SORT 42 (2) July-December 2018, 159-182 DOI: 10.2436/20.8080.02.73

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

Ramón Ferri-García and María del Mar Rueda*

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 analyse the efficiency of correction techniques in multiple situations, applying a combination of propensity score adjustment and calibration on both types of variables (correlated and not correlated with the missing data mechanism) 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.

MSC: 62D05

Keywords: Online surveys, Smartphone surveys, propensity score adjustment, calibration, simulation

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, 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–

* *Corresponding author:* mrueda@ugr.es
Department of Statistics and Operations Research, University of Granada, Avda. Fuentenueva s/n 18004 Granada, Spain. rferri@ugr.es, mrueda@ugr.es
Received: April 2018
Accepted: November 2018

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 above 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 completed 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 mode in which they are completed: 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 completed 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, 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 a 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 (e.g., 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 taken 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 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 with sociodemographic variables which could be eventually related to the outcome variables of a certain study. To mention some examples, data from the Pew Research Center (2017) reveal 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 internet penetration rate is above 90% for all age groups below 54 years of age, in citizens between 65 and 74 years old 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, Rafferty and Valliant, 2008); even though 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 behaviour 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 variable of interest 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, with population total 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), the linear distance being 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, λ, 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, Särndal and Wretman, 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 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 estimated 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, 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 completed 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 this 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 behaviour 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, z, 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, π, can be displayed as

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

for some vector γ, 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):

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

where s_{Rg} is the set of individuals in the reference sample that are in the gth class of propensity scores, and d_{Rk} is the original design weight of the k individual in the reference sample, s_{Vg} is the set of individuals in the volunteer sample that are in the gth 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_{Vg} (i. e. the individual k of the gth 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.

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

Alternatively, the approach proposed by 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)

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

where $\hat{\pi}(\mathbf{x}_k)$ is the estimated response propensity for the individual k 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 analyse 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 a 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:

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

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

- 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 at 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 the Spanish adult population (National Institute of Statistics, 2017c). These probabilities can be consulted in Table 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.
- Probability of voting for 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 for 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 for 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 for 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 how 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.

