.50 | 0.96–2.36 | 1.60 | 0.93–2.75 |
| Age (5 years) | 1.14 | 1.02–1.27 | 1.16 | 1.02–1.31 | 1.17 | 1.00–1.36 |
| Primary care | 1.16 | 0.77–1.75 | 1.24 | 0.78–1.98 | 1.37 | 0.77–2.44 |
| Nursing degree | 1.91 | 1.33–2.74 | 1.85 | 1.25–2.76 | 1.86 | 1.20–2.89 |
| Other degree | 0.92 | 0.46–1.87 | 1.07 | 0.51–2.27 | 1.17 | 0.51–2.67 |

Table 6: Regression models for poorer self-perceived health among women according to each weighting adjustment method. Reference classes for categorical variables: no health problems, never smoked, $\geq$7 h of sleep, physical exercise several days a week, normal weight or underweight, working in a specialised field of healthcare and degree in medicine ($n = 1211$ observations, Nagelkerke $R^2$=0.23).

| | No PSA Adjustment | | PSA with Logistic Regression | | PSA with Neural Net (1 Unit) | |
|---|---|---|---|---|---|---|
| Predictors | Odds ratio | 95% CI | Odds ratio | 95% CI | Odds ratio | 95% CI |
| 1—2 intercept | 6.96 | 3.65–13.2 | 6.7 | 3.24–13.8 | 7.2 | 2.97–17.5 |
| 2—3 intercept | 252.80 | 234–273 | 242.22 | 222–264 | 253.71 | 230–280 |
| 3—4 intercept | 2705.6 | 2141–3419 | 2481.2 | 1886–3264 | 2252.8 | 1643–3088 |
| 4—5 intercept | 6655.2 | 3093–14,319 | 5758.5 | 2337–14,191 | 4897.8 | 1816–13,210 |
| One health problem | 2.27 | 1.64–3.14 | 1.90 | 1.33–2.72 | 1.76 | 1.15–2.70 |
| Two or more health problems | 10.81 | 6.22–18.8 | 10.25 | 5.32–19.8 | 10.15 | 4.91–21.0 |
| Daily smoker | 1.54 | 1.07–2.23 | 1.64 | 1.10–2.45 | 1.60 | 1.02–2.51 |
| Non-daily smoker | 1.56 | 0.98–2.51 | 1.59 | 0.96–2.64 | 1.46 | 0.83–2.59 |
| Ex-smoker | 0.93 | 0.71–1.22 | 0.96 | 0.71–1.29 | 0.99 | 0.70–1.41 |
| <7 h of sleep | 1.27 | 0.97–1.65 | 1.46 | 1.09–1.97 | 1.53 | 1.10–2.13 |
| No physical activity at all | 1.94 | 1.25–3.00 | 1.54 | 0.93–2.55 | 1.48 | 0.82–2.65 |
| Occasional physical activity | 1.50 | 1.13–2.00 | 1.43 | 1.04–1.97 | 1.47 | 1.02–2.11 |
| Regular physical activity | 1.17 | 0.842–1.64 | 1.13 | 0.78–1.65 | 1.10 | 0.72–1.69 |
| Obesity | 2.14 | 1.23–3.72 | 2.10 | 1.10–4.02 | 1.84 | 0.81–4.20 |
| Overweight | 1.39 | 1.04–1.85 | 1.27 | 0.91–1.77 | 1.16 | 0.81–1.67 |
| Age (5 years) | 1.19 | 1.11–1.28 | 1.18 | 1.09–1.28 | 1.19 | 1.07–1.32 |
| Primary care | 1.24 | 0.95–1.60 | 1.21 | 0.92–1.59 | 1.29 | 0.98–1.70 |
| Nursing degree | 1.67 | 1.29–2.16 | 1.78 | 1.33–2.38 | 1.87 | 1.36–2.56 |
| Other degree | 1.93 | 1.30–2.88 | 1.99 | 1.31–3.03 | 2.20 | 1.45–3.33 |

The strongest OR for poor self-perceived health was obtained when the respondent had one or more pre-existing health problems. Thus, the prior existence of one health problem increased the likelihood of poor health by 3 and 2 times, respectively, for men and women. In the case of two or more health problems, this probability rose to 8 and 10 times, respectively, see Table 5 and Table 6. In addition, there was evidence that the presence of obesity, according to the BMI index, was significantly associated with a lower probability of good health among women (OR = 2.1).

Regarding the type of university degree held, nursing qualifications were significantly associated with poorer self-perceived health, compared with respondents with a degree in medicine, regardless of sex (OR = 1.8), or even among women those whose degree subject was reported as neither medicine nor nursing (OR = 2). However, no significant differences in OR were observed between those who worked in primary care or other level of healthcare.

In relation to lifestyle habits, smoking every day was associated with a greater likelihood of poorer self-perceived health in women; no physical activity or only occasional activity was also associated with poorer self-perception of health, especially in men, as was sleeping less than seven hours per night.

Figure 1: Confidence intervals at 95% for the odds ratio for each explanatory variable on self-perception of health, using logistic regression for the propensity score adjustment. Reference classes for categorical variables: no health problems, never smoked, ≥7 h of sleep, physical exercise several days a week, normal weight or underweight, specialised field of healthcare and degree in medicine. The x axis scale is logarithmic to facilitate interpretation of the data.

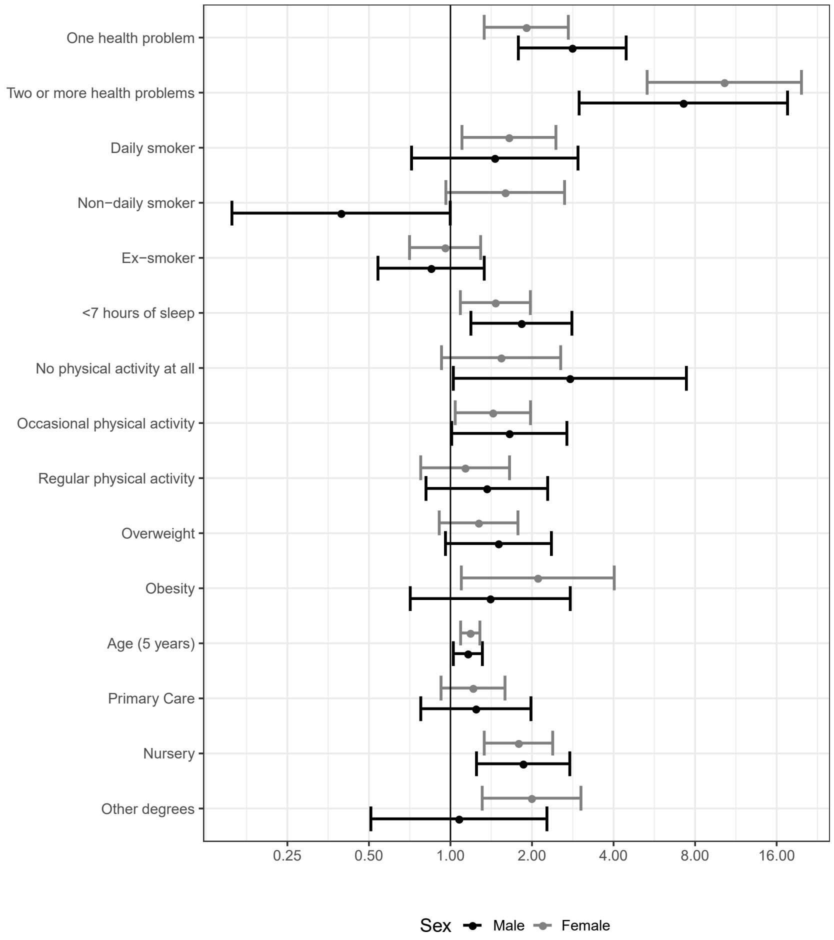


Table 7: Regression models for poorer self-perceived life satisfaction among men according to each weighting adjustment method. Reference classes for categorical variables: no health problems, never smoked, $\geq 7$ h of sleep, physical exercise several days a week, normal weight or underweight, specialised field of healthcare and degree in medicine ($n = 558$ observations, Nagelkerke $R^2 = 0.266$).

