Workshop on the `nonprobsvy` package

dr Maciej Beręsewicz
Department of Statistics, Poznań University of Economics and Business
Centre for the Methodolody of Population Studies, Statistical Office in Poznań

Workshops for Ukraine (08.01.2026)

Poznań University
of Economics
and Business

Contents

1. Introduction
   - About the workshop
   - About the team
   - About funding
   - About the `nonprobsvy` package
   - Selected literature
2. Methods implemented in the package
   - Basic setup
   - Inverse probability weighting
   - Mass imputation
   - Doubly robust estimators
   - Variable selection, variance estimators
3. Case study
4. Exercises

Contents

1. Introduction
    - About the workshop
    - About the team
    - About funding
    - About the `nonprobsvy` package
    - Selected literature

2. Methods implemented in the package

3. Case study

4. Exercises

About the workshop

- Part 1: theory
- Part 2: practice (a)
- Part 3: practice (b)
- Part 4: discussion session
- Materials: `https://github.com/ncn-foreigners/2026-workshop-for-ukraine`

The team behind the nonprobsvy package

dr Maciej Beręsewicz (left)

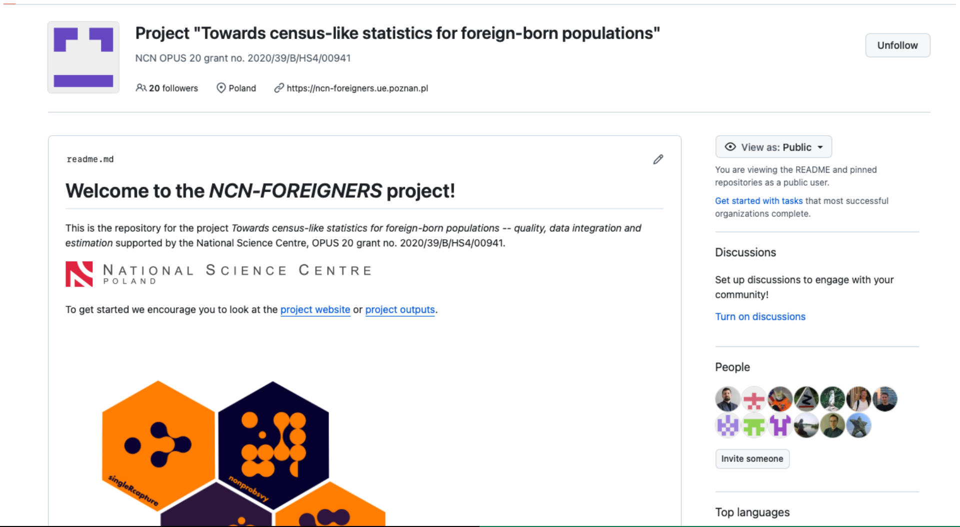
- Department of Statistics, Poznań University of Economics and Business, Poland
- Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland
- Last year Chair of the Baltic-Nordic-Ukrainian Network on Survey Statistics (see `https://wiki.helsinki.fi/xwiki/bin/view/BNU/Home/`)

mgr Łukasz Chrostowski (right) – M.Sc. in Mathematics, Analyx

About funding

- Works on this package was funded by the National Science Centre grant entitled: *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (no. 2020/39/B/HS4/00941).
- The main objective of the project is to develop methods for estimating the size and characteristics of the foreign population in Poland based on available data sources.
- `https://ncn-foreigners.ue.poznan.pl`

About funding – github



Project "Towards census-like statistics for foreign-born populations"

NCN OPUS 20 grant no. 2020/39/B/HS4/00941

20 followers Poland https://ncn-foreigners.ue.poznan.pl

readme.md

Welcome to the *NCN-FOREIGNERS* project!

This is the repository for the project *Towards census-like statistics for foreign-born populations -- quality, data integration and estimation* supported by the National Science Centre, OPUS 20 grant no. 2020/39/B/HS4/00941.

NATIONAL SCIENCE CENTRE POLAND

To get started we encourage you to look at the project website or project outputs.

You are viewing the README and pinned repositories as a public user.

Get started with tasks that most successful organizations complete.