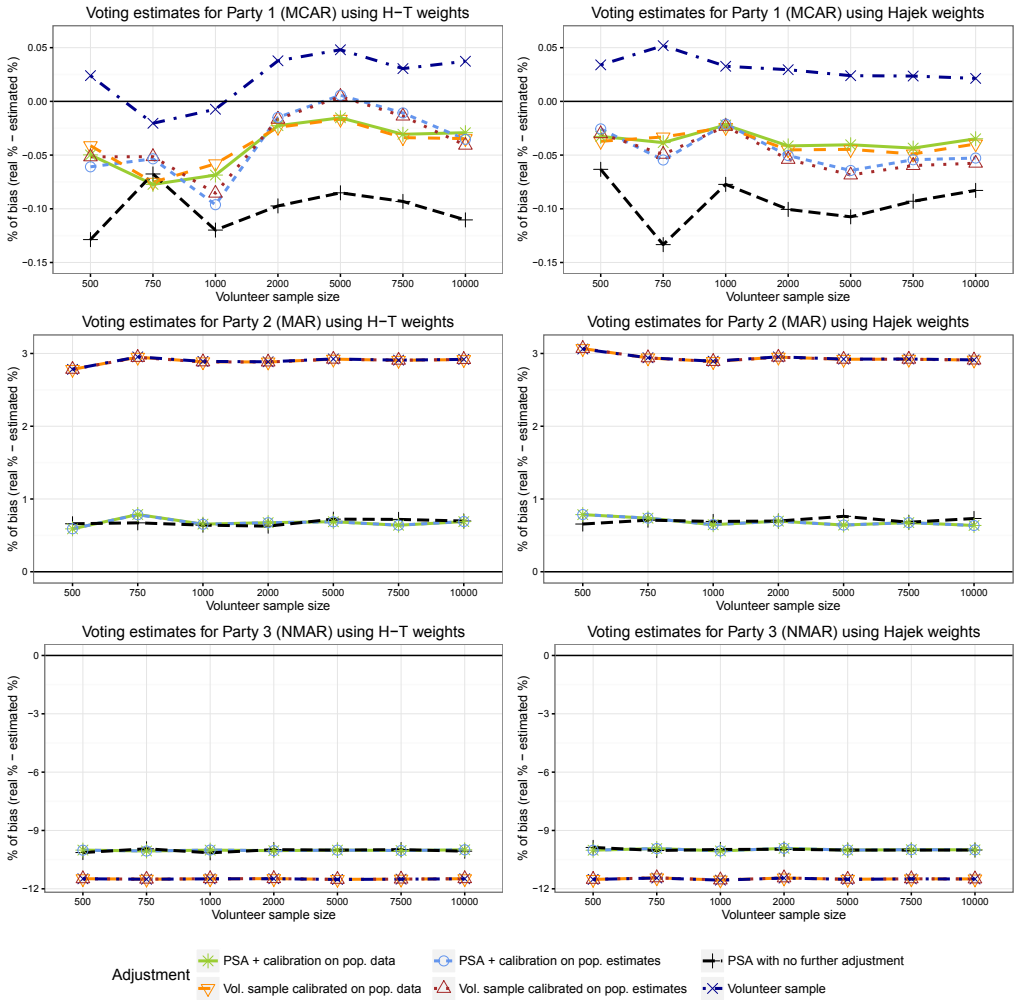

Figure 1: *Bias of each method in voting intention estimations by party in Situation 1.*

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.

