Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Advances in the use of auxiliary information for estimation from nonprobability samples.

by Ramón Ferri García

Release date: October 15, 2021

Statistics Canada
Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Advances in the use of auxiliary information for estimation from nonprobability samples.

Ramón Ferri-García[1]

Abstract

Recent developments in questionnaire administration modes and data extraction have favored the use of nonprobability samples, which are often affected by selection bias that arises from the lack of a sample design or self-selection of the participants. This bias can be addressed by several adjustments, whose applicability depends on the type of auxiliary information available. Calibration weighting can be used when only population totals of auxiliary variables are available. If a reference survey that followed a probability sampling design is available, several methods can be applied, such as Propensity Score Adjustment, Statistical Matching or Mass Imputation, and doubly robust estimators. In the case where a complete census of the target population is available for some auxiliary covariates, estimators based in superpopulation models (often used in probability sampling) can be adapted to the nonprobability sampling case. We studied the combination of some of these methods in order to produce less biased and more efficient estimates, as well as the use of modern prediction techniques (such as Machine Learning classification and regression algorithms) in the modelling steps of the adjustments described. We also studied the use of variable selection techniques prior to the modelling step in Propensity Score Adjustment. Results show that adjustments based on the combination of several methods might improve the efficiency of the estimates, and the use of Machine Learning and variable selection techniques can contribute to reduce the bias and the variance of the estimators to a greater extent in several situations

Key Words: nonprobability sampling; calibration; Propensity Score Adjustment; Matching.

1. Introduction

Traditional survey methods are facing important drawbacks, in terms of response rates, costs, and coverage errors, and new questionnaire administration methods could fix those gaps. Among these new methods, we could mention online and smartphone surveys. Passive data sources often used in the so-called Big Data analyses (such as GPS, web scrapping, mobile apps) can also be considered as new methods to obtain samples from a population.

The new methods constitute tools to obtain more timely samples at cheaper costs and with many more possibilities regarding questionnaire design or even targeting non-demographical strata that can be reach using online tools. However, they also entail non-sampling errors that should be taken into account. For example, Internet connection or questionnaire computerization problems can involve measurement error, and the absence of an interviewer may enhance survey satisficing behaviors. The most relevant source of error in new survey methods is selection bias, which comes from three different mechanisms: coverage error, response error and self-selection bias.

All of these mechanisms may lead to important amounts of bias if the individuals that participate in the samples differ on their characteristics to those that do not participate in the samples (Elliott and Valliant, 2017). The error formula developed by Meng (2018) shows that the estimation error has three sources: the amount of data drawn from the population, the variability of the data itself and the correlation between the selection mechanism and the variable of interest. If there are differences in the values of the variable of interest between sampled and non sampled individuals, there will be a high correlation that will lead to a larger estimation error. According to Meng (2018), the most efficient way to reduce the estimation error is not to increase the sample size, which will provide very slow improvements, but to reduce this correlation. For this reason, some adjustments have been proposed in literature to reduce selection bias.

[1] Ramón Ferri García, Department of Statistics and Operations Research, University of Granada, Avenida de la Fuente Nueva S/N, Granada, Spain, 18071 (rferri@ugr.es)

These adjustments are described in Section 2. In Section 3, some recent advances are described, and some future research lines are considered in Section 4.

2. Adjustments for nonprobability samples

2.1 Framework

Let U be a target population of size N, and $U_{pc} \subset U$ a potentially covered population which is composed by those that can belong to the nonprobability sample. Let s_r be a probability sample of size n_r drawn from U with a given sampling design, with design weights d^r, and let s_v be a nonprobability sample of size n_v drawn from U_{pc} with no sampling design. Let y be a variable of interest which has been measured in the nonprobability sample but not in the probability sample, and $\boldsymbol{x}$ a set of auxiliary variables that have been measured in both samples. Consider an indicator variable R which measures the inclusion of an individual in the nonprobability sample, such that

$$R_i = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases}$$

2.2 Calibration

An adjustment that does not actually require a probability sample is calibration. This approach was originally developed by Deville and Särndal (1992) to obtain more efficient estimates but then was adapted to correct for survey nonresponse, and it only requires to know the population totals of $\boldsymbol{x}$. This method is based in obtaining a new vector of weights, w, which minimizes a distance function $g(.,.)$ with respect to the design weights d of a given sample s, while respecting the calibration equations

$$\sum_{i \in s} \boldsymbol{x}_i \, w_i = \sum_{i \in U} \boldsymbol{x}_i$$

according to which, when estimating the population totals of the auxiliary variables with the new weights, they should be exactly the same as the actual population totals. Given that there is no design in nonprobability samples, there are no design weights to which the distance should be minimized. The usual choice is to assume uniform weights in this context, such that $d = N/n_v$. A simulation study by Bethlehem (2010) showed good results with this setup.