Discussions

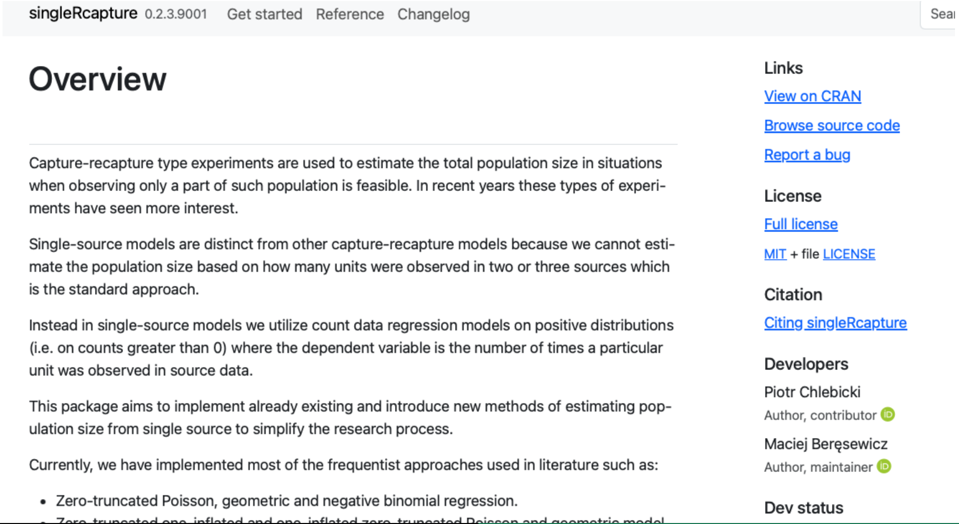
Set up discussions to engage with your community!

Turn on discussions

People

Top languages

Single-source capture-recapture models



singleRcapture 0.2.3.9001 Get started Reference Changelog Sea

Overview

Capture-recapture type experiments are used to estimate the total population size in situations when observing only a part of such population is feasible. In recent years these types of experiments have seen more interest.

Single-source models are distinct from other capture-recapture models because we cannot estimate the population size based on how many units were observed in two or three sources which is the standard approach.

Instead in single-source models we utilize count data regression models on positive distributions (i.e. on counts greater than 0) where the dependent variable is the number of times a particular unit was observed in source data.

This package aims to implement already existing and introduce new methods of estimating population size from single source to simplify the research process.

Currently, we have implemented most of the frequentist approaches used in literature such as:

- Zero-truncated Poisson, geometric and negative binomial regression.
- Zero-truncated one-inflated and one-inflated zero-truncated Poisson and geometric model

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Citation

Citing singleRcapture

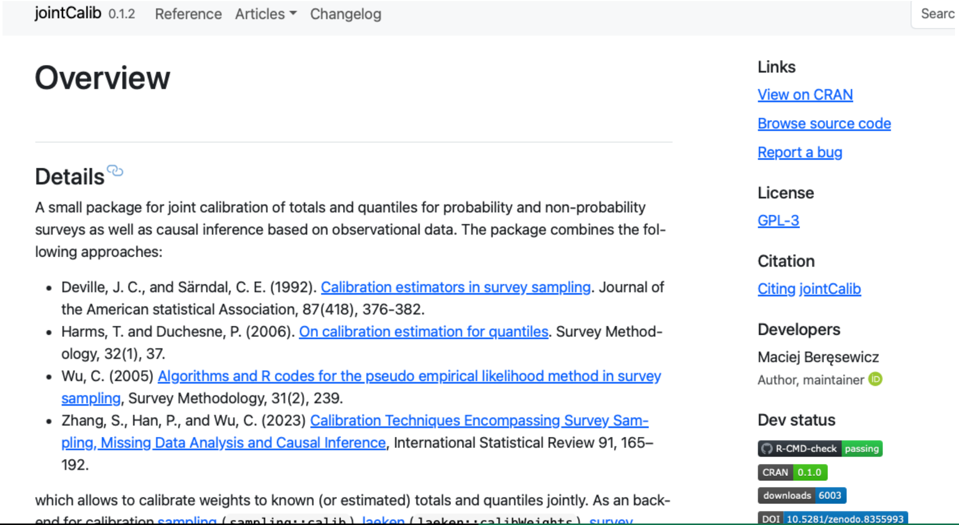
Developers

Piotr Chlebicki
Author, contributor

Maciej Beręsewicz
Author, maintainer

Dev status

Calibration for quantiles (and totals)