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

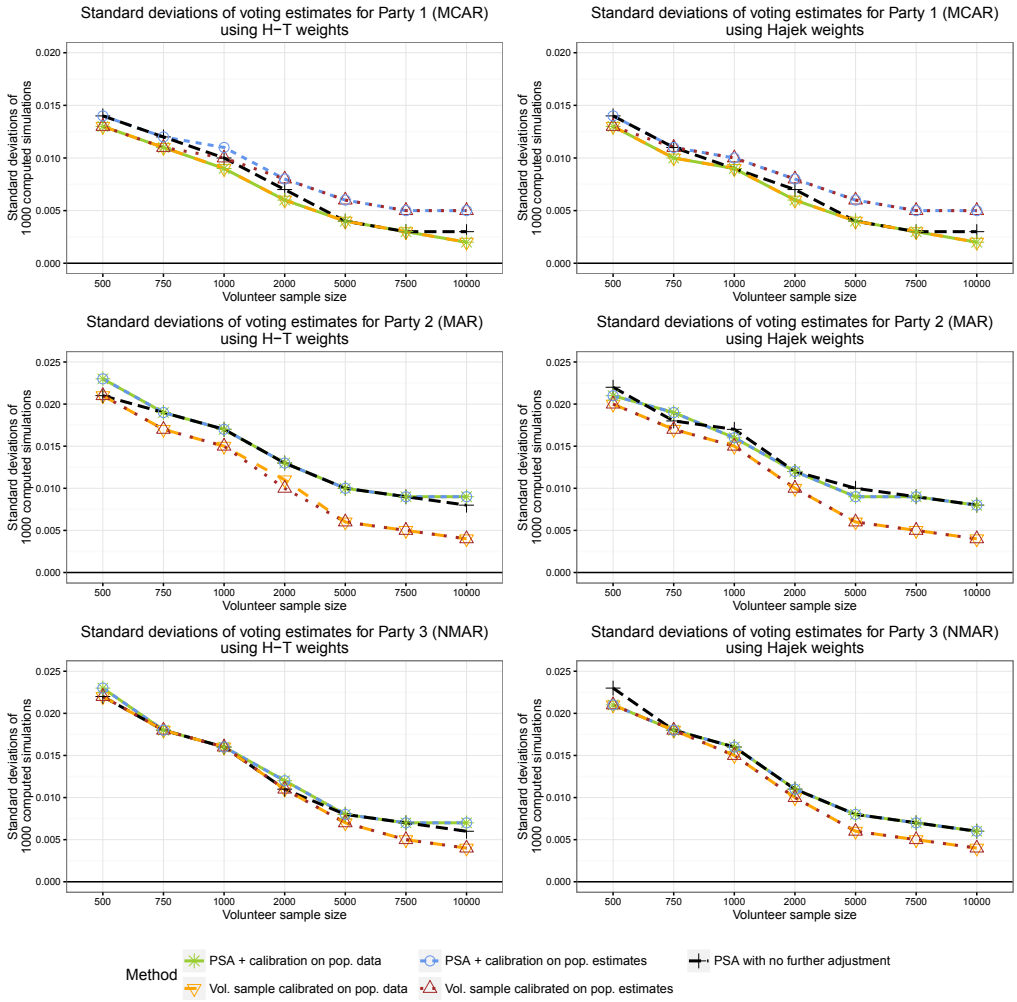

Figure 2: *Standard deviation of voting intention estimations by party provided by each method in Situation 1.*

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 mentioning 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.