2.3 Propensity Score Adjustment

Other adjustment, based in a very popular method developed by Rosenbaum and Rubin (1983) to make inference from non-randomized experiments, is Propensity Score Adjustment (PSA). The original theoretical development for online surveys was done by Lee (2006) and has continued to develop since then. In PSA, we assume that inclusion probabilities for s_v, π, are unknown but related to a set of covariates, such that

$$\pi_i = P(R = 1 \mid \boldsymbol{x}_i), \quad i \in U$$

This assumption allows us to estimate those propensities using a predictive model M (usually logistic regression) by fitting it to predict the indicator variable using both pooled samples.

$$\hat{\pi}_i^* = E_M[R^* = 1 \mid \boldsymbol{x}_i], \ \ i \in s_r \cup s_v,$$

where R^* is equivalent to the indicator variable R but only considering the pooled samples such that

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

The variable R^* can be considered a measurement of R. The estimated propensities can be transformed into weights afterwards, using the inverse probability approach $w_i = 1/\pi_i$ or the modified version $w_i = (1 - \pi_i)/\pi_i$ considered in Schonlau and Couper (2017) which takes into account the fact that s_v does not belong to the target population of s_r. Other alternatives involve the stratification of propensities, in order to classify individuals with similar propensities into a same stratum. Some alternatives have been presented in Lee and Valliant (2009) and Valliant and Dever (2011).

2.4 Superpopulation modeling estimators

Both of the adjustments described in previous sections are design-based adjustments, focused on modeling the participation propensities, but some model-based approaches have been also developed in literature. In such approaches, we assume that the population U is a realization of a superpopulation random variable which is related to some covariates via a given function with a model m and an error term:

$$y_i = m(\boldsymbol{x}_i) + e_i, \;\; i \in U, \; e \sim N(0, \, \sigma^2)$$

For the class of estimators described in this subsection, we assume that a complete census of U is available for $\boldsymbol{x}$, meaning that we can consider that the probability sample s_r observed in our study is a census. In this situation, we can use data from the nonprobability sample s_v to fit a predictive model SP on the variable of interest y, and then use it to predict its values for the whole population, such that we obtain the vector of new predicted values, $\hat{y}$:

$$\hat{y}_j = E_{SP}\left[y_j \middle| \boldsymbol{x}_j\right], \;\; j \in U$$

The predicted values can be used to obtain estimates of the population total of y, Y, using several formulas, by replacing the original probability samples, for which these estimators were developed, with the nonprobability sample:

- The model-based estimator (Royall, 1970):

$$\hat{Y}_{MB} = \sum_{i \in s_v} y_i + \sum_{i \in U - s_v} \hat{y}_i$$

- The model-assisted estimator (Cassel et al., 1976):

$$\hat{Y}_{MA} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s_v} w_i \, (y_i - \; \hat{y}_i)$$

- The model-calibrated estimator (Wu and Sitter, 2001):

$$\hat{Y}_{MC} = \sum_{i \in s_v} w^{MC} \, y_i,$$

  where w^{MC} are calibration weights which minimize the distance with the design weights of s_v while respecting the calibration equations

$$\frac{1}{N} \sum_{i \in s_v} w^{MC} = 1, \;\; \sum_{i \in s_v} w^{MC} \, \hat{y}_i = \sum_{i \in U} \hat{y}_i$$

  The restriction $\frac{1}{N} \sum_{i \in s_v} w^{MC} = 1$ can be dropped out, leading to an alternative estimator $\hat{Y}^*{}_{MC}$.

