1.

Combining Statistical Matching and Propensity Score Adjustment for Inference from Non-Probability Surveys

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

[a]*Department of Statistics and Operations Research, University of Granada*

Abstract

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

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

*Corresponding author

Email addresses: luiscastro193@ugr.es (Luis Castro-Martín), mrueda@ugr.es (María del Mar Rueda), rferri@ugr.es (Ramón Ferri-García)

1. Introduction

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

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

Inferences from Big Data and nonprobability surveys have important sources of error. Given the characteristics of these data collection procedures, selection bias is particularly relevant. Following notation from [**?**], in a situation where U is the target population to which survey results are supposed to be generalized, a nonprobability selection ensures that sample individuals will be drawn from a population of potentially covered individuals, $U_{pc} \subset U$. This is the case of internet and smartphone surveys, where the population with the necessary devices for taking part in the survey are a subset of the total population. The bias produced by this issue is commonly known as coverage error. In addition, if the participation in the survey is conditioned to a selection mechnism, the sample will be eventually drawn from an actually covered population, $U_{ac} \subset U_{pc}$. Following the previous example, internet surveys with an opt-in scheme (such as

snowball samples in social media websites) would recruit volunteer respondents willing to participate, hence not all of the potentially covered population would have a non-zero probability of being drawn. This is commonly known as self-selection bias.

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

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

tiple sources for estimation in nonprobability survey sampling has the potential to overcome current methods in terms of bias reduction and efficiency of the estimators.

The remainder of the article is organized in four sections. After introducing the problem of estimation in Section 2, in Section 3, new estimators are proposed based on different approach to integrate data. In Section 4, we propose the use of resampling techniques for the variance estimation of the quantile estimators proposed in the previous section. Some simulation experiments are carried out to check the finite size sample properties of the proposed estimators in Section 6. Finally, Section 7 presents the concluding remarks.

2. The problem of estimation with non-probability samples

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

A matching estimator is defined by:

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

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

The key is how to predict the values y_k. Formal working linear regression models, relating the study variable y to the vector of auxiliary variables are usually considered to develop efficient estimators of the total Y. Suppose a working population model, $E_m(y_i) = m(x_i, \beta) = m_i$ for $i \in U$ is assumed to

hold for the sample s_V where E_m denotes model expectation and the mean function m_i is specified. Using the data from the sample s_V we obtain an estimator $\hat{\beta}$ which is consistent for β if the model is correctly specified and thus the estimator $\hat{Y}_{SM}$ is consistent if the model for the study variable is correctly specified but the estimator will be biased if the model for the study variable is incorrectly specified. Parametric models require assumptions regarding variable selection, the functional form and distributions of variables, and specification of interactions. Contrary to statistical modelling approaches that assume a data model with parameters estimated from the data, more advanced machine learning algorithms aim to extract the relationship between an outcome and predictor without an a priori data model. These methods have been recently applied in the statistical matching context in [**?**].

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

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

$$I_i = \begin{array}{ll} 1 & i \in s_V \\ 0 & i \notin s_V \end{array}, i = 1, 2, ..., N \tag{1}$$

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

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

The probability reflects the selection mechanism of the non-probability sample. Depending on the mechanism, the conditional probability might vary. If the selection is Missing Completely At Random (MCAR), then $P(I_i = 1|\mathbf{x}_i, y_i) =$

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

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

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

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

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

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

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

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

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

3.1. Shrinkage

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

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

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

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

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

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

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

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

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

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

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

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

$$K_{opt} = \frac{V_2 - C}{V_1 + V_2 - 2C}.$$

The second derivative is

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

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

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

The optimal coefficient K_{opt} depends on population variances, which are usually unknown in practice, and so $\hat{Y}_{srk_{opt}}$ cannot be calculated. By substituting V_1 and V_2 by its sample-based analogues

The following estimator can be defined

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

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

3.2. Double robust estimator

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

The total Y can be written as:

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

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

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

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

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

3.3. Training data with PSA weights

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

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

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

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

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

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

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

Thus the proposed estimator will be obtained with the algorithm 1:

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

4. Simulation study

4.1. Data

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

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

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

companys income tax, the size (small, medium or big) and whether it is ISO certified.

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

4.2. Simulation

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

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

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

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

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

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

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

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

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

4.3. Results

The results obtained for the bias and RMSE can be consulted in Tables 1 and 2 respectively. Table 3 contains the summary comparing each method. Both shrinkage estimators are referred to as K_1, for $K_1 = s_r/(s_r + s_v)$, and K_2, for $K_2 = V(\hat{\theta}_{PSA})/(V(\hat{\theta}_{PSA}) + V(\hat{\theta}_{SM}))$. The estimator based on the idea of Chen et al. (2019) is referred to as *Chen*. The estimator which uses PSA weights when training the Statistical Matching model is referred to as *Training*.

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

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

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

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

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

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

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

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

5. Conclusions

Selection bias, a growing issue in survey sampling and empirical sciences due to new questionnaire administration methods, appears when a sample is drawn from a potentially covered population which is different on its composition to the target population. If a sample drawn from the target population is available, some methods can be applied to adjust for selection bias in the nonprobability sample. Propensity Score Adjustment (PSA) and Statistical Matching are the most important methods up to date, both of them showing an increase in efficiency when applied to the estimation of a population parameter. In this context, it is feasible that a combination of both methods could result in an advantage in terms of bias and error reduction, especially given that they can be complemented as they have different outcomes (weights in PSA and predictions in Matching). Previous work by [**?**] proved that a doubly-robust estimator could provide acceptable results, with good properties.

In this study, shrinkage methods to combine two estimates, doubly-robust estimation and the use of PSA weights in the training of models to be used for Statistical Matching are compared in terms of bias and RMSE. The results are obtained from simulations with three different datasets to enable the study of the behavior of such methods under different conditions. Results show a

certain advantage of the training method developed in this paper over the model-assisted estimator, and an advantage of both of them over Statistical Matching. Shrinkage and PSA stand far below, although they offer competitive results under certain circumstances.

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

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

Acknowledgements

This work is partially supported by Ministerio de Economía y Competitividad of Spain (grants MTM2015-63609-R and PID2019-106861RB-I00).

References