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

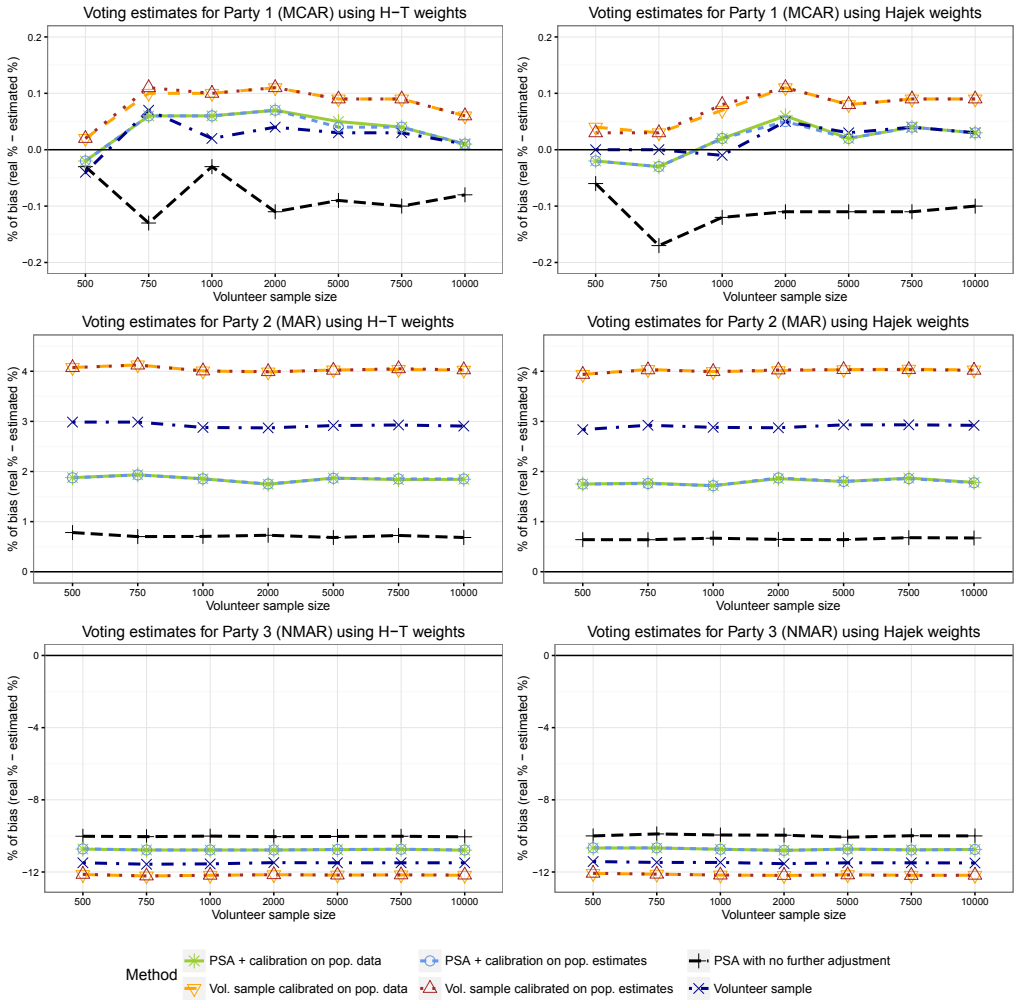

Figure 3: *Bias of each method in voting intention estimations by party in Situation 2.*

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.

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 mentioning 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

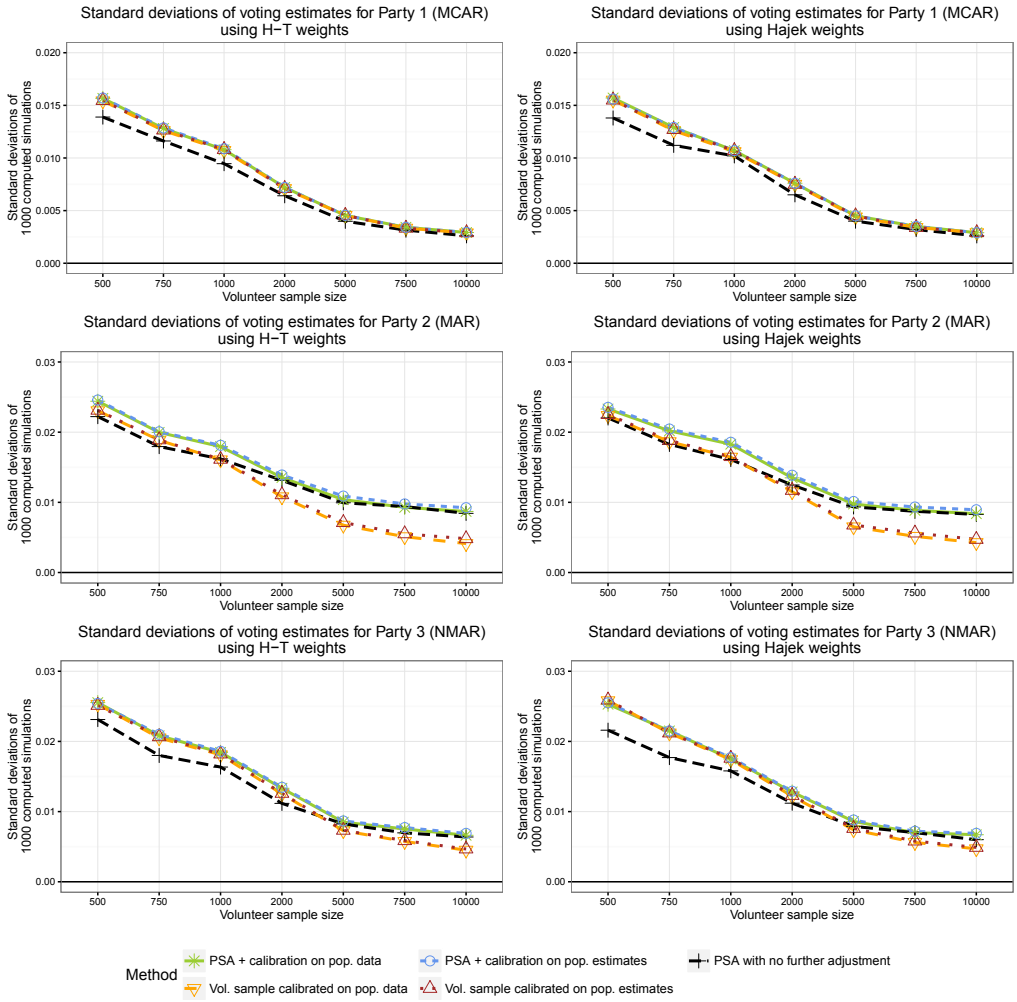

Figure 4: *Standard deviation of voting intention estimations by party provided by each method in Situation 2.*

Hajek-type estimators have a greater variance than Horvitz-Thompson-type estimators, specially when the volunteer sample size is relatively small.