2.5 Sample Matching (Mass Imputation)

In the case where the probability sample s_r is not the complete census but only a subset of the population (drawn according to a probability sampling design), we can apply the same approach based on predicting the values of y in

s_r, which is known in literature as Statistical Matching, Sample Matching or Mass Imputation. This approach was firstly introduced by Rivers (2007) and Vavreck and Rivers (2008), and it is based on fitting a predictive model SM using data from the nonprobability sample s_v and apply it in the probability sample s_r to predict the unobserved values of the variable of interest y:

$$\hat{y}_j = E_{SM}\left[y_j \middle| \boldsymbol{x}_j, R_j\right],\ \ j \in s_r$$

The predicted values can be treated as true values in the usual estimators of the population mean, $\bar{Y}$, and population total with the design weights of the probability sample:

$$\hat{Y}_{SM} = \sum_{i \in s_r} d_i^r \hat{y}_i\,,\ \ \hat{\bar{Y}}_{HT}^{SM} = \frac{1}{N}\sum_{i \in s_r} d_i^r \hat{y}_i\,,\ \ \hat{\bar{Y}}_{H}^{SM} = \frac{\sum_{i \in s_r} d_i^r \hat{y}_i}{\sum_{i \in s_r} d_i^r}$$

2.6 Doubly Robust estimation

A recent approach by Chen et al. (2020) considers the combination of both the design-based and the model-based approach in their doubly robust estimator. In this estimation, we predict the values of the target variable in the probability sample and use it as in Sample Matching, but we add a term that takes into account the prediction errors, as observed in the nonprobability sample, and each prediction error is elevated by weighting it with the inverse propensity estimated with PSA. The Doubly Robust estimator of the population mean can be defined as follows:

$$\hat{\bar{Y}}^{DR} = \frac{1}{N}\sum_{i \in s_v} \frac{y_i - \hat{y}_i}{\hat{\pi}_i} + \frac{1}{N}\sum_{i \in s_r} d_i^r \hat{y}_i,$$

where N can be substituted by $\hat{N} = \sum_{i \in s_r} d_i^r$ to obtain the Hajek estimator. This approach is doubly robust because it is robust to misspecifications in the prediction of $\hat{\pi}$ done in PSA or in the prediction of $\hat{y}$ done in Sample Matching.

3. Recent advances

3.1 Combination of PSA and calibration

Considering this theoretical framework, some contributions have been done based on several methodologies. The first one (Ferri-García and Rueda, 2018) studies the combination of PSA and calibration, substituting design weights in calibration by the propensity weight obtained with PSA. This combination had been studied in literature (Lee and Valliant, 2009) but in this contribution we considered the use of estimated population totals in calibration, instead of actual ones, such that the calibration equations are

$$\sum_{i \in s_v} \boldsymbol{x}_i\, w_i = \sum_{i \in s_r} \boldsymbol{x}_i$$

We also considered two different procedures to transform propensities into weights: the inverse probability weighting $d_i = (1 - \hat{\pi}_i^*)/\hat{\pi}_i^*, i \in s_v$ and the propensity stratification proposed in Lee and Valliant (2009). The simulation featured a fictitious population with four covariates and a variable of interest representing the vote to three fictitious political parties. The estimation of the voting percentage to each one of them was subject to Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) mechanisms. Several configurations for the covariates were also considered.

Results of the mean root mean square error (RMSE) observed in the simulation can be consulted in Table 3.1-1. RMSE of the combination of methods for MAR data was slightly lower than using PSA alone, although the error varied significantly across the scenarios considered in the simulation. The results suggest that using estimated totals can work as well as using actual population totals, and that the combination is helpful if the right covariates are used.

Table 3.1-1
Mean RMSE of the estimates in the scenarios considered in Ferri-García and Rueda (2018)

RMSE of estimates	Unweighted	PSA	PSA + calib. (real totals)	PSA + calib. (estimated totals)
MAR variable	0.0367	0.0226	0.0224	0.0225
MNAR variable	0.1171	0.1013	0.1023	0.1023

3.2 Combination of PSA and Matching

The combination of PSA and Sample Matching has been also considered (Castro-Martín et al., 2021) in a slightly different way than the Doubly Robust estimator, by using the estimated propensities to build weighted models (where the input weights are $w_i = \frac{1}{\hat{\pi}_i^*}, i \in s_v$) which could then be used for predicting the values of the target variable. This means that the estimator is the same as in Sample Matching, but the model used for prediction is trained using weights already obtained in PSA. Our simulation study involved three pseudo-populations with two different sampling schemes to obtain s_v in each one of them. More details can be consulted in Castro-Martín et al. (2021). The results regarding Mean Square Error (MSE) observed in the simulations are summarized in Table 3.2-1, suggesting that the described combination of PSA and Matching could achieve the same levels of efficiency than the doubly robust estimator or even slightly higher.