Overview

Details

A small package for joint calibration of totals and quantiles for probability and non-probability surveys as well as causal inference based on observational data. The package combines the following approaches:

- Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.
- Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. Survey Methodology, 32(1), 37.
- Wu, C. (2005) Algorithms and R codes for the pseudo empirical likelihood method in survey sampling, Survey Methodology, 31(2), 239.
- Zhang, S., Han, P., and Wu, C. (2023) Calibration Techniques Encompassing Survey Sampling, Missing Data Analysis and Causal Inference, International Statistical Review 91, 165–192.

which allows to calibrate weights to known (or estimated) totals and quantiles jointly. As an backend for calibration sampling (`sampling::calib`), laeken (`laeken::calibWeights`), survey

Links

View on CRAN

Browse source code

Report a bug

License

GPL-3

Citation

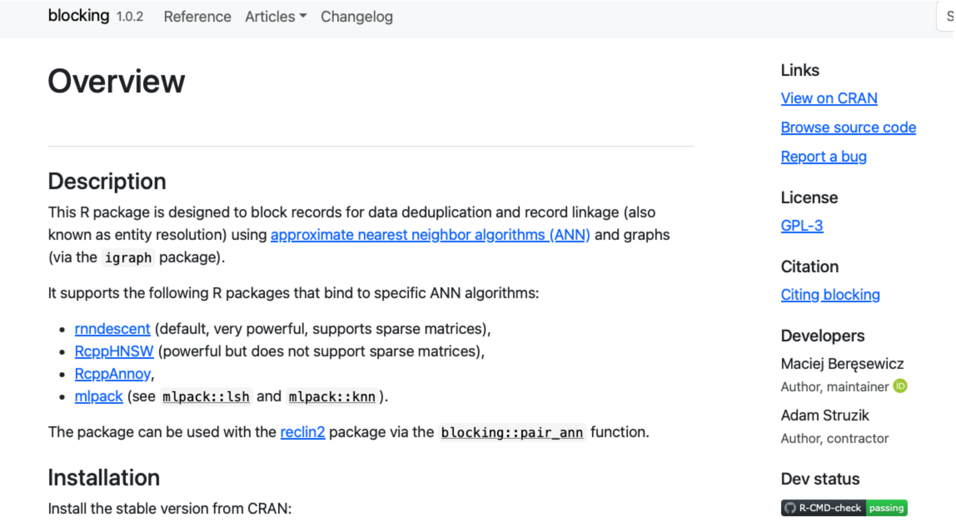
Citing jointCalib

Developers

Maciej Beręsewicz
Author, maintainer

Dev status

Blocking using ANN and graphs



Overview

Description

This R package is designed to block records for data deduplication and record linkage (also known as entity resolution) using approximate nearest neighbor algorithms (ANN) and graphs (via the `igraph` package).

It supports the following R packages that bind to specific ANN algorithms:

- rnndescent (default, very powerful, supports sparse matrices),
- RcppHNSW (powerful but does not support sparse matrices),
- RcppAnnoy,
- mlpack (see `mlpack::lsh` and `mlpack::knn`).

The package can be used with the reclin2 package via the `blocking::pair_ann` function.

Installation

Install the stable version from CRAN:

Links

View on CRAN

Browse source code

Report a bug

License

GPL-3

Citation

Citing blocking

Developers

Maciej Beręsewicz
Author, maintainer

Adam Struzik
Author, contractor

Dev status

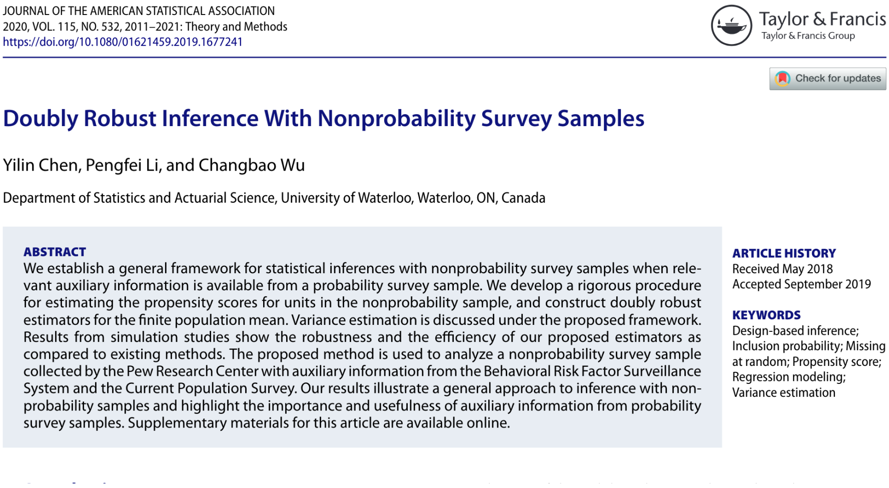
Why another package for non-probability samples?