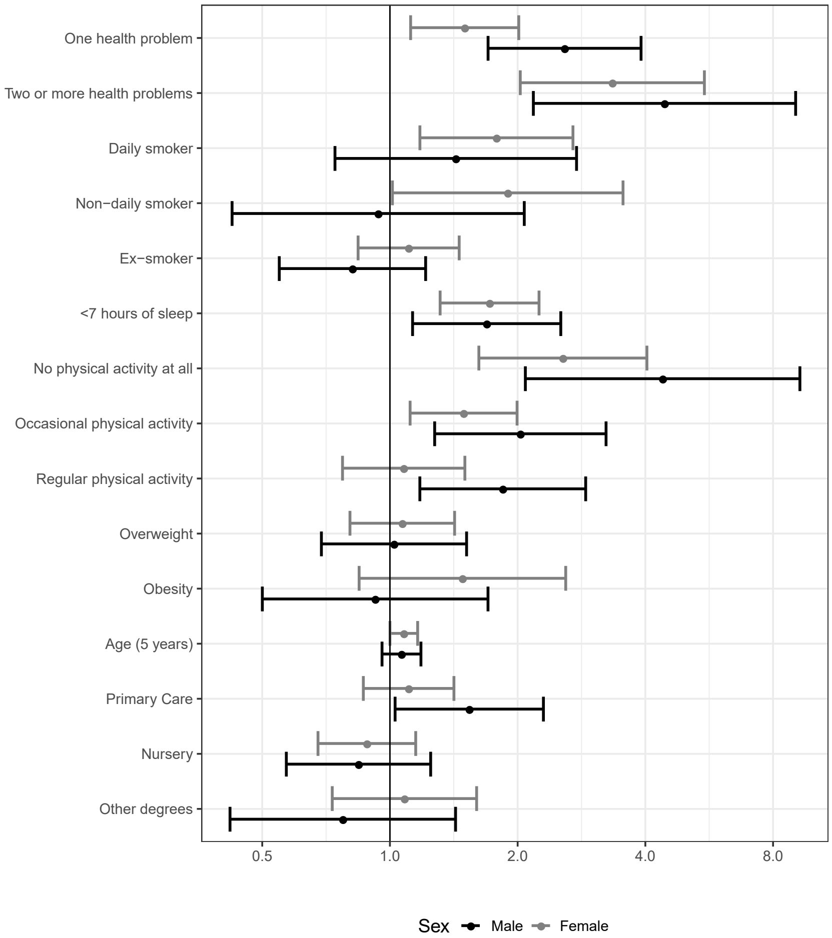
| | No PSA Adjustment | | PSA with Logistic Regression | | PSA with Neural Net (1 Unit) | |
|---|---|---|---|---|---|---|
| Predictors | Odds ratio | 95% CI | Odds ratio | 95% CI | Odds ratio | 95% CI |
| 1—2 intercept | 0.35 | 0.14–0.87 | 0.21 | 0.08–0.57 | 0.24 | 0.07–0.80 |
| 2—3 intercept | 3.05 | 2.60–3.58 | 2.43 | 2.05–2.89 | 3.25 | 2.69–3.93 |
| 3—4 intercept | 18.54 | 16.4–20.9 | 15.01 | 13.1–17.1 | 20.24 | 17.4–23.6 |
| 4—5 intercept | 91.95 | 77.6–109 | 75.94 | 63.1–91.4 | 100.31 | 80.4–125 |
| 5—6 intercept | 227.57 | 162–320 | 180.36 | 121–269 | 255.27 | 156–418 |
| 6—7 intercept | 509.65 | 296–877 | 442.63 | 245–799 | 577.23 | 300–1111 |
| 7—8 intercept | 1597.5 | 776–3290 | 1938.1 | 975–3852 | 2483.3 | 1187–5196 |
| 8—9 intercept | 1597.6 | 612–4175 | 2281.3 | 838–6212 | 2919.1 | 1047–8136 |
| 9—10 intercept | 3223.0 | 790–13,165 | 4045.2 | 853–19,181 | 4846.0 | 993–23,640 |
| One health problem | 2.60 | 1.79–3.77 | 2.58 | 1.70–3.91 | 2.65 | 1.67–4.20 |
| Two or more health problems | 3.98 | 2.13–7.44 | 4.44 | 2.18–9.04 | 3.65 | 1.38–9.70 |
| Daily smoker | 1.53 | 0.84–2.76 | 1.43 | 0.74–2.75 | 1.41 | 0.69–2.89 |
| Non-daily smoker | 0.91 | 0.41–2.03 | 0.94 | 0.43–2.07 | 0.74 | 0.29–1.89 |
| Ex-smoker | 0.81 | 0.57–1.16 | 0.82 | 0.55–1.21 | 0.65 | 0.40–1.07 |
| <7 h of sleep | 1.51 | 1.07–2.14 | 1.69 | 1.13–2.53 | 1.87 | 1.15–3.05 |
| No physical activity at all | 5.10 | 2.73–9.51 | 4.39 | 2.09–9.26 | 3.69 | 1.26–10.8 |
| Occasional physical activity | 1.95 | 1.29–2.96 | 2.03 | 1.27–3.23 | 2.09 | 1.21–3.61 |
| Regular physical activity | 1.94 | 1.30–2.90 | 1.84 | 1.18–2.89 | 1.94 | 1.12–3.35 |
| Obesity | 0.99 | 0.58–1.70 | 0.92 | 0.50–1.70 | 1.03 | 0.52–2.02 |
| Overweight | 1.20 | 0.83–1.73 | 1.02 | 0.69–1.52 | 1.07 | 0.68–1.68 |
| Age (5 years) | 1.08 | 0.98–1.19 | 1.06 | 0.96–1.18 | 1.11 | 0.98–1.27 |
| Primary care | 1.43 | 1.00–2.05 | 1.54 | 1.03–2.30 | 1.42 | 0.89–2.25 |
| Nursing degree | 1.00 | 0.71–1.41 | 0.84 | 0.57–1.25 | 0.75 | 0.49–1.15 |

Table 8: Regression models for poorer self-perceived life satisfaction among women according to each weighting adjustment method. Reference classes for categorical variables: no health problems, never smoked, $\geq$7 h of sleep, physical exercise several days a week, normal weight or underweight, specialised field of healthcare and degree in medicine ($n = 1211$ observations, Nagelkerke $R^2 = 0.159$).

| | No PSA Adjustment | | PSA with Logistic Regression | | PSA with Neural Net (1 Unit) | |
|---|---|---|---|---|---|---|
| Predictors | Odds ratio | 95% CI | Odds ratio | 95% CI | Odds ratio | 95% CI |
| 1—2 intercept | 0.23 | 0.12–0.42 | 0.24 | 0.12–0.49 | 0.26 | 0.11–0.59 |
| 2—3 intercept | 1.48 | 1.32–1.66 | 1.52 | 1.33–1.72 | 1.58 | 1.37–1.83 |
| 3—4 intercept | 7.20 | 6.62–7.84 | 7.55 | 6.86–8.31 | 7.93 | 7.08–8.88 |
| 4—5 intercept | 32.10 | 28.8–35.8 | 35.79 | 31.7–40.5 | 40.47 | 35.0–46.8 |
| 5—6 intercept | 79.91 | 64.3–99.3 | 81.36 | 63.5–104 | 88.66 | 66.7–118 |
| 6—7 intercept | 177.46 | 127–248 | 185.18 | 127–269 | 194.14 | 127–298 |
| 7—8 intercept | 437.33 | 273–700 | 441.97 | 259–755 | 411.03 | 223–758 |
| 8—9 intercept | 820.65 | 365–1847 | 938.97 | 396–2224 | 949.42 | 385–2344 |
| 9—10 intercept | 2901.2 | 1147–7338 | 2710.6 | 892–8233 | 2447.5 | 690–8684 |
| One health problem | 1.57 | 1.19–2.07 | 1.50 | 1.12–2.01 | 1.55 | 1.13–2.13 |
| Two or more health problems | 3.71 | 2.33–5.92 | 3.34 | 2.03–5.51 | 3.74 | 2.20–6.33 |
| Daily smoker | 1.83 | 1.25–2.66 | 1.78 | 1.18–2.70 | 1.57 | 0.99–2.49 |
| Non-daily smoker | 1.92 | 1.13–3.25 | 1.90 | 1.01–3.54 | 1.40 | 0.71–2.77 |
| Ex-smoker | 1.21 | 0.94–1.54 | 1.11 | 0.84–1.46 | 1.07 | 0.79–1.47 |
| <7 h of sleep | 1.80 | 1.42–2.30 | 1.72 | 1.31–2.25 | 1.86 | 1.34–2.58 |
| No physical activity at all | 2.47 | 1.65–3.70 | 2.56 | 1.62–4.03 | 2.17 | 1.33–3.56 |
| Occasional physical activity | 1.57 | 1.21–2.04 | 1.49 | 1.12–1.99 | 1.28 | 0.93–1.77 |
| Regular physical activity | 1.10 | 0.82–1.50 | 1.08 | 0.77–1.50 | 1.05 | 0.71–1.55 |
| Obesity | 1.54 | 0.94–2.51 | 1.48 | 0.85–2.60 | 1.43 | 0.84–2.43 |
| Overweight | 1.11 | 0.87–1.43 | 1.07 | 0.81–1.42 | 1.13 | 0.82–1.57 |
| Age (5 years) | 1.06 | 0.99–1.13 | 1.08 | 1.00–1.16 | 1.10 | 1.00–1.21 |
| Primary care | 1.10 | 0.87–1.38 | 1.11 | 0.87–1.42 | 1.04 | 0.81–1.34 |
| Nursing degree | 0.88 | 0.70–1.11 | 0.88 | 0.68–1.15 | 0.87 | 0.65–1.16 |
| Other degree | 1.14 | 0.78–1.66 | 1.08 | 0.73–1.60 | 1.08 | 0.71–1.63 |

The results obtained from the analysis of self-perceived life satisfaction are detailed in Table 7 and Table 8 and illustrated in Figure 2. As in the case of self-perceived health, the strongest negative association with life satisfaction was measured for prior health problems, and this relationship became significantly stronger for both male and female respondents as the number of pre-existing health problems increased. For men, furthermore, working in primary rather than other levels of healthcare was also associated with less life satisfaction. Another important factor was that of physical inactivity, which was also associated with lower levels of life satisfaction, especially among men, although the differences with women in this respect were not statistically significant. Thus, male and female HCPs who performed no physical activity at all were 5 and 2.5 times, respectively, more likely to have less satisfaction with life than their more physically active counterparts. With respect to tobacco consumption, women who smoked (whether every day or less frequently) were more likely to report lower levels of life satisfaction than those who had never smoked. Finally, HCPs who slept less than seven hours per night were around 1.5 and 1.8 times (for men and women, respectively) more likely to report low levels of life satisfaction than those who slept for longer, assuming all other variables remained constant.

Figure 2: Confidence intervals at 95% for the odds ratio for each explanatory variable on self-perceived life satisfaction after applying logistic regression to the propensity score adjustment. The following reference classes are assumed for the qualita-tive variables: no health problems, never smoked, seven or more hours of sleep per night, physical exercise several days a week, normal weight or underweight, working in specialised healthcare and holding a degree in medicine. The xaxis scale is logarithmic to facilitate interpretation of the data.