Table 3.2-1
Mean and median efficiency (%) of each method and times it has been among the best (MSE less than 1% greater than the minimum MSE) in the simulations performed in Castro-Martín et al. (2021)

Method	Mean	Median	Best
PSA + Matching	65.8	66.4	18
Doubly Robust	64	65.2	18
Matching	61.8	64.2	14
PSA	46.6	53.9	6

3.3 Machine Learning in PSA

Regarding PSA, we studied the use of Machine Learning (ML) classification algorithms for the estimation of propensities as an alternative to logistic regression (Ferri-García and Rueda, 2020). Two simulation studies were performed for the matter using the same fictitious population as in Ferri-García and Rueda (2018) and a pseudo-population based in real data from the Spanish Living Conditions Survey. In both studies, propensities were estimated with a wide range of predictive algorithms.

Results of MSE from the first simulation study can be consulted in Table 3.3-1. When the selection mechanism is MCAR, decision trees offer better results in terms of mean square error, especially the Classification And Regression Trees (CART) algorithm. When the selection mechanism is MAR, it can be observed that, although logistic regression achieves good results, K-nearest neighbors (K-NN) and especially Gradient Boosting Machines (GBM) could be helpful in a range of situations. Finally, when the selection mechanism is MNAR, the best choice for estimation of propensities is Random Forest.

Table 3.3-1
Average MSE provided by each algorithm and times it has been among the best (MSE less than 1% greater than the minimum MSE) in the simulation using a fictitious population from Ferri-García and Rueda (2020)

Variable of interest	Measure	Logistic regression	C4.5	C5.0	CART	K-NN	Naïve Bayes	Random Forest	GBM
MCAR variable	Mean MSE	1.3	1.2	1.0	0.7	1.5	3.1	11.7	1.0
	Best	0	4	3	11	0	0	0	5
MAR variable	Mean MSE	3.7	38.0	42.8	61.3	4.5	16.5	67.5	26.6
	Best	12	0	0	0	1	0	0	10
MNAR variable	Mean MSE	102.9	166.9	175.7	201.3	90.5	73.2	75.3	144.9
	Best	0	0	0	0	2	1	19	0

3.4 Machine Learning in superpopulation modeling

The use of ML in superpopulation modeling was considered in Ferri-García et al. (2021) for the case where a complete census of the population is available. These estimators are strongly dependent on the specification of the model, and in fact it is observed that the choice of the model was the most important factor, regardless of other factors like sample size or the formula used in estimation. In a simulation study using the same pseudo-populations as in Castro-Martín et al. (2021), we found that penalized regression models such as Ridge or Elastic Net (Glmnet) were better than other alternatives, although some algorithms like K-NN were good for some specific situations.

Table 3.4-1

Mean and median efficiency (%) of each method with each algorithm and times it has been among the best (MSE less than 1% greater than the minimum MSE) in the simulations performed in Ferri-García et al. (2021). MA = Model-assisted; MB = Model-based; MC = Model-calibrated.

Method	Algorithm	Mean	Median	Best	Method	Algorithm	Mean	Median	Best
MA	Ridge regression	62.2	64.3	13	MC	Glmnet	61.5	63	12
MB	Ridge regression	61.9	64.1	12	MB	Glmnet	61.3	63	9
MA	GLM	61.7	64.3	12	MC	K-NN	59.1	53.1	7
MB	GLM	61.7	64.1	12	MA	K-NN	58.5	52.7	7
MC	GLM	61.7	64.3	12	MC	Bayesian LASSO	58.5	61.3	10
MC	Ridge regression	61.6	62.8	11	MA	Bayesian LASSO	58.2	61.2	11
MA	Glmnet	61.6	62.8	11	MB	Bayesian-regularized neural networks	57.9	61.8	8

3.5 Automated variable selection prior to PSA

The importance of the choice of covariates has been stated in literature of propensity score weighting (Hirano and Imbens, 2001; Brookhart et al., 2006), concluding that the best choice is to include covariates related to the variable of interest. However, we might not always know the associations or the causal relationships prior to the estimation procedure, so we must learn from data in order to select some variables and discard others. For this reason, we performed a study on the use of automated variable selection algorithms to improve PSA results (Ferri-García and Rueda, 2021). We used some selectors for linear regression (StepWise, LASSO), filters (Chi-Square, CFS, OneR) and importance measures (Random Forest, Boruta) and applied them in two simulation studies, using fictitious data and real data from a survey conducted by the Spanish Centre for Sociological Research.