- It should be noted that there are some packages that can be used for non-probability samples, such as `NonProbEst`, `WeightIt` or `GJRM`.
- These packages are limited in terms of the approaches that can be employed and the variance estimation that can be carried out.
- None of these packages are integrated with the `survey` package.
- There is a lack of implementation of the current solutions presented in the literature.
- The motivation for this project is as follows: The development of a tool that enables the consistent application of different estimation techniques.
- The package is on CRAN and under development so we encourage testing and commenting and tracking on github.

What is unique about the `nonprobsvy` package?

- It provides an easy to use `nonprob` function that mimics existing R functions (e.g. uses formulas).
- It has a full integration with the `survey` package.
- It implements both analytical and bootstrap variance estimators recently proposed in the literature.
- It extends state-of-the-art methods in various ways.

JASA



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 2011–2021: Theory and Methods
https://doi.org/10.1080/01621459.2019.1677241

Doubly Robust Inference With Nonprobability Survey Samples

Yilin Chen, Pengfei Li, and Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

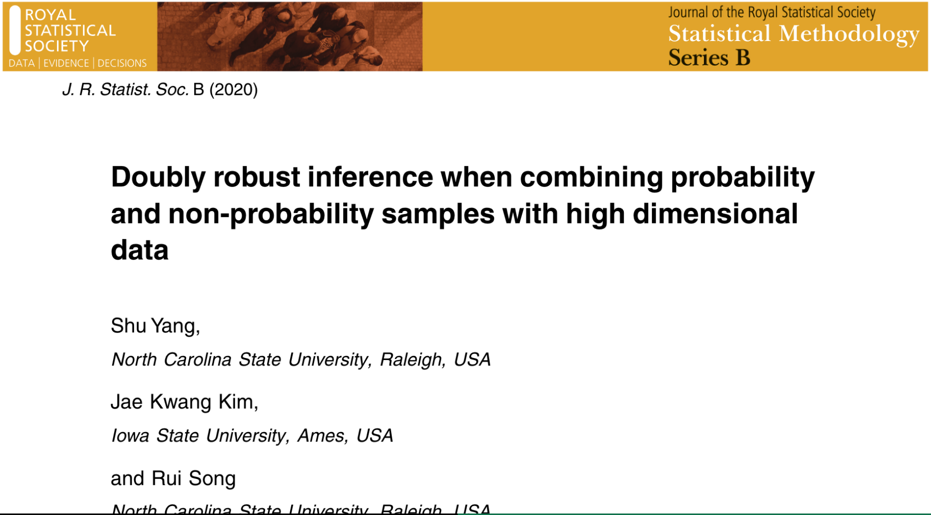
ABSTRACT

We establish a general framework for statistical inferences with nonprobability survey samples when relevant auxiliary information is available from a probability survey sample. We develop a rigorous procedure for estimating the propensity scores for units in the nonprobability sample, and construct doubly robust estimators for the finite population mean. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and the efficiency of our proposed estimators as compared to existing methods. The proposed method is used to analyze a nonprobability survey sample collected by the Pew Research Center with auxiliary information from the Behavioral Risk Factor Surveillance System and the Current Population Survey. Our results illustrate a general approach to inference with nonprobability samples and highlight the importance and usefulness of auxiliary information from probability survey samples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2018
Accepted September 2019

KEYWORDS
Design-based inference; Inclusion probability; Missing at random; Propensity score; Regression modeling; Variance estimation

JRSSB



Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)

Doubly robust inference when combining probability and non-probability samples with high dimensional data