Figures 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.

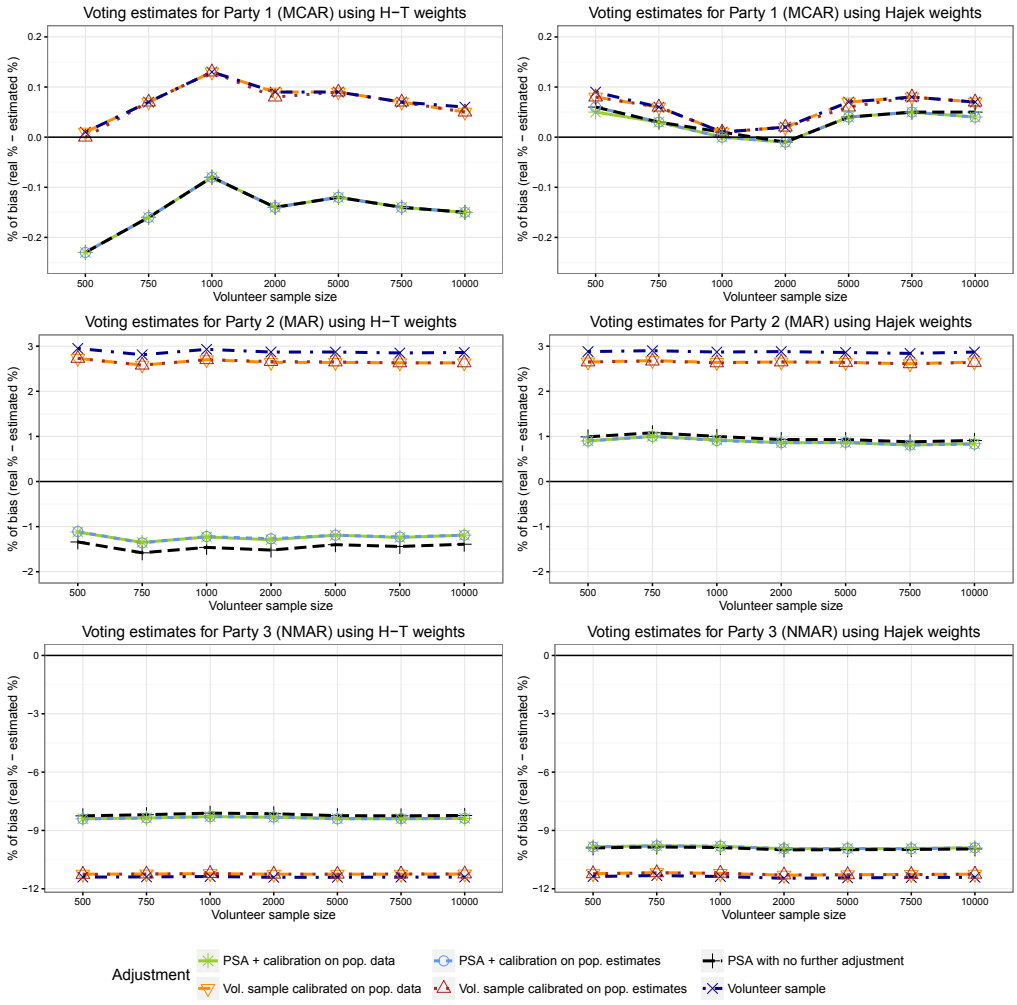

Figure 5: *Bias of each method in voting intention estimations by party in Situation 3.*

