The Survey Statistician, 2024, Vol. 90, 40–47.

Book and Software Review

Software review for inference with non-probability surveys

Beatriz Cobo[1], **Ramón Ferri-García**[2], **Jorge L. Rueda-Sánchez**[3], **María del Mar Rueda**[4],

[1,2,3,4]University of Granada, Spain
[1]beacr@ugr.es, [2]rferri@ugr.es, [3]jorgerueda@ugr.es, [4]mrueda@ugr.es

Abstract

Implementing probability sampling methods has become more challenging as there has been a noticeable decline in response rates with a consequent increase in survey costs. Furthermore, new data sources that have emerged in recent years could be considered alternatives to survey data. Examples include large data sets from sources such as registries or geolocation and web surveys that have the potential to provide estimates, as well as offer easier access to data and lower data collection costs in comparison to traditional probability sampling, leading to larger sample sizes. Given these new forms of sampling, specific software is needed to support theoretical development. We are going to carry out a review of the existing software for this purpose in the most used programming languages in this field (R and Python), indicating the strengths of each of them.

Keywords: non-probability surveys, inference, software, R, Python.

1 Introduction

Survey methodologies are currently in flux due to social and technological changes that have led to a significant increase in refusals to participate and difficulties in accessing individuals to interview. However, the development of new technologies has facilitated the emergence of new data acquisition techniques, such as web surveys, that present great advantages in terms of speed in obtaining data, reduced costs and the possibility of accessing specific population sectors.
Web surveys have replaced face-to-face and computer-assisted telephone interviews as the primary mode of data collection in most countries, and this trend was reinforced as a result of restrictions related to the COVID-19 pandemic. Although web surveys can be probabilistic, in practice many operate via self-selection, meaning that the principles of probability sampling are not applied.

Non-probability surveys have serious problems in calculating estimates because the principles of probability sampling inference are invalid. The first problem is the lack of an adequate sampling frame that allows the selection of samples from the general population in a probabilistic manner, causing selection bias if the covered population differs from the target population (Elliott and Valliant, 2017), therefore, the generalization of the results under these biases is compromised. However, these types of surveys can be useful in some situations, such as complementing a small probability sample with a larger non-probability sample to improve the efficiency of estimates or mitigating selection biases by intentionally focusing on respondent profiles that tend to be underrepresented.

Powerful new methodologies have been developed to infer parameters using data from non-probability samples, and this research has been reviewed by Buelens et. al. (2018), Rao (2022), Valliant (2020), Yang and Kim (2020), among others. The methods considered include propensity score adjustment (Lee and Valliant, 2009), tree-based inverse propensity weighting (Chu and Beaumont, 2019), propensity-adjusted probability prediction (Elliott and Valliant, 2017), inverse sampling (Kim and Wang, 2019), mass imputation or statistical matching (Rivers, 2007), doubly robust methods (Chen et. al., 2020), kernel smoothing methods (Kern et. al., 2021), superpopulation modelling (Buelens et. al., 2018) and combinations of these techniques (Castro et. al., 2022; Liu and Valliant, 2023). In the following we will briefly describe some of the open source software that implement some of these techniques (R and Python) and their main features.

2 R Software

Over the years, software has been developed to carry out estimates in probability surveys, but it cannot be used directly with non-probability surveys. To fill this gap, researchers have developed new functions in which they consider that the samples are obtained without a probability sampling design. If we focus on the free software Rwe find some packages **NonProbEst** (Martín et. al., 2020), **nonprobsy** (Chrostowski and Beręsewicz, 2024), **nppR** (Beaumont and Dhushenthen, 2024), and **KWML** (Wang and Kern, 2023), that implement different estimation methods with non-probability surveys.

2.1 ”NonProbEst” Package

The R package **NonProbEst** (Martín et. al., 2020) is stored in the official R repository `https://cran.r-project.org/web/packages/NonProbEst/index.html` and is explained in detail in Rueda et. al. (2020). This package was the first that implemented the estimation of linear parameters using Propensity Score Adjustment (PSA) (Valliant and Dever, 2011). The package computes estimates on the propensity to participate in the convenience sample based on classification models to be selected by the user. From these propensities the program calculates pseudo-weights for the non-probability sample units using 4 methods: inverting the propensity ($valliant\_weights$), inverting the propensity minus 1 as in Schonlau and Couper (2017) ($sc\_weights$), using the propensity score averaging design weights formula introduced in Lee and Valliant (2009) ($lee\_weights$), and using the propensity stratification averaging formula introduced in Valliant and Dever (2011) ($vd\_weights$). These weights can be downloaded for use in any subsequent statistical analysis that the researcher wishes to perform with his non-probability data.