Shu Yang,

North Carolina State University, Raleigh, USA

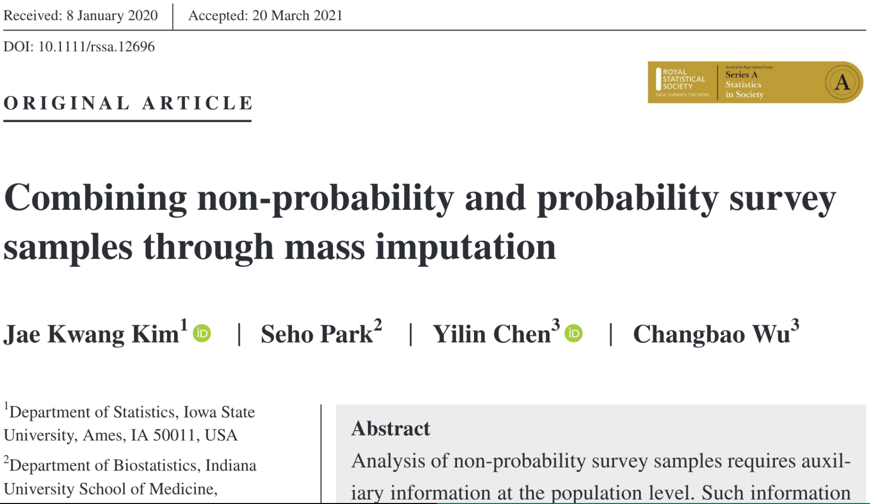
Jae Kwang Kim,

Iowa State University, Ames, USA

and Rui Song

North Carolina State University, Raleigh, USA

JRSSA



Received: 8 January 2020 | Accepted: 20 March 2021

DOI: 10.1111/rssa.12696

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Series A
Statistics
in Society

ORIGINAL ARTICLE

Combining non-probability and probability survey samples through mass imputation

Jae Kwang Kim[1] | Seho Park[2] | Yilin Chen[3] | Changbao Wu[3]

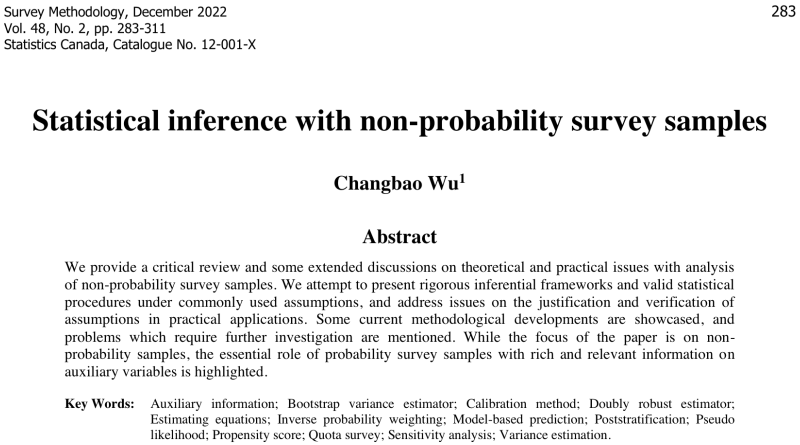
[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[2]Department of Biostatistics, Indiana University School of Medicine,

Abstract

Analysis of non-probability survey samples requires auxiliary information at the population level. Such information

Survey Methodology 1



Survey Methodology, December 2022
Vol. 48, No. 2, pp. 283-311
Statistics Canada, Catalogue No. 12-001-X