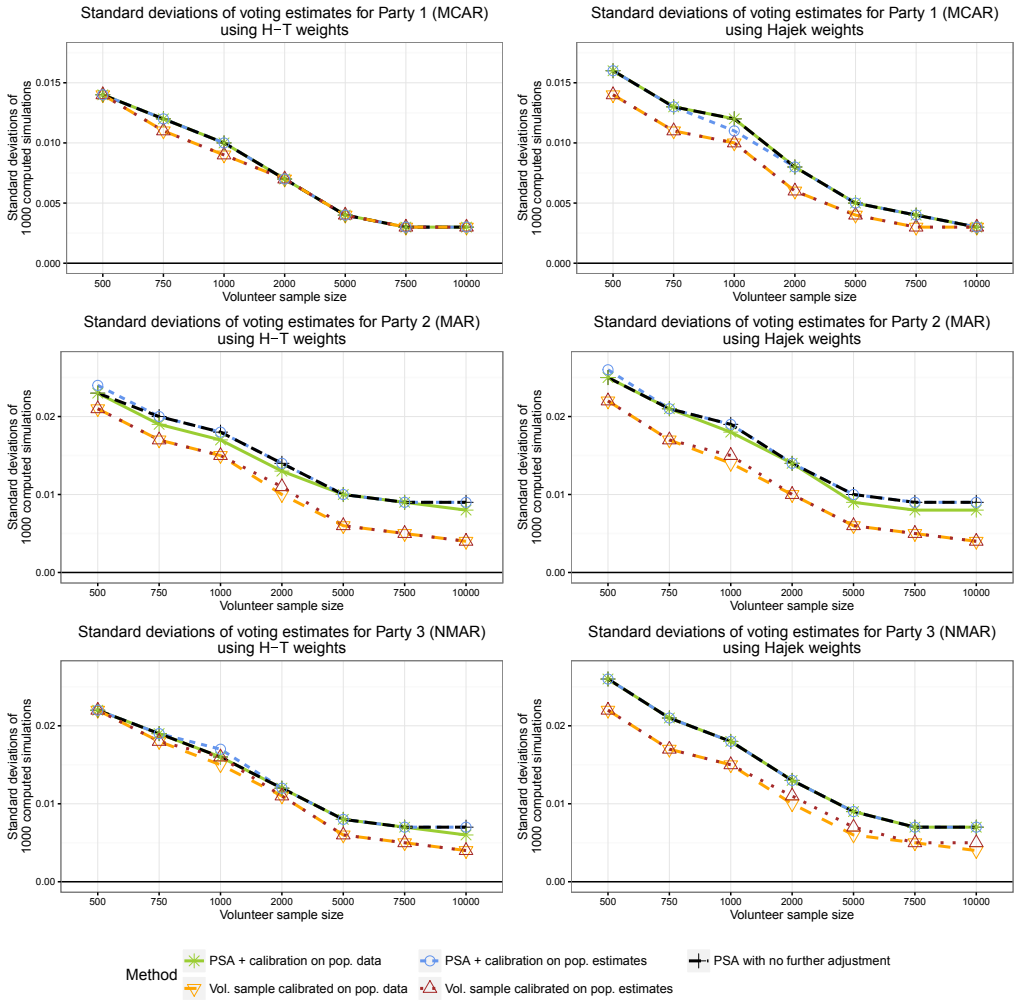

Figure 6: *Standard deviation of voting intention estimations by party provided by each method in Situation 3.*