The package implements functions to calculate the estimator of the total, mean, and the proportion using these four versions of PSA, as well as other techniques as PSA plus calibration (Lee and Valliant, 2009; Ferri and Rueda, 2018), mass imputation (Rivers, 2007; Beaumont, 2020), or superpopulation models (Buelens et. al., 2018), including model-based, model-calibrated, and model-assisted estimator. One of the main features of the package is that a wide range of statistical models and machine learning (ML) algorithms can be used to leverage the information provided by the auxiliary variables, because all the functions that compute the estimators have an argument in which we can include the machine learning model of our choice from those available in the *train()* function of the **caret** package (Kuhn, 2008).

NonProbEst package, in addition to offering the estimates, also allows users to calculate the variance using two alternatives of Leave-One-Out jackknife (Quenouille, 1956), with and without reweighting in each iteration, and the confidence intervals. It also provides a data set of a simulated fictitious population of 50000 individuals, a probability sample (drawn with simple random sampling from the simulated population and sample size 500) and a non-probability sample (drawn from the simulated population and sample size 1000), to test the implemented functions.

2.2 “nonprobsvy” Package

The newly released package **nonprobsvy** (Chrostowski and Beręsewicz, 2024) is deposited in `https://github.com/ncn-foreigners/nonprobsvy` and is fully explained in Chrostowski and Beresewicz (2023). The goal of this R package is to carry out statistical inferences with non-probability survey samples (including big data) when auxiliary information from external sources like probability samples or population totals/means are available.

The R package **nonprobsvy** implement the first alternative of PSA, also called inverse probability weighting estimators (IPW) (see Chen et. al., 2020), with possible calibration constraints and using as predictive model for propensity scores (selection model) logistic regression (GLM) with logit, probit, and log-log functions; mass imputation (Yang et. al., 2021), using logistic regression (GLM), nearest neighbour (NN) and predictive mean matching (PMM) (Kim et. al., 2021); and the doubly robust estimator (Yang et. al., 2020). The package also allows variable selection in high-dimensional spaces using SCAD (Yang et. al., 2020), Lasso and MCP penalty (via the ncvreg, Rcpp, RcppArmadillo packages), estimation of variance using analytical and bootstrap approaches (Wu, 2022) and computation of the estimated covariance matrix for model coefficients. The software therefore stands out for the wide variety of current techniques implemented for parameter estimation. It also contains a simulated population, described in Chen et. al. (2020), with size desired by the user, and from this population it extracts a non-probability sample, also with size selected by the user.

2.3 “nppR” Package

The R package **nppR** (Beaumont and Dhushenthen, 2024) is stored on the Github site `https://github.com/StatCan/nppR`. This package provides us with some tools to carry out estimations with non-probability samples, using the information of a probability sample in a complementary way. In particular, this package is revolutionary because allows us to use in our experiments the tree-based inverse propensity weighting (TrIPW) estimator, which has not yet been developed in any software. The package was developed in 2022 and is based on the method explained in Chu and Beaumont (2019), which estimates the inclusion probabilities of individuals in the non-probability sample using a modification of the Classification And Regression Tree (CART) algorithm, using information from both samples to more accurately predict whether an individual belongs to the volunteer sample or not. This package also allows us to implement the doubly robust estimator, developed by Chen et. al. (2020), although we can only use linear regression as a predictive model for predicting the variable of interest and logistic regression for estimating inclusion probabilities of the non-probability sample. Finally, the package offers the possibility of creating a synthetic population and probability and non-probability samples from this population.

2.4 ”KWML” Package

The R package **KWML** (Wang and Kern, 2023) is deposited in `https://rdrr.io/github/chkern/KWML/` and is explained in detail in Kern et. al. (2023). This package of functions mainly allows us to compute pseudo-weights for non-probability samples, especially using kernel weighting (KW), with the possibility of estimating propensities with logistic regression models or with ML techniques. As ML techniques, we can use conditional random forests (CRF), gradient boosting machines (GBM) and model-based recursive partitioning (MOB). This package also allows to compute the pseudo KW weights given the previously computed propensity scores, so that we can use any ML technique we want if we have previously estimated the inclusion probabilities of the non-probability sample using the method of our choice. It also contains a dataset containing a simulated non-probability (size 2000) and probability (size 2000) sample.

3 Python

Python is a free software that is being increasingly used by researchers who have to analyze data. If we focus on the Python programming language, the package **inps** (Castro-Martín, 2024) is the first library in Python that implements the main bias adjustment techniques in non-probability surveys. The package is available from the Python Package Index (PyPI) at `https://pypi.org/project/inps/` and is fully detailed in `https://github.com/luiscastro193/inps`.

