Maciej Beręsewicz, PhD
Poznań University of Economics and Business
Department of Statistics
Poland
maciej.beresewicz@ue.poznan.pl

Cover Letter May 24, 2025

Dear Editors of the Journal of Statistical Software,

we are writing to resubmit an revised manuscript (no. 5793) of our article entitled ***nonprobsvy*** – *An R package for modern methods for non-probability surveys* for review to the Journal of Statistical Software. Replies to the comments are provided at the end of this cover letter.

Official statistics traditionally rely on probability surveys, censuses, or administrative registers that cover the entire population. However, rising non-response rates and survey costs have increased interest in non-probability data sources such as opt-in web panels, social media, scanner data, mobile phone data, and voluntary registers. Since these sources lack known selection mechanisms, standard design-based inference methods cannot be directly applied. That is why we have developed the `nonprobsvy` package in the R language (version 0.2.2; available on CRAN).

The `nonprobsvy` package has several advantages over packages currently available in R or Python. In particular, the novelty and our contribution can be summarised as follows:

- the package implements state-of-the-art methods recently proposed in the literature, along with valid statistical inference procedures (i.e. analytical and boostrap variance estimators),
- the package implements various approaches, such as calibrated inverse probability weighting, mass imputation, and doubly robust estimators, with our contributions that extend existing literature,
- the package supports the functions included in the `survey` package to account for the design of the probability sample (if is available).

We provide a user-friendly API that mimics `glm`, `svydesign` and other functions known in R, together with the main function to specify the approach and estimators.

The package has been developed since 2022 and the full history can be found at the GitHub repository `https://github.com/ncn-foreigners/nonprobsvy`. The package has been cited multiple times (cf. `https://scholar.google.com/scholar?q=nonprobsvy`), including in the review paper by *Cobo et al. (2024). Software review for inference with non-probability surveys. The Survey Statistician, 90, 40-47.* Our package is also included in *West et al. (2025). Applied Survey Data Analysis (3rd ed.). Taylor & Francis.*

As far as we know, the `nonprobsvy` is the only software (open-access or commercial) that offers such functionalities. That is why we believe that the paper and software will be of interest to the readership of the Journal of Statistical Software.

Thank you for your consideration of this manuscript.

Sincerely,

Maciej Beręsewicz, Łukasz Chrostowski & Piotr Chlebicki

Replies:

- The article "nonprobsvy – An R package for modern methods for non-probability surveys" presents an R package implementing various methods for inference based on non-probability samples. This is the second submission of this article. Most of the previous remarks have been convincingly addressed by the authors but we found two minor issues that we recommand to fix before this can be sent into review. Please address these issues and resubmit your work as a new submission along with point-by-point answer.
- Reply: We have updated the package and changes include:
  - minor changes to the code, e.g. `control_out(eps=1e-8)`
  - fixing a bug in the bootstrap variance estimator the `method_nn` and `method_pmm`
  - fixing bootstrap for doubly robust estimators
  - more unit-tests for doubly robust estimators and other methods
  - more informative vignette for `method_glm`

- Titles in help(package = "nonprobsvy") are still not in title style. `Propensity score model functions` should be `Propensity Score Model Functions`.
- Reply: Corrected.

- We found slightly different results running the replication script than what is displayed in the article in one of the output:
  ```
  mean SE
  NN 0.6799537 0.01568503
  PMM 0.7337228 0.02231784
  ```
  instead of the result shown page 21
- Reply: This took us some time to understand what is going on. Based on the simulations and we conclude that is actually the issue of the `eps` from the `glm.control` function and the character of our example. In the paper we present data that contains only categorical variables and the bootstrap estimator is based on resampling this data (see my codes and the comments in this issue `https://github.com/jefferislab/RANN/issues/35`).
  Therefore, to avoid this numerical problem where only categorical variables are present in the dataset we decided not to use the bootstrap in this example and provided the following comment in the paper (after the creation of the `mi_est1` object):
  However, in this example we have decided not to use bootstrap as all variables are categorical and this results in a large number of NNs having the same distance and the numerical approximation defined in `control_out(eps)` may give slightly different results between platforms. Therefore the reported variance is based on the probability sample S_B only.