Some efficiency results from the second simulation study can be observed in Table 3.5-1. In terms of efficiency, some variable selection algorithms, especially OneR or CFS, are associated to more efficient estimates. This advantage can be explained by the smaller variance of the estimators, which are even less biased than the case where all variables are used when PSA is combined with calibration.

Table 3.5-1

Mean and median Efficiency of the estimates (MSE of algorithm/MSE using all vars.) and number of times its Efficiency has been below 1 or 0.9 in the simulation study using real performed in Ferri-García and Rueda (2021). Raking calibration applied after PSA.

Algorithm	Mean	Median	Efficiency < 1	Efficiency < 0.9
Boruta	1.028	1.004	22	1
CFS	0.950	0.943	38	11
Chi-Square	0.968	0.942	36	11

Gain ratio	0.983	0.955	37	9
LASSO	0.976	0.956	34	7
StepWise	1.015	1.010	21	2
OneR	0.965	0.943	42	14
Random Forest importance	0.991	0.973	34	4

4. Future research lines

Some research lines should be considered in further research on the estimation from nonprobability samples using auxiliary information. The inclusion of design weights in predictive models for propensity estimation should be studied. Although a consistent estimator involving design weights was developed by Chen et al. (2020) for logistic regression, other weighting strategies might be more suitable for other parametric and nonparametric models. Another important issue is the mitigation of bias produced by MNAR mechanisms; the bias produced by this mechanism is the most difficult to deal with, according to current research. Other research lines include the development of theoretical properties and the inclusion of other data preprocessing strategies, such as class balancing or hyperparameter tuning, which are common in data science and could be helpful in estimation from nonprobability samples.

References

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

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

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976), “Some results on generalized difference estimation and generalized regression estimation for finite populations”, *Journal of the American Statistical Association*, 63(3), pp. 615-620.

Castro-Martín, L., Rueda, M., and Ferri-García, R. (2021), “Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys”, *Journal of Computational and Applied Mathematics* (in press).

Chen, Y., Li, P., and Wu, C. (2020), “Doubly robust inference with nonprobability survey samples”, *Journal of the American Statistical Association*, 115(532), pp. 2011-2021.

Deville, J. C., and Särndal, C. E. (1992), “Calibration estimators in survey sampling”, J*ournal of the American statistical Association*, 87(418), pp. 376-382.

Elliott, M. R. and Valliant, R. (2017), “Inference for nonprobability samples”, *Statistical Science*, 32(2), pp. 249-264.

Ferri-García, R., and Rueda, M. (2018), “Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys”, *SORT*, 42(2), pp. 159-182.

Ferri-García, R., and Rueda, M. (2020), “Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys”, *PloS one*, 15(4), e0231500.

Ferri-García, R., and Rueda, M. (2021), “Variable selection in Propensity Score Adjustment to mitigate selection bias in online surveys”, *Statistical Papers* (in revision).

Ferri-García, R., Castro-Martín, L., and Rueda, M. (2021), “Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling”, *Mathematics and Computers in Simulation*, 186, pp. 19-28.

Hirano, K., and Imbens, G. W. (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), pp. 259-278.

Lee, S. (2006), “Propensity score adjustment as a weighting scheme for volunteer panel web surveys”, *Journal of official statistics*, 22(2), pp. 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), pp. 319-343.

Meng, X. L. (2018), “Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election”, *The Annals of Applied Statistics*, 12(2), pp. 685-726.

Rivers, D. (2007), “Sampling for web surveys”, paper presented at Joint Statistical Meetings, Salt Lake City, Utah.

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

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

Schonlau, M., and Couper, M. P. (2017), “Options for conducting web surveys”, *Statistical Science*, 32(2), pp. 279-292.

Valliant, R., and Dever, J. A. (2011), “Estimating propensity adjustments for volunteer web surveys”, *Sociological Methods & Research*, 40(1), pp. 105-137.

Vavreck, L., and Rivers, D. (2008), “The 2006 cooperative congressional election study”, *Journal of Elections, Public Opinion and Parties*, 18(4), pp. 355-366.

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