# 4 Discussion

The stress of addressing the COVID-19 pandemic is having significant ill effects on HCPs'mental and physical health [44]. In consequence, the analysis of relevant data compiled before the present crisis is of crucial assistance to efforts to maintain and/or improve HCPs'well-being and to facilitate the application of more effective supportive interventions targeting policies, institutions and individuals [45]. In this regard, attention to personal welfare and service quality is of the utmost importance [46].

Regarding the methodological aspects of this study, in the analysis of non-probability samples, any inference drawn must take into account the selection bias inherent in the sampling procedure, which in most internet surveys is equivalent to self-selection bias. Propensity score adjustment can be a useful means of overcoming the effects of this kind of bias, although additional calibration may be needed to remove the bias completely [47,48]. In our study, PSA alone produced no substantial changes in the estimates except for the effect of certain variables on the indicators of health and life satisfaction. From this, we conclude that either the original sample was sufficiently representative of the target population or the variables in question did not properly model the self-selection mechanism.

The outcomes from algorithms used to estimate prevalences, as an alternative to logistic regression, did not differ from those obtained by assigning weights to decision trees and 5-unit neural networks. In the first case, this was because the algorithm was unable to grow any branch for the tree, as it did not detect any variable enabling it to classify an individual, either in the self-selected or in the reference sample. In the second case, the feed-forward technique achieved convergence in the first iteration, and therefore no adjustment was needed (see Appendix A for further information). Either or both of these cases might reflect a lack of predictability in the covariates available for both samples. On the other hand, the Horvitz–Thompson weights, which were also obtained for each PSA performed, had to be discarded as they resulted in a higher variance of the estimators and produced unstable and misleading point estimates.

The study has several limitations that have to be pointed out. First of all, there were no available measures to assess whether the bias removal had been successful or not. It is reasonable to assume that adjustments to mitigate selection bias may have a significant effect; however, model misspecification in PSA can increase the bias of the estimates, although the logistic regression model that was used as the reference result showed a relative robustness to changes in the covariates or sample size [23]. Further studies could consider the use of estimators that ensure robustness against model misspecifications, such as the doubly robust estimator proposed in [49].

Moreover, the available covariates did not show a very different behaviour in the online sample in comparison with the full population. This can indicate that the online sample was fairly representative of the population but can also indicate that the available covariates failed to capture the differences between the sampled and the non-sampled population, which could reduce the potential of PSA to mitigate the selection bias.

It was also observed that PSA increased the variance of the estimators in comparison with the unadjusted case. As stated in Section 1, it is known that PSA can reduce the selection bias at the cost of increasing the variance because of the complexity added by the predictive models. However, the bias–variance trade-off is often positive, as the mean square error gets reduced after the application of PSA in certain situations, according to literature [11,14,15,23,24].

Our analysis shows that, although there were no significant differences between male and female HCPs regarding self-rated health and dissatisfaction with life, male personnel had significantly poorer lifestyle habits than their female counterparts, together with a higher prevalence of chronic disease, of disability and of health problems. A different tendency was observed in sleep, chronic disease and health problems when comparedwith the general population. Further research is needed in this area in order to justify interventions which encourage male HCPs to modify their lifestyle habits in order to prevent problems from spiralling through the burnout cascade stages of reduced activity, distress and despair [50].

In our survey, members of the general population reported significantly poorer health and less satisfaction with life than the HCPs consulted. Although female HCPs consumed alcohol at least once in a month in a significantly higher frequency than those in the general population, they were only half as likely to suffer chronic disease. A limitation of that result is that the quantity of consumed alcohol was not reported in the survey. Other studies have also found a lower prevalence of chronic diseases among physicians than in the general population, with similar percentages to ours, ranging from 13–44% [51,52]. Nevertheless, further detailed, up-to-date research is needed in this area.

Among HCPs, the prior existence of health problems was the factor most strongly associated with worsening self-perceived health and decreased life satisfaction, while obesity had an important negative impact on female practitioners' self-perceived health. Our study did not include work environment, workplace characteristics and other factors such as quality of management, professional development and colleague support/team spirit. Allof those factors have a stronger positive association with HCPs' satisfaction compared with personal and intrinsic factors [53].

## 5 Conclusions

For almost all of the explanatory variables, any misestimations caused by the non-probabilistic nature of the sampling process for the online survey were corrected by reweighting. There were some differences across the estimations provided by different adjustments and estimators, although several groups of algorithms for PSA with similar behaviours could be spotted according to the weights that they provided. Horvitz–Thompson estimates had larger estimated variances, and tree-based bagging algorithms provided more skewed weights, which contributed to an increase in the variance of the estimates. The point estimates finally considered were similar, meaning that they probably removed bias to the same extent, but some adjustments presented lower variances, which made them more desirable in terms of reducing estimation error.According to our analysis, male HCPs reported poorer lifestyle habits and health conditions than their female counterparts, although men and women had similar perceptions of health and life satisfaction. All HCPs self-reported much better health conditions and life satisfaction than the general population. The prevalence of chronic disease among female HCPs was half that of the prevalence measured among the general population but that of disability among all HCPs was almost twice that of the general population. Prior health problems, sleeping for less than seven hours per night, physical inactivity and smoking (by women) were all associated with the perception of poorer health, while obesity (among women), working as a nurse or in primary healthcare (among male HCPs) were associated with less satisfaction with life. Accurate knowledge of HCPs' self-perceived health, life satisfaction and associated factors is essential to enabling policy makers and healthcare managers to design and implement effective programmes to improve the attention paid to human resources. The study results we report can be used as a baseline for monitoring the health effects produced in HCPs by the COVID-19 pandemic and for assessing interventions to benefit the welfare of these professionals, whose current role makes them priority beneficiaries of such attention.

## Author Contributions

Conceptualisation, A.C.-L.; methodology, R.F.-G., M.d.M.R. and A.C.-L.; software, R.F.-G.; validation, R.F.-G., M.d.M.R. and A.C.-L.; formal analysis, R.F.-G.; investigation, A.C.-L.; resources, R.F.-G., M.d.M.R. and A.C.-L.; data curation, A.C.-L.; writing—original draft preparation, R.F.-G., M.d.M.R. and A.C.-L.; writing—review and editing, R.F.-G., M.d.M.R. and A.C.-L.; visualisation, R.F.-G.; supervision, M.d.M.R. and A.C.-L.; project administration, M.d.M.R. and A.C.-L.; funding acquisition, M.d.M.R. All authors have read and agreed to the published version of the manuscript.

## Funding

This research was funded by the Ministerio de Ciencia e Innovación, Spain, under project PID2019-106861RB-I00/AEI/10.13039/501100011033 and, in terms of the first author, a FPU grant from the Spanish Ministry of Science, Innovation and Universities, grant number FPU17/02177.

## Institutional Review Board Statement

Ethical review and approval were waived for this study, as it had an observational design with no personal data involved.

## Informed Consent Statement

Study participants voluntarily enrolled in the online course approved by the Andalusian Health Quality Agency of the Junta de Andalucía (March 2014). They cannot be identified, their responses were anonymous.

## Data Availability Statement

Data for the HCP sample and the Hajek and Horvitz–Thompson weights are available from the OSF home database in the link: https://osf.io/tj6a7/?view_only=7d820d6b242147cd87836100b0b4fa1a (accessed on 5 April 2021).

## Conflicts of Interest

The authors declare no conflict of interest.

## Appendix A

Descriptive statistics of weights obtained through PSA with Horvitz–Thompson weighting applying each predictive algorithm can be observed in Table A1.

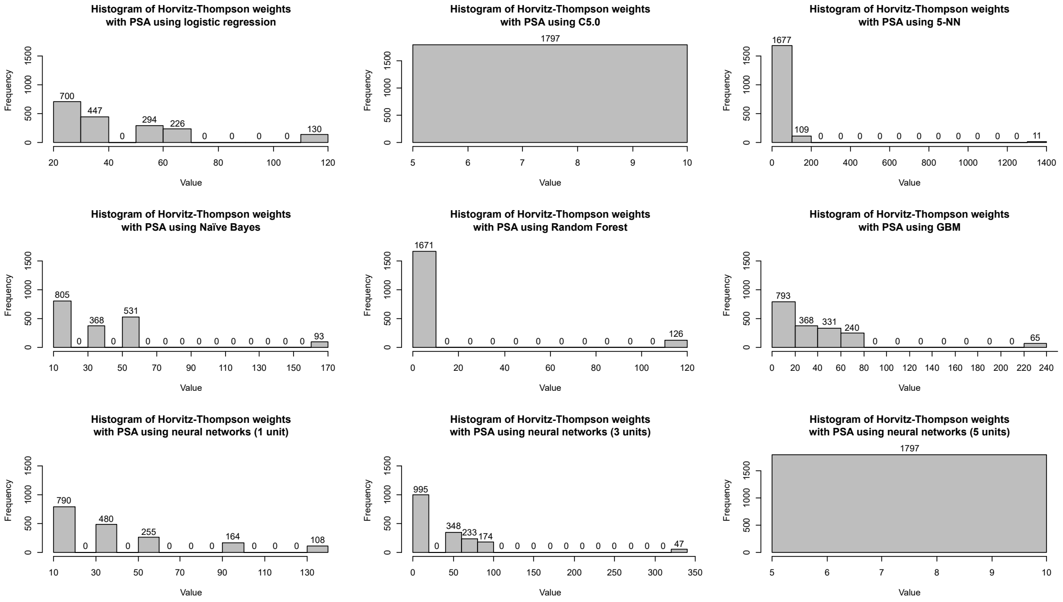
Table A1: Descriptive statistics for Horvitz–Thompson weights.