The library allows statistical inference from non-probability samples and emphasizes an innovative integration of advanced statistical models and machine learning algorithms for bias correction. More precisely, **inps** implements some of the most promising methods for selection bias mitigation such as propensity score adjustment, calibration, methods based on superpopulation modelling (mass imputation and model-based, model-adjusted and model-calibrated estimators), the doubly robust estimator and the PSA-weighted mass imputation estimator from Castro et. al. (2022). Moreover, all of the implemented methods can be easily applied with any machine learning model with a standard API. This allows us to immediately benefit from the huge pool of models already implemented (and those to be added in the future) in Python libraries such us SciPy or scikit-learn. This package also includes a function to pre-process an estimator, which is essential for those models that require numerical data with no missing values. The implementation of these methods is unprecedented in Python and has the potential to allow researchers and practitioners to leverage all the statistical functionalities that Python offers while being able to use the same methods for inference in non-probability samples that were already present in other software.

Six datasets are available in the package. The first three (*nonprobEd, nonprobCovid* and *nonprobHealth*) correspond to three real nonprobability surveys carried out in Spain that address different topics, such as eating disorders, attitudes towards the coronavirus pandemic and satisfaction with life of health professionals. Datasets *probEd* and *probCovid* contain two well-designed probability surveys conducted by official agencies in the same study populations that *nonprobEd* and *nonprobCovid* respectively attempted to cover. They are intended to be used as reference surveys in each of the two cases to adjust for selection biases. The last dataset, *censusHealth*, contains the population frame of units from which the non-probability survey *nonprobHealth* was obtained. These examples show the potential of the package in tackling selection bias and the wide usage possibilities for non-probability sampling estimation for both academics and practitioners.

4 Discussion

In the last few years, a plethora of methods have emerged to mitigate selection bias, employing techniques based on superpopulation models or based on the quasi-aleatorization of the sample, or even the combination of both, which use auxiliary information provided by alternative sources of data. In practice, however, the majority of applications where non-probability samples are considered do not apply any of these corrections, making at most only simple adjustments using basic sociodemographic variables such as gender or age. The scarce use of these methods in practice may be due to their difficult implementation for people who are not experts in sampling. In the last years, several free-use packages have appeared that can contribute to increasing its use in real surveys by researchers.

In this work we have briefly introduced the main packages available in R and Python to obtain estimates from non-probability samples. Now we show a summary table to have all the previously mentioned information concentrated in an easy-to-understand table.

Table 1: Summary table of packages for non-probability bias reduction (see the text for the following abbreviations)

Packages	Software	Techniques	ML algorithms	Datasets	Extra Features
NonProbEst	R	Calibration	“caret” package algorithms	Simulated population with 50000 individuals	Variance and CI with re-sampling techniques
		PSA (4 alternatives)			
		Superpopulation Models			
		Mass Imputation			
		PSA + Calibration			
nonprobsvy	R	IPW	GLM	Simulated population and non-probability sample with selected sizes	Variable selection, Variance with analytical and bootstrap approach, Control parameters for outcome, selection, and inference model, Weights adjustment with probit, logit, and log-log models, Different links for outcome variables, ...
		Mass Imputation	GLM, NN, PMM		
		Doubly Robust	GLM / GLM, NN, PMM		
		IPW + Calibration	GLM		
nppR	R	Doubly Robust	GLM / GLM	Creation of a synthetic population and drawing of probability and non-probability samples	None
		TrIPW	CART		
KWML	R	KW	GLM, CRF, GBM, MOB	Simulated probability and non-probability samples	Can use any ML algorithm if you compute propensity scores separately
		IPW	GLM		
inps	Python	Calibration	“scikit-learn” API compatible packages algorithms	Six datasets (three non-probability samples, two probability samples, and one real population)	Data preprocessing
		PSA (4 alternatives)			
		Superpopulation Models			
		Mass Imputation			
		Doubly Robust			
		KW			
		PSA + Mass Imputation			

We can see in Table 1 that packages that implement the largest amount of techniques are **NonProbEst** package and **inps**, especially the Python package, as it also implements the kernel weighting (KW) estimator, the doubly robust estimator, and mass imputation with PSA weights. This package is unique because it allows such techniques to be used in Python, it is the first to implement the PSA-weighted mass imputation estimator (Castro et. al., 2022) and includes a function to pre-process data for an estimator. We highlight that both packages allow researchers to use their estimators with the vast majority of machine learning (ML) algorithms as predictive models. The **nonprobsvy** package allows estimation using Inverse Propensity Weighting (IPW) estimator, innovating since logit, log-log and probit functions can be used to predict propensity scores; mass imputation estimator, emphasising predictive mean matching (PMM) algorithm as predictive models for the variable of interest; and Doubly Robust estimator, with the aforementioned innovative predictive algorithms for the IPW and Mass Imputation estimators. This package also allows other functionalities such as variable selection before calculating the estimator (SCAD, Lasso and MCP penalty) or variance estimation with the bootstrap technique and analytical formulas, both unique to this package. In the case of the **nppR** package, the doubly robust estimator can be calculated and it is worth noting that we can calculate the tree-based inverse propensity weighting estimator (TrIPW), something that cannot be done with any other package. If we want to compute the KW estimator in R software, we have to use the **KWML** package, which also allows us to use any ML model if the propensity scores have been computed beforehand (for example, with the **NonProbEst** package) with the desired ML algorithm.

