IOCMA 2026 Conference

The 2nd International Online Conference on Mathematics and Applications

10–12 June 2026 | Online

MDPI

Statistical Methods for Inference from Non-Probability Samples and their Application to a Health Barometer

Jorge L. Rueda-Sánchez[1], María del Mar Rueda[1], Pablo Ruiz[2]

[1]Department of Statistics and Operational Research, University of Granada, Granada, 18071, Spain.

[2]Faculty of Science, University of Granada, Granada, 18071, Spain.

INTRODUCTION & AIM

We live in the information age, where online surveys, social media, and Big Data provide access to large samples quickly and with minimal resources. However, because these sources are usually non-probability samples (unknown or non-random sampling design), making accurate estimates from them is a major challenge. Without a proper probabilistic sampling design, sample representativeness cannot be guaranteed.

To address this issue, various statistical techniques have been developed to reduce bias. These methods rely heavily on auxiliary information. Currently, three main challenges remain in this field: There is no gold standard, as no single method works best for every scenario; no standardized bias metrics have been developed to date; and there is room for improvement, as methods depend on assumptions that are difficult to hold in practice. Consequently, their performance remains highly context-dependent and varies by specific study.

METHOD

To reduce bias of the estimate of a non-probability sample $s_{np}$with variable of interest y and vector of covariates x but unknown sampling design (unknown design weights d_{np}); there are several statistical techniques that depends on auxiliary information available (Wu, 2022):

- **Population totals of the covariates**: Calibration.
- **Covariates for each individual in the target population**: Superpopulation models.
- **Probability sample s_p (matched x but y unknown):** Propensity Score Adjustment (PSA), Kernel Weighting Method (KW), Statistical Matching (SM) and Doubly Robust estimator (DR).

For this application we focus on the **Statistical Matching (SM)** (Rivers, 2007): Assuming a superpopulation model, the vector of $y = (y_1, \dots, y_N)$ is a realization of the vector of random variables $Y = (Y_1, \dots, Y_N)$, which can be modeled (m) as follows:

$$Y_i = m(x_i, \beta) + e_i, \quad e \sim N(0,1), \quad \forall i = 1, \dots, N.$$

Then, the variable of interest y is modeled with the information from s_{np}, y can be predicted, and we construct the final estimator using the design weights d_p from s_p:

$$\hat{y}_i = m(x_i, \hat{\beta}); \quad \overline{y}_{SM} = \frac{1}{N} \sum_{i \in s_p} d_{p,i} \hat{y}_i.$$

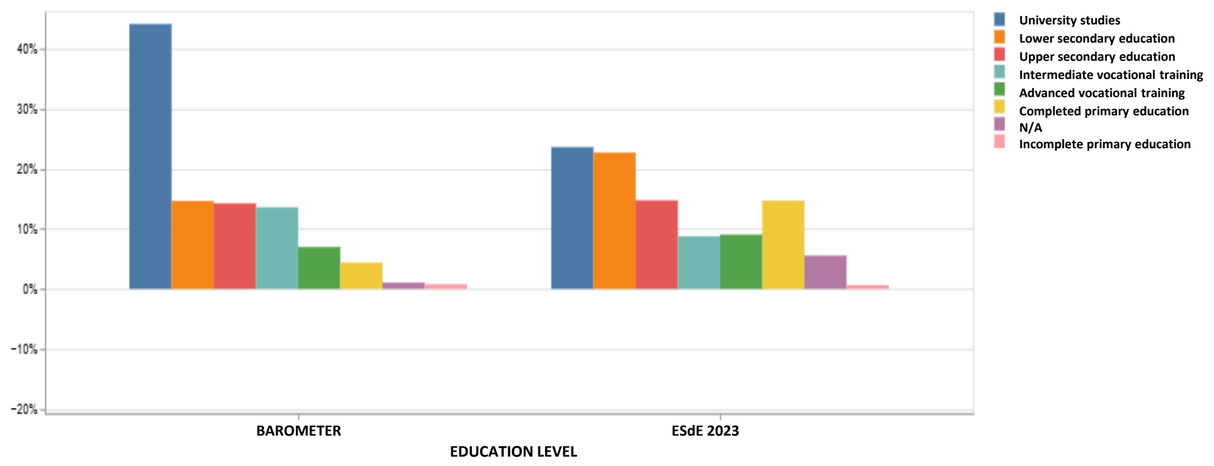
The SM estimations can be improved if the units used to fit the model m are weights using an estimation of d_{np} from the PSA approach (Castro-Martín et. al, 2022).