| | Logistic Regression | C5.0 | 5-NN | Naïve Bayes | Random Forest | GBM | Neural Net (1 Unit) | Neural Net (3 Units) | Neural Net (5 Units) |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 40.67 | 7.38 | 40.67 | 40.88 | 15.67 | 40.88 | 40.67 | 40.67 | 7.38 |
| Std. Dev. | 26.22 | 0 | 109.39 | 33.7 | 28.07 | 40.65 | 33.44 | 52.35 | 0 |
| CV | 0.64 | 0 | 2.69 | 0.82 | 1.79 | 0.99 | 0.82 | 1.29 | 0 |
| Minimum | 20.19 | 7.38 | 13.02 | 17.52 | 7.96 | 17.93 | 17.87 | 14 | 7.38 |
| Q1 | 20.19 | 7.38 | 13.02 | 17.52 | 7.96 | 17.93 | 17.87 | 14 | 7.38 |
| Median | 31.62 | 7.38 | 13.02 | 38.39 | 7.96 | 33.24 | 30.24 | 14 | 7.38 |
| Q3 | 50.85 | 7.38 | 36.8 | 54.9 | 7.96 | 49 | 56.25 | 64.61 | 7.38 |
| Maximum | 115.89 | 7.38 | 1373.45 | 166.92 | 117.85 | 231.42 | 139.74 | 323.55 | 7.38 |
| MAD | 16.94 | 0 | 0 | 27.75 | 0 | 22.7 | 18.33 | 0 | 0 |
| IQR | 30.66 | 0 | 23.78 | 37.38 | 0 | 31.07 | 38.38 | 50.62 | 0 |
| Skewness | 1.63 | NaN | 11.1 | 2.58 | 3.36 | 3.67 | 1.75 | 4.12 | NaN |
| Kurtosis | 2.15 | NaN | 131.95 | 7.24 | 9.32 | 14.52 | 2.22 | 19.39 | NaN |

It can be noticed that weights obtained using C5.0 and neural networks with 5 units in the hidden layer for propensity estimation provide constant weights as a result, equivalent to not doing any adjustment at all and using design weights. The rest of the weights move around the same values given the similarity of means (except for weights using random forest in PSA), but the variability is not the same for all of them. More precisely, variability of weights after using logistic regression is relatively smaller, as well as after the use of naïve Bayes, neural networks with 1 unit in the hidden layer or gradient boosting machines. Variability begins to be relatively high when 3 units are placed in the hidden layer in neural networks and very high when using random forest and 5-NN. In these last two cases, very significant outliers are present. All of the weightings present a high skewness, along with a high kurtosis in a majority of the cases.

Histograms and boxplots for each weighting can be observed in Figure A1 and Figure A2, where some of the patterns detected in the descriptive statistics are notorious. Positive skew is present in all weights, but although some of them are more uniform (such as weights using logistic regression in PSA), positive skew is more pronounced in others and even attributable exclusively to outliers. For example, when using GBM in PSA, most of the weights are below 80, except for only 65 of those weights (3.6% of the individuals) which take values over 220. However, the most notorious cases are those provided by random forest and 5-NN. In the case of random forest, all of the individuals have a weight of 7.96, except for 126 individuals (around 7% of the sample) that take a value of 117.85, much higher than the rest, leading to an increase of the skewness and the variability. On the other hand, weighting using 5-NN in PSA provides weights under 200 (with most of them being under 36.8, as described in Table A1), while a small subset of 11 individuals (0.6% of the sample) has a weight of almost 1400. This disposition largely increases variability, as well as skewness.

Figure A1: Histograms of Horvitz–Thompson weights.



Histogram of Horvitz-Thompson weights with PSA using logistic regression

Histogram of Horvitz-Thompson weights with PSA using C5.0

Histogram of Horvitz-Thompson weights with PSA using 5-NN

Histogram of Horvitz-Thompson weights with PSA using Naïve Bayes

Histogram of Horvitz-Thompson weights with PSA using Random Forest

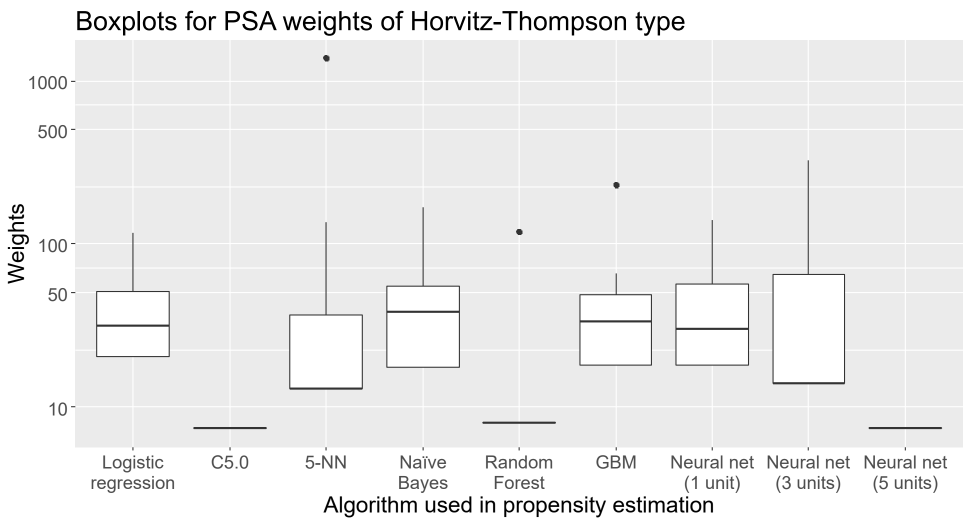
Histogram of Horvitz-Thompson weights with PSA using GBM

Histogram of Horvitz-Thompson weights with PSA using neural networks (1 unit)

Histogram of Horvitz-Thompson weights with PSA using neural networks (3 units)

Histogram of Horvitz-Thompson weights with PSA using neural networks (5 units)

Figure A2: Boxplots of Horvitz–Thompson weights.



Descriptive statistics of weights obtained through PSA with Hajek weighting applied to each predictive algorithm can be observed in Table A2.

Table A2: Descriptive statistics for Hajek weights.

| | Logistic Regression | C5.0 | 5-NN | Naïve Bayes | Random Forest | GBM | Neural Net (1 Unit) | Neural Net (3 Units) | Neural Net (5 Units) |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.00056 |
| Std. Dev. | 0.00030 | 0 | 0.00063 | 0.00040 | 0.00015 | 0.00031 | 0.00052 | 0.00063 | 0 |
| CV | 0.53 | 0 | 1.12 | 0.71 | 0.27 | 0.56 | 0.93 | 1.14 | 0 |
| Minimum | 0.00022 | 0.00056 | 0.00001 | 0.00019 | 0.0000002 | 0.00011 | 0.00025 | 0.0000067 | 0.00056 |
| Q1 | 0.00032 | 0.00056 | 0.00019 | 0.00027 | 0.0005985 | 0.00032 | 0.00027 | 0.000217 | 0.00056 |
| Median | 0.00049 | 0.00056 | 0.00033 | 0.00046 | 0.0005985 | 0.00050 | 0.00028 | 0.0003098 | 0.00056 |
| Q3 | 0.00071 | 0.00056 | 0.00071 | 0.00064 | 0.0005985 | 0.00068 | 0.00069 | 0.0008019 | 0.00056 |
| Maximum | 0.00202 | 0.00056 | 0.00710 | 0.00340 | 0.0005985 | 0.00365 | 0.00392 | 0.0048296 | 0.00056 |
| MAD | 0.00028 | 0 | 0.00027 | 0.00028 | 0 | 0.00027 | 0.00004 | 0.000314 | 0 |
| IQR | 0.00039 | 0 | 0.00052 | 0.00037 | 0 | 0.00036 | 0.00042 | 0.0005849 | 0 |
| Skewness | 1.42 | NaN | 3.28 | 2.28 | -3.26 | 2.57 | 3.51 | 4.28 | NaN |
| Kurtosis | 2.77 | NaN | 16.88 | 8.29 | 8.70 | 14.08 | 16.91 | 25.10 | NaN |

Weights obtained for Hajek estimators are more stable than those obtained for Horvitz–Thompson ones. In each weighting, values are around the same numbers (mean is identical in all cases), and the coefficient of variation is, in all cases, relatively low and below its counterpart for Horvitz–Thompson weights. Skewness coefficients again show that weights tend to be right-skewed, except for weighting with PSA using random forest, which provides very left-skewed values. Kurtosis coefficients are high as well, showing leptokurtic distributions.

Figure A3 and Figure A4 show histograms and boxplots for Hajek weights obtained with each algorithm in PSA. In this case, skewness appears in a smoother manner as propensities were not grouped in strata as was done with Horvitz–Thompson weights. This allows weights to be closer to the arithmetic mean, which results in the decrease in variability previously mentioned. The use of 5-NN or random forest provides the most unstable situations because of the presence of outliers.

Figure A3: Histograms of Hajek weights.