Some things are missing in the various packages, for example, a mean square error estimation procedure that takes into account all the sources of randomization that involve the various methods and that are valid for any machine learning technique or estimators to integrate probability and non-probability data (Kim and Tam, 2021; Rueda et. al., 2023; Rueda et. al., 2024). A measure of the final bias in the estimate and its reduction compared to estimators that do not use these adjustment methods would also be useful.

In our opinion, these software described here can serve to advance inference in non-probability sampling, offering a very broad set of specific tools that can be useful both for academic research and for practical implementation.

References

Beaumont, J. F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology, Statistics Canada*, **46**1. `http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-eng.htm`

Beaumont J. and Dhushenthen J. (2024). nppR: Inference on non-probability sample data via integrating probabilty sample data. R package version 1.13.003.

Castro-Martín, L. (2024). INPS: Inference from Non-Probability Samples. Python package version 1.0.

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

Castro L., Rueda M. and Ferri-García R. (2022). Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys. *Journal of Computational and Applied Mathematics*, **404**, 113414.

Chen, Y., Li, P. and Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, **115**(532), 2011–2021.

Chrostowski L. and Beręsewicz M. (2023). nonprobsvy: Package for Inference Based on Nonprobability Samples. R package version 0.1.0. `https://cran.r-project.org/web/packages/nonprobsvy/index.html`

Chrostowski L. and Beręsewicz M (2024). nonprobsvy: Inference Based on Non-Probability Samples. R package version 0.1.0, ¡`https://CRAN.R-project.org/package=nonprobsvy`¿.

Chu, K. C. K. and Beaumont, J. F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting, Calgary, AB, Canada **26**.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Stat. Sci.* **32** , 249–264.

Ferri-García, R. and Rueda, M. d. M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Stat. Oper. Res. T.*, **42**(2), 159–162.

Kern, C., Li, Y. and Wang, L. (2021). Boosted Kernel Weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, **9**,(5),1088—1113. https://doi.org/10.1093/jssam/smaa028.

Kern, C., Li, Y. and Wang, L. (2023). KWML: Boosted Kernel Weighting - Using Statistical Learning to Improve Inference from Nonprobability Samples. `https://rdrr.io/github/chkern/KWML/`

Kim, J. K. and Tam, S. M. (2021). Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference.*International Statistical Review,* **89**, (2), 382—401.

Kim, J. K., Park S., Chen Y. and Wu C. (2021). Combining Non-Probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society* **184** (3): 941—963. https://doi.org/10.1111/rssa.12696.

Kim, J. K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review,* **87**, 177–191.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. `https://doi.org/10.18637/jss.v028.i05`

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

Liu, Z. and R. Valliant. (2023). Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration. *Journal of Official Statistics* **39** (1), 45—78.
DOI: http://dx.doi.org/10.2478/JOS-2023-0003.

Martín L. C., García, R. F. and Rueda, M. d. M. (2020). _NonProbEst: Estimation in Nonprobability Sampling_. R package version 0.2.4, ¡`https://CRAN.R-project.org/package=NonProbEst`¿.

Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika, 43(3/4), 353-—360. [p408]

Rao, J. N. K. (2022) On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, **83**, 242–272.

Rivers, D. (2007). Sampling for web surveys. *In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA*

Rueda, M. d. M., Cobo, B., Rueda-Sánchez, J. L., Ferri-García, R. and Castro-Martín, L. (2024). Kernel Weighting for blending probability and non-probability survey samples. *SORT,* **48**(1), 1–32.

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

Rueda, M. d. M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L. and Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, **65**(2), 2200035.

Schonlau, M., Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science,* **32**(2), 279–292.

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

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

Wang L., Kern, C. (2023). KWML: Boosted Kernel Weighting. R package version 1.0.1.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology, Statistics Canada*, **48**(2), 283—311.

Yang, S., Kim, J. K. and Hwang, Y. (2021). Integration of Data from Probability Surveys and Big Found Data for Finite Population Inference Using Mass Imputation. *Survey Methodology* **47**(1), 29—58. `https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-eng.html`

Yang, S., Kim, J. K. and Song R. (2020). Doubly Robust Inference When Combining Probability and Nonprobability Samples with High Dimensional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **82**(2), 445-–65. `https://doi.org/10.1111/rssb.12354`