Statistical inference with non-probability survey samples

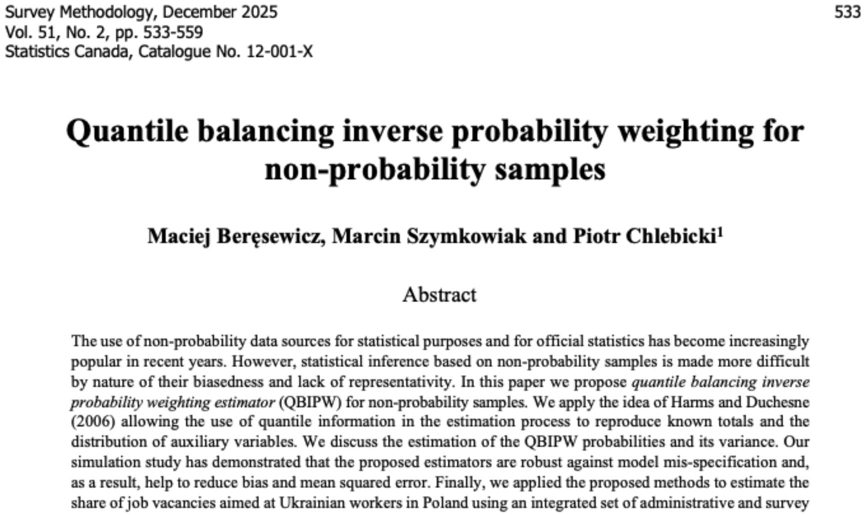
Changbao Wu[1]

Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

Key Words: Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

Survey Methodology 2



Survey Methodology, December 2025
Vol. 51, No. 2, pp. 533-559
Statistics Canada, Catalogue No. 12-001-X

533

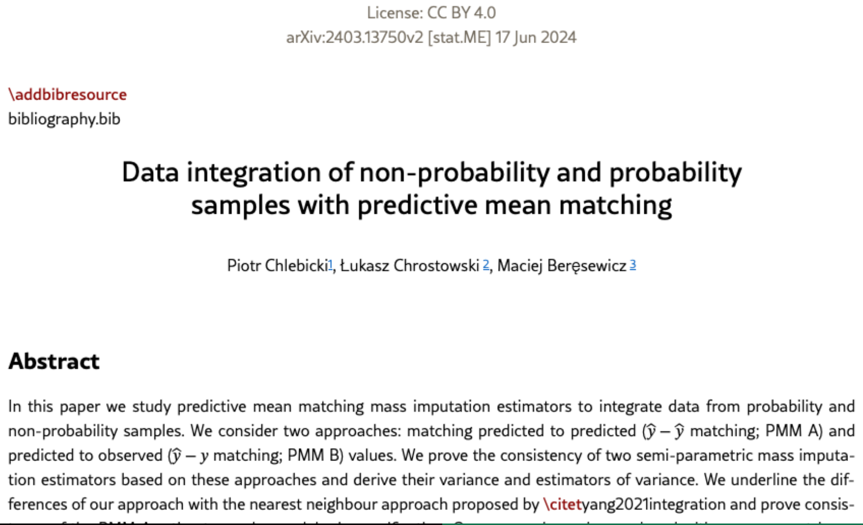
Quantile balancing inverse probability weighting for non-probability samples

Maciej Beręsewicz, Marcin Szymkowiak and Piotr Chlebicki[1]

Abstract

The use of non-probability data sources for statistical purposes and for official statistics has become increasingly popular in recent years. However, statistical inference based on non-probability samples is made more difficult by nature of their biasedness and lack of representativity. In this paper we propose *quantile balancing inverse probability weighting estimator* (QBIPW) for non-probability samples. We apply the idea of Harms and Duchesne (2006) allowing the use of quantile information in the estimation process to reproduce known totals and the distribution of auxiliary variables. We discuss the estimation of the QBIPW probabilities and its variance. Our simulation study has demonstrated that the proposed estimators are robust against model mis-specification and, as a result, help to reduce bias and mean squared error. Finally, we applied the proposed methods to estimate the share of job vacancies aimed at Ukrainian workers in Poland using an integrated set of administrative and survey

Working papers



\addbibresource
bibliography.bib

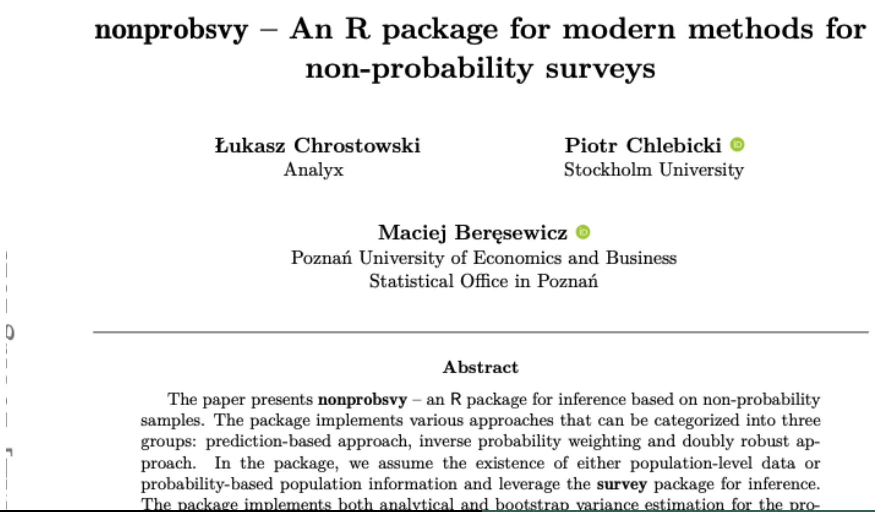
Data integration of non-probability and probability samples with predictive mean matching