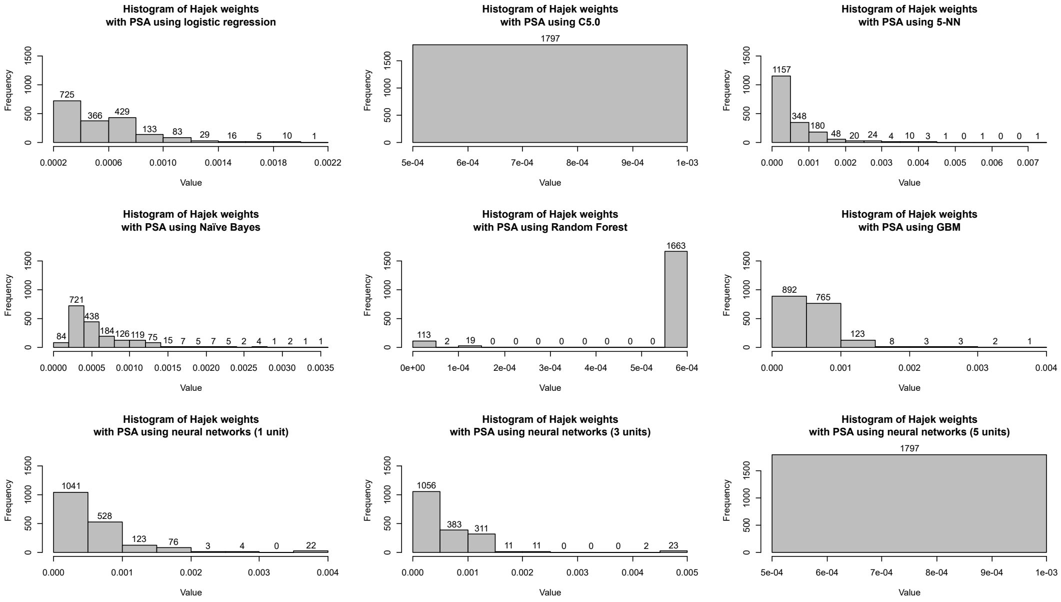
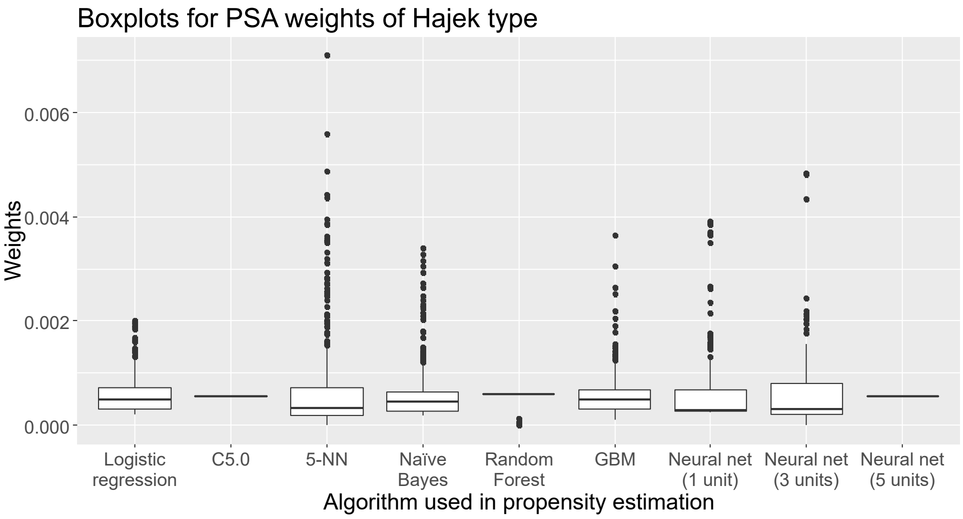


Figure A4: Boxplots of Hajek weights.



Following one-dimensional analysis, Pearson bivariate correlations between weights were analysed. Results of correlations can be observed in Figure A5 and Figure A6.

Figure A5: Representation of Pearson correlations between weights. The darker and larger the circle, the closer the correlation is to 1 (in caseswith a blue circle) or -1 (in caseswith a red circle).

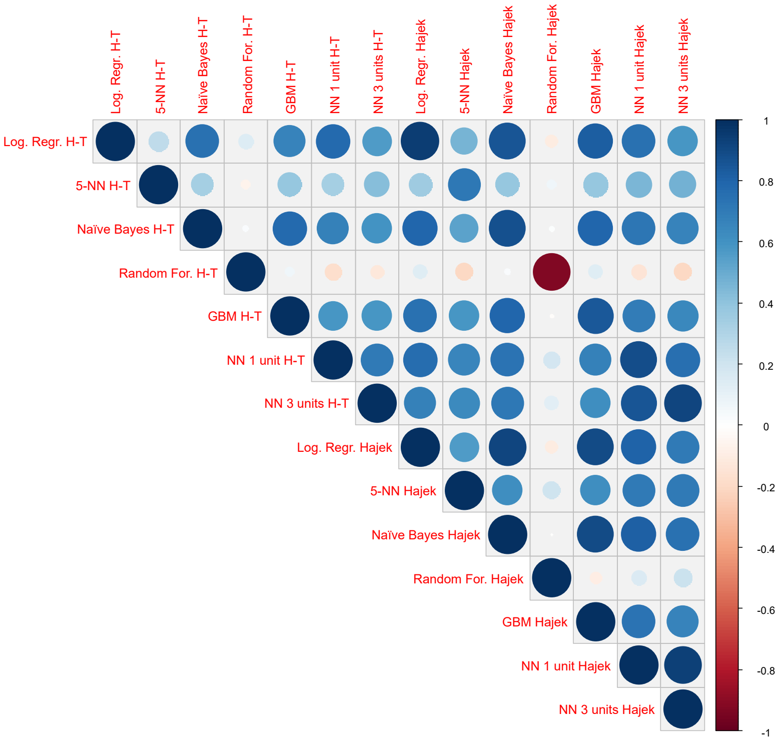
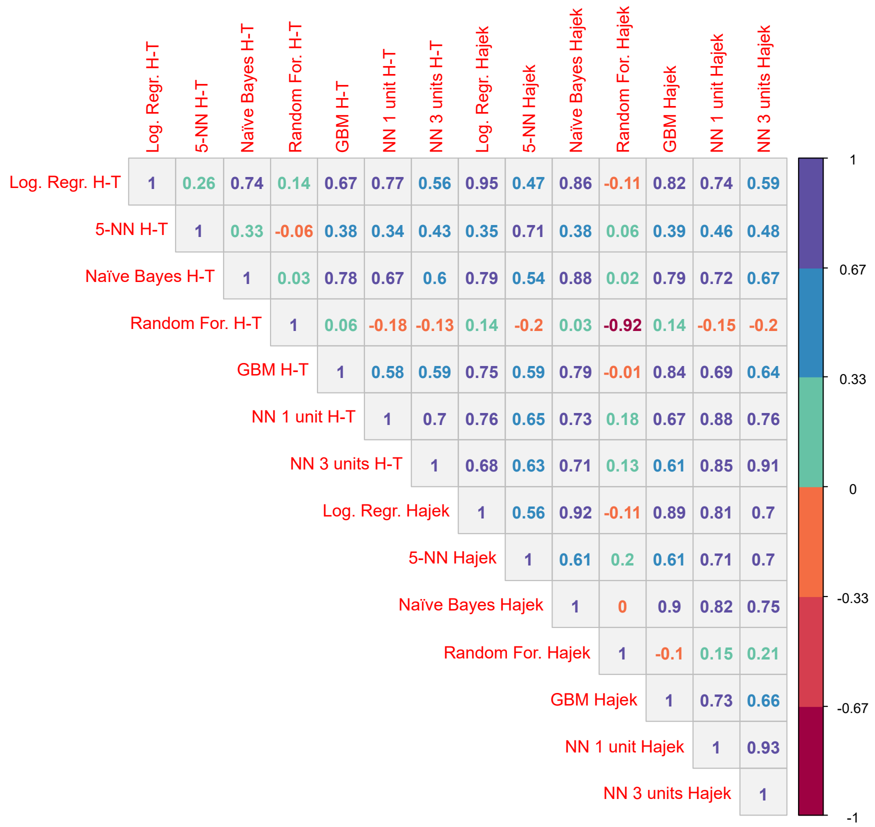


Figure A6: Pearson's bivariate correlations between weights.