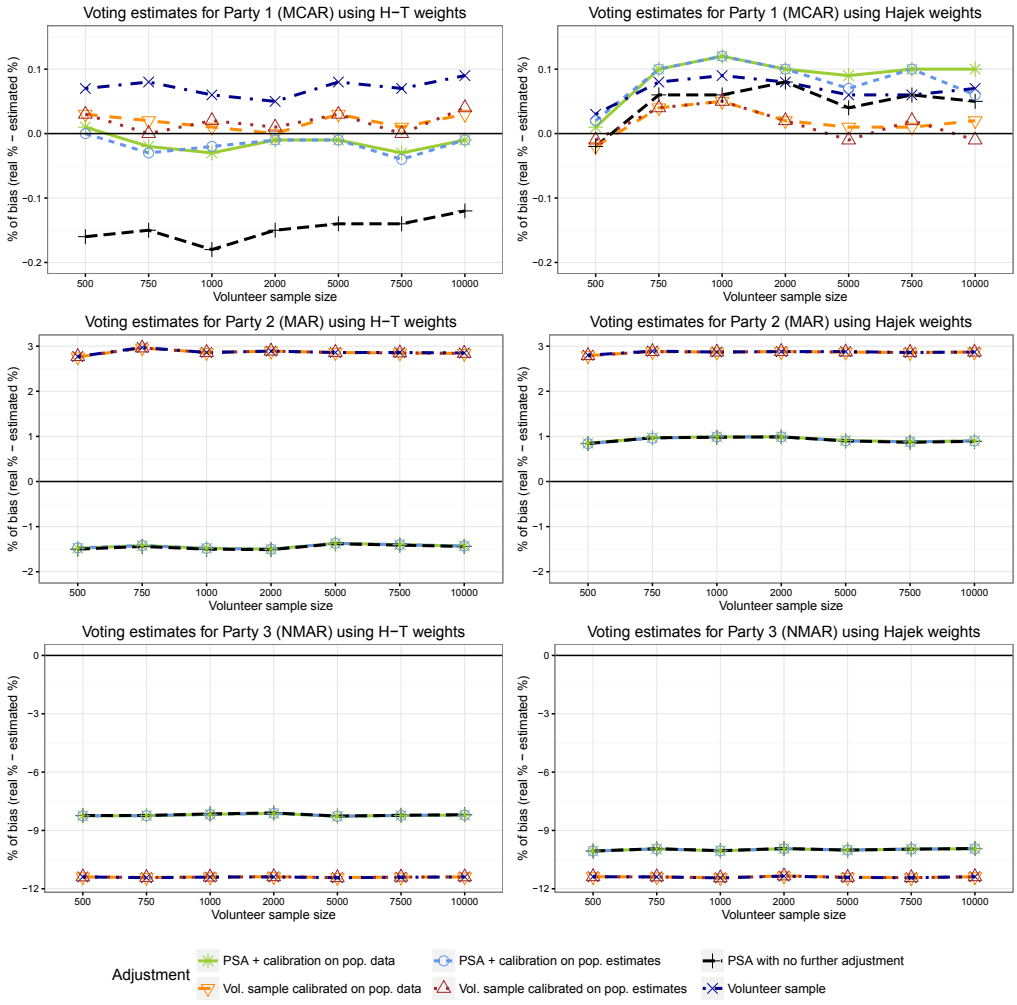

Figure 7: *Bias of each method in voting intention estimations by party in Situation 4.*