RESULTS & DISCUSSION

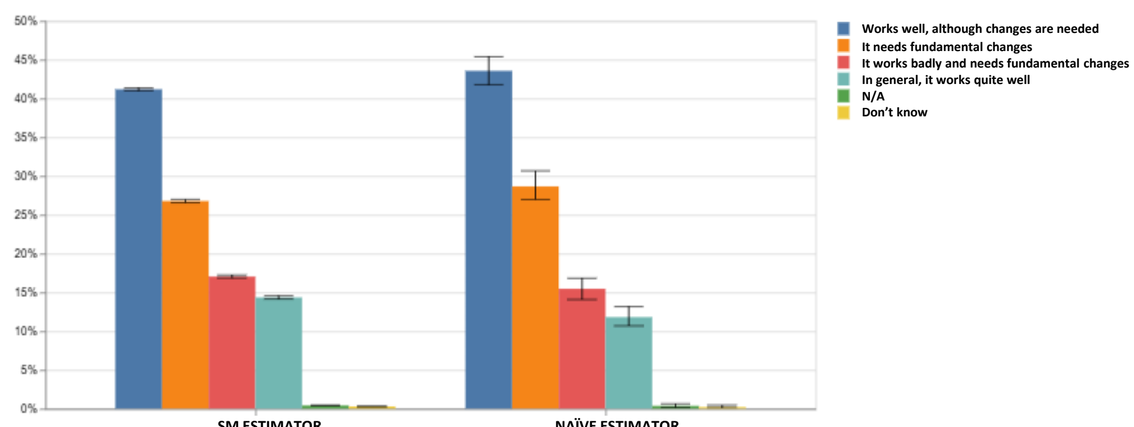
A real-world application is presented applying the SM estimator (analysis done with webpage: https://easp.es/info/bettersurveys/).

- ***Population*:** Spanish population aged ≥18 in 2024.
- s_{np}**:** Health Barometer of April 2024 (CIS 2024), quota sampling by sex and age
- Spanish National Health Survey (ESdE 2023), stratified three-stage sampling ($n_p = 21{,}032$)
- x**:** Sex, Age, Nationality, Education level , Marital status, Employment status, Residents by CCAA, and Religion,…
- y**:** Satisfaction scale with the functioning of the Spanish healthcare system (Parameter: proportion of each response).

Both samples present different values for the covariate vector (s_{np} lacks population representativeness). For education level variable:



We apply the SM estimator using the estimated d_{np}, and we compared its results with the naïve estimator (no bias adjustment):



CONCLUSION

Results show different estimates between SM and the naïve estimator. Although this difference is relatively small, it could be far more severe for other variables, datasets, or estimators.

Estimates from non-probability samples are biased and must be treated with caution, and nowadays, almost all datasets are from this nature

FUTURE WORK / REFERENCES

References: 1. Castro-Martín, L., Rueda, M. d. M. and Ferri-García, R. Combining statistical matching and propensity score adjustment for inferen-ce from non-probability surveys. Journal of Computational and AppliedMathematics, vol. 404, p. 113414, 2022.
2. Rivers, D. Sampling for web surveys. In Joint Statistical Meetings, p. 4.2007.
3. Wu, C. Statistical inference with non-probability survey samples. SurveyMethodology, vol. 48(2), pp. 283–311, 2022.

This study was partially supported by: grant **PID2024-157156OB-I00**, funded by the MCIN/AEI/, Spain; grant **FEDER C-EXP-153-UGR23** and **P24_00565** funded by the Consejeria de Universidad, Investigación e Innovacion and by the ERDF Andalusia Program 2021–2027, Spain; and **FPU grant FPU23/02908** funded by Agencia Estatal de Investigación, Ministerio de Ciencia, Innovación y Universidades, Gobierno de España