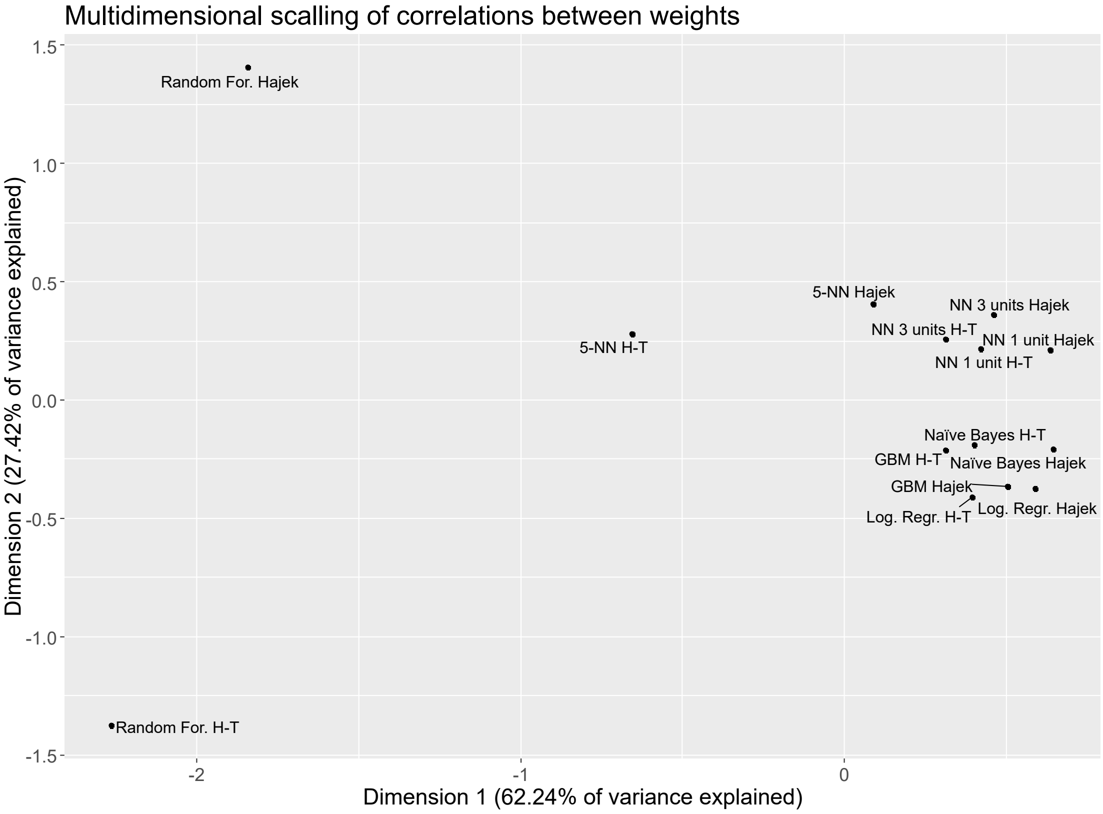


|  | Log. Regr. H-T | 5-NN H-T | Naïve Bayes H-T | Random For. H-T | GBM H-T | NN 1 unit H-T | NN 3 units H-T | Log. Regr. Hajek | 5-NN Hajek | Naïve Bayes Hajek | Random For. Hajek | GBM Hajek | NN 1 unit Hajek | NN 3 units Hajek |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Log. Regr. H-T | 1 | 0.26 | 0.74 | 0.14 | 0.67 | 0.77 | 0.56 | 0.95 | 0.47 | 0.86 | -0.11 | 0.82 | 0.74 | 0.59 |
| 5-NN H-T |  | 1 | 0.33 | -0.06 | 0.38 | 0.34 | 0.43 | 0.35 | 0.71 | 0.38 | 0.06 | 0.39 | 0.46 | 0.48 |
| Naïve Bayes H-T |  |  | 1 | 0.03 | 0.78 | 0.67 | 0.6 | 0.79 | 0.54 | 0.88 | 0.02 | 0.79 | 0.72 | 0.67 |
| Random For. H-T |  |  |  | 1 | 0.06 | -0.18 | -0.13 | 0.14 | -0.2 | 0.03 | -0.92 | 0.14 | -0.15 | -0.2 |
| GBM H-T |  |  |  |  | 1 | 0.58 | 0.59 | 0.75 | 0.59 | 0.79 | -0.01 | 0.84 | 0.69 | 0.64 |
| NN 1 unit H-T |  |  |  |  |  | 1 | 0.7 | 0.76 | 0.65 | 0.73 | 0.18 | 0.67 | 0.88 | 0.76 |
| NN 3 units H-T |  |  |  |  |  |  | 1 | 0.68 | 0.63 | 0.71 | 0.13 | 0.61 | 0.85 | 0.91 |
| Log. Regr. Hajek |  |  |  |  |  |  |  | 1 | 0.56 | 0.92 | -0.11 | 0.89 | 0.81 | 0.7 |
| 5-NN Hajek |  |  |  |  |  |  |  |  | 1 | 0.61 | 0.2 | 0.61 | 0.71 | 0.7 |
| Naïve Bayes Hajek |  |  |  |  |  |  |  |  |  | 1 | 0 | 0.9 | 0.82 | 0.75 |
| Random For. Hajek |  |  |  |  |  |  |  |  |  |  | 1 | -0.1 | 0.15 | 0.21 |
| GBM Hajek |  |  |  |  |  |  |  |  |  |  |  | 1 | 0.73 | 0.66 |
| NN 1 unit Hajek |  |  |  |  |  |  |  |  |  |  |  |  | 1 | 0.93 |
| NN 3 units Hajek |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |

It is noticeable how correlations are generally positive and relatively high except for two cases: Horvitz–Thompson weighting using 5-NN in PSA and using random forest. In the former case, correlations with the rest of weights are positive but weaker than the rest of the cases (it only shows a slightly stronger relationship when the same algorithm is used but weights are developed for Hajek estimator instead). The random forest case is more remarkable: correlations with any other set of weights are very low, except with Hajek weights using the same algorithm where the correlation is highly negative. It is likely that this lack of correspondence is caused by the propensities estimated by the random forest algorithm, which assigns probabilities very close to the limits 0 and 1, and therefore correlation depends almost exclusively on the few individuals that have been assigned probabilities far from those limits.

In order to better visualise the existent relationships between weights, the correlation matrix was used as an input for multidimensional scaling (MDS) in two dimensions, which explains 89.65% of the total variance. Results of the analysis can be observed in Figure A7.

Figure A7: Multidimensional scaling for two dimensions of the correlations between weights.



Thanks to the scaling, the existence of two differentiated groups can be noted: the group composed of weights obtained using PSA with logistic regression, GBM and naïve Bayes and another group composed of those obtained with neural networks and 5-NN (for Hajek estimators). For 5-NN, if Horvitz–Thompson weighting is used, weights separate from the groups previously mentioned but are closer to the second group than to the first one. Weights obtained with PSA using random forest are very separated from the rest of the weights, no matter which estimator weights were developed for.

# Appendix B

Table A3: Point estimate, variance and difference from the non-adjusted case of estimators of prevalence in male HCPs for each propensity score adjustment (PSA) (algorithms are sorted from the least to the most complex).

| Algorithm Used in PSA | Poor Self-Perceived Health | | | | Dissatisfied with Life (Score of 6 or Less) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.088 | 0.00014 | Estimate | Variance | 0.1002 | 0.00016 | Estimate | Variance |
| Logistic regression | 0.084 | 0.00016 | -4.34% | 17% | 0.1031 | 0.00023 | 2.93% | 45% |
| 5-NN | 0.086 | 0.00029 | -2.29% | 103% | 0.1019 | 0.00041 | 1.68% | 159% |
| Naïve Bayes | 0.081 | 0.00017 | -8.24% | 21% | 0.1049 | 0.00031 | 4.67% | 98% |
| Random Forest | 0.087 | 0.00015 | -1.12% | 8% | 0.1026 | 0.00018 | 2.38% | 11% |
| GBM | 0.082 | 0.00016 | -6.12% | 11% | 0.0965 | 0.00020 | -3.68% | 28% |
| Neural net (1 unit) | 0.087 | 0.00023 | -0.58% | 62% | 0.1090 | 0.00043 | 8.84% | 174% |
| Neural net (3 units) | 0.086 | 0.00025 | -1.77% | 76% | 0.1190 | 0.00061 | 18.75% | 285% |
| Algorithm Used in PSA | Alcohol once a week | | | | <7 h of sleep | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.6232 | 0.00041 | Estimate | Variance | 0.3093 | 0.00038 | Estimate | Variance |
| Logistic regression | 0.6234 | 0.00053 | 0.02% | 29% | 0.3118 | 0.00049 | 0.82% | 30% |
| 5-NN | 0.5940 | 0.00095 | -4.69% | 129% | 0.3252 | 0.00083 | 5.12% | 121% |
| Naïve Bayes | 0.6240 | 0.00066 | 0.12% | 59% | 0.3055 | 0.00059 | -1.25% | 56% |
| Random Forest | 0.6145 | 0.00046 | -1.40% | 10% | 0.3136 | 0.00041 | 1.40% | 10% |
| GBM | 0.6107 | 0.00058 | -2.02% | 40% | 0.3034 | 0.00048 | -1.91% | 27% |
| Neural net (1 unit) | 0.6004 | 0.00085 | -3.66% | 106% | 0.3395 | 0.00085 | 9.76% | 126% |
| Neural net (3 units) | 0.5942 | 0.00109 | -4.67% | 163% | 0.3609 | 0.00114 | 16.69% | 204% |
| Algorithm Used in PSA | Disability (physical. mental or sensorial) | | | | Chronic disease | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.0645 | 0.00011 | Estimate | Variance | 0.3369 | 0.00040 | Estimate | Variance |
| Logistic regression | 0.0695 | 0.00016 | 7.74% | 46% | 0.3179 | 0.00048 | -5.63% | 19% |
| 5-NN | 0.0587 | 0.00020 | -8.96% | 81% | 0.3280 | 0.00082 | -2.66% | 104% |
| Naïve Bayes | 0.0688 | 0.00017 | 6.64% | 54% | 0.3065 | 0.00055 | -9.03% | 37% |
| Random Forest | 0.0574 | 0.00011 | -10.98% | -3% | 0.3412 | 0.00044 | 1.27% | 10% |
| GBM | 0.0707 | 0.00016 | 9.51% | 45% | 0.3211 | 0.00050 | -4.70% | 26% |
| Neural net (1 unit) | 0.0584 | 0.00015 | -9.42% | 43% | 0.3065 | 0.00065 | -9.03% | 63% |
| Neural net (3 units) | 0.0506 | 0.00013 | -21.56% | 16% | 0.2974 | 0.00077 | -11.73% | 91% |
| Algorithm Used in PSA | One health problem | | | | Two or more health problems | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.2742 | 0.00036 | Estimate | Variance | 0.1072 | 0.00017 | Estimate | Variance |
| Logistic regression | 0.2630 | 0.00044 | -4.09% | 22% | 0.1037 | 0.00019 | -3.23% | 13% |
| 5-NN | 0.2487 | 0.00067 | -9.30% | 89% | 0.1158 | 0.00038 | 8.06% | 128% |
| Naïve Bayes | 0.2527 | 0.00048 | -7.83% | 35% | 0.1003 | 0.00020 | -6.43% | 21% |
| Random Forest | 0.2684 | 0.00038 | -2.12% | 8% | 0.1084 | 0.00019 | 1.12% | 10% |
| GBM | 0.2634 | 0.00045 | -3.95% | 26% | 0.1059 | 0.00020 | -1.23% | 20% |
| Neural net (1 unit) | 0.2361 | 0.00054 | -13.90% | 51% | 0.1048 | 0.00024 | -2.22% | 41% |
| Neural net (3 units) | 0.2235 | 0.00062 | -18.49% | 74% | 0.1044 | 0.00025 | -2.58% | 51% |