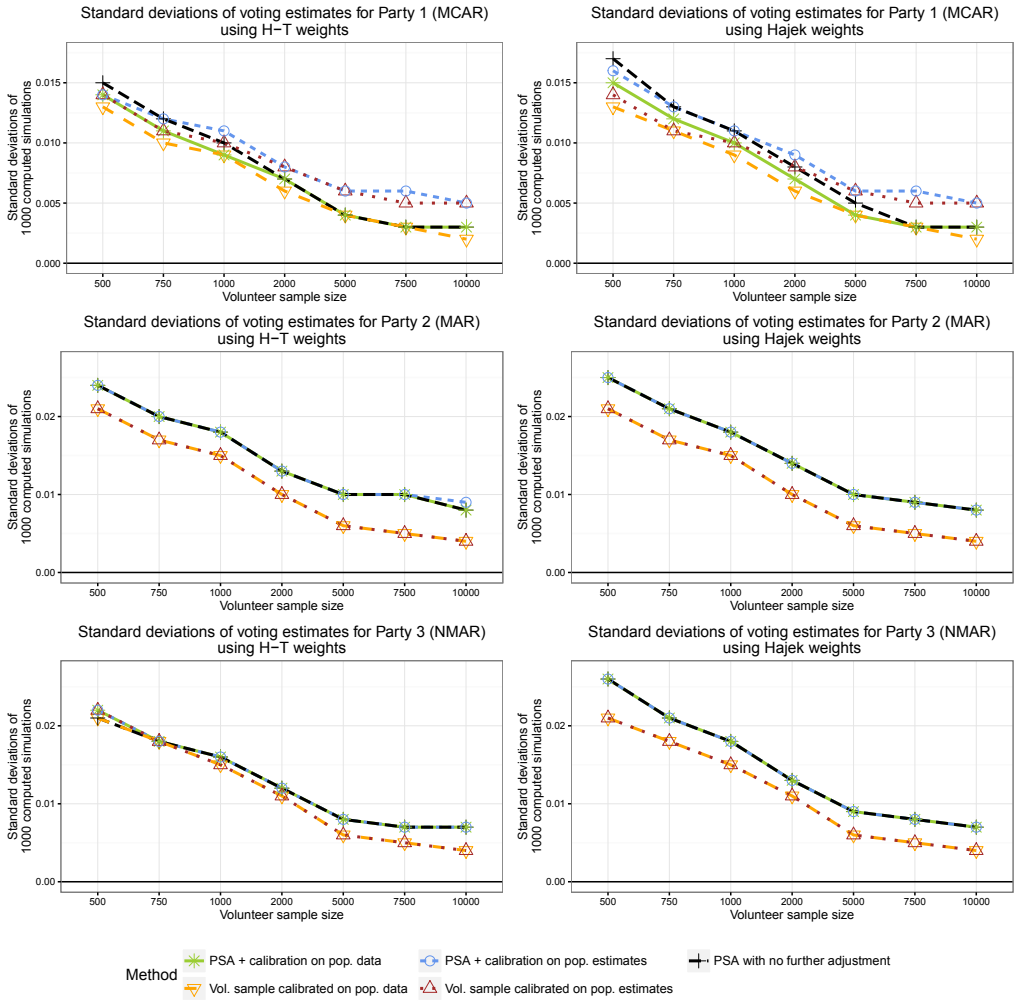

Figure 8: *Standard deviation of voting intention estimations by party provided by each method in Situation 4.*

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 $\pm$ 3.3% given the sample size and a confidence level of 95%. Respondents had to complete 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 completed 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 showed 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.

The difference in gender proportions between both samples is statistically significant ($p = 0.0012$), 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

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

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.

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 the jackknife method.

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 Estimated	Std. Err.	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 %

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 non-probabilistic 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 analyse 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, 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, 295–313.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 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, 357–377.

Dever, J. A., Rafferty, A. and Valliant, R. (2008). Internet surveys: can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2, 47–62.

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

Díaz de Rada, V. (2012). Ventajas e inconvenientes de la encuesta por internet. *Papers*, 97, 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, 337–348.

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

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 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, 111–121.

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

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22, 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, 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, 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 (españoles/extranjeros) por edad (grupos quinquenales), sexo y año. 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). España 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 años). Retrieved from http://www.ine.es/ss/Satellite?c=INESeccionC&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&cid=1259925481659&L=0l (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, 41–55.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 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, 99–119.

Schonlau, M. and Couper, M. (2017). Options for conducting web surveys. *Statistical Science*, 32, 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, 291–318.

Pasadas-del-Amo, S. (2018). Cell phone-only population and election forecasting in Spain: The 2012 regional election in Andalusia. *Revista Española de Investigaciones Sociológicas (REIS)*, 162, 55–72.

Taylor, H. (2000). Does internet research work? *International Journal of Market Research*, 42, 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, 127–135.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105–137.