Table A4: Point estimate, variance and difference from the non-adjusted case of estimators of prevalence in female HCPs for each propensity score adjustment (PSA) (algorithms are sorted from the least to the most complex).

| Algorithm Used in PSA | Poor Self-Perceived Health | | | | Dissatisfied with Life (Score of 6 or Less) | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.088 | 0.00014 | Estimate | Variance | 0.1002 | 0.00016 | Estimate | Variance |
| Logistic regression | 0.084 | 0.00016 | -4.34% | 17% | 0.1031 | 0.00023 | 2.93% | 45% |
| 5-NN | 0.0597 | 0.00006 | -28.78% | -3% | 0.1234 | 0.00022 | 2.41% | 158% |
| Naïve Bayes | 0.0774 | 0.00009 | -7.71% | 41% | 0.1270 | 0.00015 | 5.34% | 68% |
| Random Forest | 0.0833 | 0.00007 | -0.74% | 7% | 0.1183 | 0.00009 | -1.82% | 6% |
| GBM | 0.0753 | 0.00006 | -10.22% | 3% | 0.1261 | 0.00013 | 4.62% | 45% |
| Neural net (1 unit) | 0.087 | 0.00023 | -0.58% | 62% | 0.1090 | 0.00043 | 8.84% | 174% |
| Neural net (3 units) | 0.0638 | 0.00007 | -23.90% | 6% | 0.1292 | 0.00025 | 7.16% | 187% |
| Algorithm Used in PSA | Alcohol once a week | | | | <7 h of sleep | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.6232 | 0.00041 | Estimate | Variance | 0.3093 | 0.00038 | Estimate | Variance |
| Logistic regression | 0.6234 | 0.00053 | 0.02% | 29% | 0.3118 | 0.00049 | 0.82% | 30% |
| 5-NN | 0.4451 | 0.00048 | 5.42% | 139% | 0.2574 | 0.00038 | -3.65% | 138% |
| Naïve Bayes | 0.4277 | 0.00031 | 1.28% | 56% | 0.2607 | 0.00023 | -2.40% | 43% |
| Random Forest | 0.4239 | 0.00021 | 0.38% | 8% | 0.2671 | 0.00017 | 0.02% | 8% |
| GBM | 0.4251 | 0.00026 | 0.67% | 33% | 0.2599 | 0.00020 | -2.70% | 23% |
| Neural net (1 unit) | 0.6004 | 0.00085 | -3.66% | 106% | 0.3395 | 0.00085 | 9.76% | 126% |
| Neural net (3 units) | 0.4227 | 0.00049 | 0.12% | 144% | 0.2503 | 0.00034 | -6.27% | 113% |
| Algorithm Used in PSA | Disability (physical. mental or sensorial) | | | | Chronic disease | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.0645 | 0.00011 | Estimate | Variance | 0.3369 | 0.00040 | Estimate | Variance |
| Logistic regression | 0.0695 | 0.00016 | 7.74% | 46% | 0.3179 | 0.00048 | -5.63% | 19% |
| 5-NN | 0.0583 | 0.00010 | -7.10% | 114% | 0.2353 | 0.00036 | 5.55% | 151% |
| Naïve Bayes | 0.0605 | 0.00008 | -3.67% | 64% | 0.2224 | 0.00022 | -0.27% | 54% |
| Random Forest | 0.0612 | 0.00005 | -2.44% | 5% | 0.2219 | 0.00015 | -0.48% | 7% |
| GBM | 0.0618 | 0.00008 | -1.51% | 56% | 0.2241 | 0.00020 | 0.52% | 37% |
| Neural net (1 unit) | 0.0584 | 0.00015 | -9.42% | 43% | 0.3065 | 0.00065 | -9.03% | 63% |
| Neural net (3 units) | 0.0627 | 0.00014 | -0.06% | 183% | 0.2273 | 0.00038 | 1.94% | 162% |
| Algorithm Used in PSA | One health problem | | | | Two or more health problems | | | |
| | Estimate | Variance | Diff. from no adj. (%) | | Estimate | Variance | Diff. from no adj. (%) | |
| No adjustment | 0.2742 | 0.00036 | Estimate | Variance | 0.1072 | 0.00017 | Estimate | Variance |
| Logistic regression | 0.2630 | 0.00044 | -4.09% | 22% | 0.1037 | 0.00019 | -3.23% | 13% |
| 5-NN | 0.2164 | 0.00032 | 1.95% | 133% | 0.0566 | 0.00011 | 0.65% | 150% |
| Naïve Bayes | 0.2013 | 0.00019 | -5.16% | 39% | 0.0616 | 0.00008 | 9.63% | 96% |
| Random Forest | 0.2136 | 0.00015 | 0.66% | 8% | 0.0542 | 0.00004 | -3.58% | 3% |
| GBM | 0.2047 | 0.00017 | -3.56% | 21% | 0.0607 | 0.00008 | 8.08% | 81% |
| Neural net (1 unit) | 0.2361 | 0.00054 | -13.90% | 51% | 0.1048 | 0.00024 | -2.22% | 41% |
| Neural net (3 units) | 0.2152 | 0.00034 | 1.41% | 150% | 0.0575 | 0.00013 | 2.26% | 205% |

Figure A8: The 95% confidence intervals for the prevalence of variables related to self-perceived health and lifestyle satisfaction among male HCPs, according to the algorithms used in the propensity score adjustment (facets are sorted by confidence interval values in order to obtain common yaxis limits in each row).

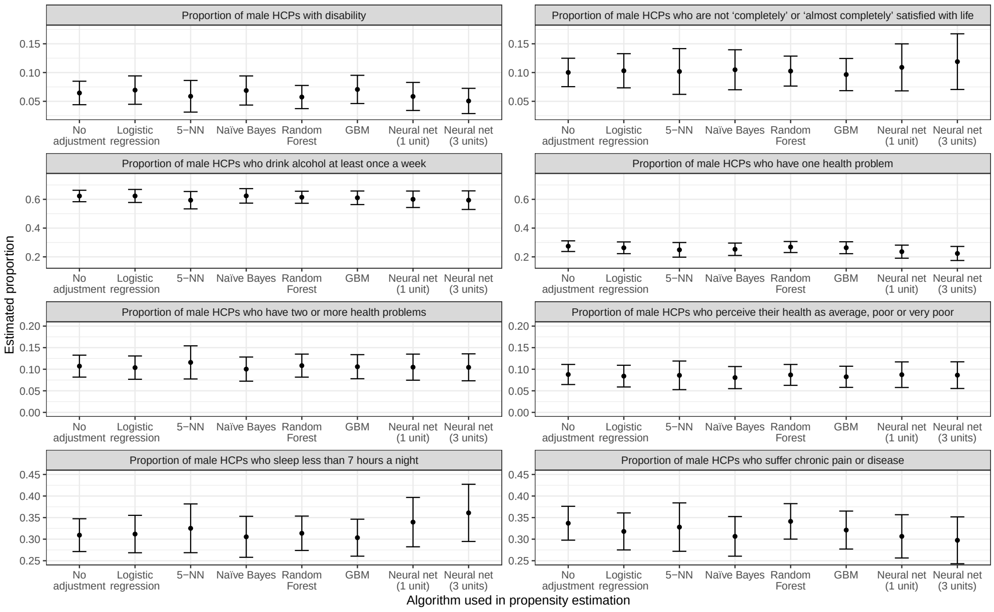
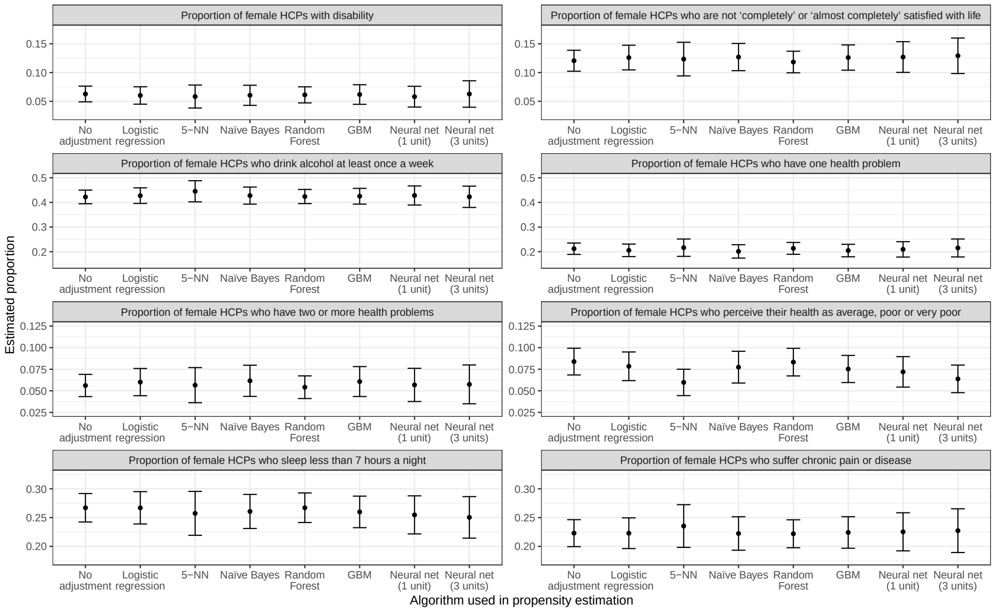


Figure A9: The 95% confidence intervals for the prevalence of variables related to self-perceived health and lifestyle satisfaction among female HCPs, according to the algorithms used in the propensity score adjustment (facets are sorted by confidence interval values in order to obtain common yaxis limits in each row).



# References

1. Parsa-Parsi, R.W. The revised Declaration of Geneva: A modern day physician's pledge. J. Am. Med. Assoc. 2017, 318, 1971–1972.

2. Hall, L.H.; Johnson, J.; Watt, I.; Tsipa, A.; O'Connor, D.B. Healthcare staff wellbeing, burnout, and patient safety: A systematic review. PLoS ONE 2016, 11, 1–12.

3. Jadad, A.R.; Jadad Garcia, T.M. From a digital bottle: A message to ourselves in 2039 2. J. Med. Internet Res. 2019, 21, e16274.

4. Albuquerque, J.; Tulk, S. Physician suicide. Can. Med. Assoc. J. 2019, 191, E505.

5. Mumba, M.; Kraemer, K. Substance use disorders among nurses in medical-surgical, long-term care, and outpatient services. Medsurg. Nurs. 2019, 28, 118.

6. Angie, C.C.; Leung, T. Substance use disorders. In The Art and Science of Physician Wellbeing: A Handbook for Physicians and Trainees; Weiss Roberts, L., Trockel, M., Eds.; Springer International Publishing: Berlin, Germany, 2019.

7. World Medical Association. WMA Statement on Physicians Well-Being. Adopted by the 66th WMA General Assembly, Moscow, 2015. Available online: https://www.wma.net/policies-post/wma-statement-on-physicians-well-be (accessed on 27 November 2019).

8. Howell, R.T.; Rodzon, K.S.; Kurai, M.; Sánchez, A.M. A validation of well-being and happiness surveys for administration via the Internet. Behav. Res. Methods 2010, 42, 775.

9. Ekman, A.; Klint, A.; Dickman, P.W.; Adami, H.; Litton, J. Optimizing the design of web-based questionnaires—Experience from a population-based study among 50,000 women. Eur. J. Epidemiol. 2007, 22, 293.

10. Beaumont, J.F.; Rao, J.N.K. Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? Surv. Stat. 2021, 83, 11–22.

11. Castro-Martín, L.; Rueda, M.; Ferri-García, R. Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys. J. Comput. Appl. Math. 2021, 113414.

12. Erens, B.; Burkill, S.; Couper, M.P.; Conrad, F.; Clifton, S.; Tanton, C.; Phelps, A.; Datta, J.; Mercer, C.H.; Sonnenberg, P.; et al. Nonprobability Web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. J. Med. Internet Res. 2014, 16, e276, PMCID:PMC4275497.

13. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. Biometrika 1983, 70, 41–55.

14. Lee, S. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J. Off. Stat. 2006, 22, 329–349.

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

16. 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.

17. Flouris, A.D.; Duffy, J. Applications of artificial intelligence systems in the analysis of epidemiological data. Eur. J. Epidemiol. 2006, 21, 167–170.

18. Keil, A.P.; Edwards, J.K. You are smarter than you think: (super) machine learning in context. Eur. J. Epidemiol. 2018, 33, 437–440.

19. Naimi, A.I.; Balzer, L.B. Stacked generalization: An introduction to super learning. Eur. J. Epidemiol. 2018, 33, 459–464.

20. Bentley, R.; Baker, E.; Simons, K.; Simpson, J.A.; Blakely, T. The impact of social housing on mental health: Longitudinal analyses using marginal structural models and machine learning-generated weights. Int. J. Epidemiol. 2018, 1414–1422.

21. Mayr, A.; Weinhold, L.; Hofner, B.; Titze, S.; Gefeller, O.; Schmid, M. The betaboost package—A software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data. Int. J. Epidemiol. 2018, 1383–1388.

22. Torres, J.M.; Rudolph, K.E.; Sofrygin, O.; Glymour, M.M.; Wong, R. Longitudinal associations between having an adult child migrant and depressive symptoms among older adults in the Mexican Health and Aging Study. Int. J. Epidemiol. 2018, 1432–1442.

23. 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.

24. 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.

25. WHO. Body Mass Classification; World Health Organization: Geneva, Switzerland, 2015; Available online: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed on 1 April 2021).

26. Cabrera-León, A.; Cantero-Braojos, M.; Garcia-Fernandez, L.; de Hoyos Guerra, J.A. Living with disabling chronic pain: Results from a face-to-face cross-sectional population-based study. BMJ Open 2018, 8, e020913.

27. Schonlau, M.; Couper, M. Options for conducting web surveys. Stat. Sci. 2017, 32, 279–292.

28. Quinlan, R. C4.5: Programs for Machine Learning; Morgan Kaufmann: San Francisco, CA, USA, 1993.

29. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2018; Available online: https://www.R-project.org/ (accessed on 1 April 2021).

30. Tillé, Y.; Matei, A. Sampling: Survey Sampling; R Package Version 2.7; R Foundation for Statistical Computing: Vienna, Austria, 2015; Available online: http://CRAN.R-project.org/package=sampling (accessed on 1 April 2021).

31. Lumley, T. Survey: Analysis of Complex Survey Samples, R package version 3.30; R Foundation for Statistical Computing: Vienna, Austria, 2014.

32. Kuhn, M.; Quinlan, R. C50: C5.0 Decision Trees and Rule-Based Models. R Foundation for Statistical Computing: Vienna, Austria, 2018. Available online: https://CRAN.R-project.org/package=C50 (accessed on 1 April 2021).

33. Liaw, A.; Wiener, M. Classification and regression by random forest. R News 2002, 2, 18–22.

34. Greenwell, B.; Boehmke, B.; Cunningham, J.; Developers, G. Package 'gbm'. R Foundation for Statistical Computing: Vienna, Austria, 2018. Available online: https://cran.r-project.org/web/packages/gbm/gbm.pdf (accessed on 1 April 2021).

35. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. e1071: Misc Functions of the Department of Statistics (e1071); R Foundation for Statistical Computing: Vienna, Austria, 2018; Available online: https://cran.rproject.org/web/packages/e1071 (accessed on 1 April 2021).

36. Kuhn, M. Caret: Classification and Regression Training. Available online: https://cran.rproject.org/web/packages/caret/index.html (accessed on 1 April 2021).

37. Venables, W.N.; Ripley, B.D. Modern Applied Statistics with S, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.

38. Quenouille, M.H. Notes on bias in estimation. Biometrika 1956, 43, 353–360.

39. Tibshirani, R.; Leisch, F. Bootstrap: Functions for the Book "An Introduction to the Bootstrap". 2017. Available online: https://cran.r-project.org/web/packages/bootstrap/index.html (accessed on 1 April 2021).

40. Stine, R.A. Graphical interpretation of variance inflation factors. Am. Stat. 1995, 49, 53–56.

41. Hair, J.F.; Black, W.C.; Babin, B.; Anderson, R.E. Multivariate Data Analysis, 7th ed.; Prentice Hall: Hoboken, NJ, USA, 2010.

42. Edwards, B.; Pollock, P. Poliscidata: Datasets and Functions Featured in Pollock and Edwards, An R Companion to Essentials of Political Analysis. 2018. Available online: https://CRAN.R-project.org/package=poliscidata (accessed on 1 April 2021).

43. Wickham, H. ggplot2—Elegant Graphics for Data Analysis, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2016; Available online: https://github.com/hadley/ggplot2-book (accessed on 1 April 2021).

44. Lim, R.; Aarsen, K.; Van Aarsen, K.; Gray, S.; Rang, L.; Fitzpatrick, J.; Fischer, L. Emergency medicine physician burnout and wellness in Canada before COVID19: A national survey. Can. J. Emerg. Med. 2020, 22, 603–607.

45. Khan, A.; Vinson, A.E. Physician well-being in practice. Anesth. Analg. 2020, 131, 1359–1369.

46. López-Cabarcos, M.Á.; López-Carballeira, A.; Ferro-Soto, C. New ways of working and public healthcare professionals' well-being: The response to face the covid-19 pandemic. Sustainability 2020, 12, 8087.

47. Valliant, R.; Dever, J.A. Estimating propensity adjustments for volunteer web surveys. Sociol. Method Res. 2011, 40, 105–137.

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

49. Chen, Y.; Li, P.; Wu, C. Doubly robust inference with nonprobability survey samples. J. Am. Stat. Assoc. 2020, 115, 2011–2021.

50. Weber, A.; Jaekel-Reinhard, A. Burnout syndrome: A disease of modern societies? Occup. Med. 2000, 50, 512–517.

51. Campbell, S.; Delva, D. Physician do not heal thyself. Survey of personal health practices among medical residents. Can. Fam. Phys. 2003, 49, 1121–1127.

52. Robert Koch-Institut. Gesundheitstrends Bei Erwachseneninin Deutschland Zwischen 2003 und 2012. In DatenundFakten: Ergebnisse der Studie 'Gesundheit in Deutschland Aktuell 2012'; Beiträge zur Gesundheitsberichterstattung des Bundes; Robert Koch-Institut, Ed.; RKI: Berlin, Germany, 2014; pp. 13–33.

53. Domagała, A.; Bała, M.M.; Storman, D.; Peña-Sánchez, J.N.; Świerz, M.J.; Kaczmarczyk, M.; Storman, M. Factors associated with satisfaction of hospital physicians: A systematic review on european data. Int. J. Environ. Res. Public Health 2018, 15, 2546.