Piotr Chlebicki[1], Łukasz Chrostowski [2], Maciej Beręsewicz [3]

Abstract

In this paper we study predictive mean matching mass imputation estimators to integrate data from probability and non-probability samples. We consider two approaches: matching predicted to predicted ($\hat{y} - \hat{y}$ matching; PMM A) and predicted to observed ($\hat{y} - y$ matching; PMM B) values. We prove the consistency of two semi-parametric mass imputation estimators based on these approaches and derive their variance and estimators of variance. We underline the differences of our approach with the nearest neighbour approach proposed by \citetyang2021integration and prove consis-

the nonprobsvy paper



nonprobsvy – An R package for modern methods for non-probability surveys

Łukasz Chrostowski
Analyx

Piotr Chlebicki
Stockholm University

Maciej Beręsewicz
Poznań University of Economics and Business
Statistical Office in Poznań

Abstract

The paper presents **nonprobsvy** – an R package for inference based on non-probability samples. The package implements various approaches that can be categorized into three groups: prediction-based approach, inverse probability weighting and doubly robust approach. In the package, we assume the existence of either population-level data or probability-based population information and leverage the **survey** package for inference. The package implements both analytical and bootstrap variance estimation for the pro-

Literature (selected)

- Chen, Yilin, Pengfei Li, and Changbao Wu. 2020. "Doubly Robust Inference With Nonprobability Survey Samples." Journal of the American Statistical Association 115 (532): 2011–21. https://doi.org/10.1080/01621459.2019.1677241.
- Kim, Jae Kwang, Seho Park, Yilin Chen, and Changbao Wu. 2021. "Combining Non-Probability and Probability Survey Samples Through Mass Imputation." Journal of the Royal Statistical Society Series A: Statistics in Society 184 (3): 941–63. https://doi.org/10.1111/rssa.12696
- Wu, Changbao. 2023. "Statistical Inference with Non-Probability Survey Samples." Survey Methodology 48 (2): 283–311. https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm.
- Yang, Shu, Jae Kwang Kim, and Youngdeok Hwang. 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/p
001-x/2021001/article/00004-eng.htm.
- Yang, Shu, Jae Kwang Kim, and Rui Song. 2020. "Doubly Robust Inference When Combining Probability and Non-Probability 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.

What is not (yet) implemented?

- Overlapping samples (in the development phase).
- Using replicated weights from the probability sample (in the development phase).
- Model calibration approach (e.g. using empirical likelihood, LASSO).
- GAM or other non-parametric methods for mass imputation /doubly robust estimators.
- Inference for quantiles (to be developed).
- Situations where target variable (Y) is observed in both sources (non-probability and probability sample, i.e. mixed-mode).
- Pseudo-population bootstrap for variance estimation.

If you have other ideas please let us know!

Contents

1. Introduction
2. Methods implemented in the package
   - Basic setup
   - Inverse probability weighting
   - Mass imputation
   - Doubly robust estimators
   - Variable selection, variance estimators
3. Case study
4. Exercises

Basic setup

The package allows two approaches, assuming unit-level data are available from the non-probability sample:

- only population=level data are available (through some vector of totals, means, and population size),
- survey data is available from the reference probability sample (a `survey::svydesign` object can be specified).

Basic setup

Tabela 1: Two sample setting

Sample	ID	Sample weight $d = \pi^{-1}$	Covariates $\boldsymbol{x}$	Study variable y
Non-probability sample (S_A)	1	?	✓	✓
	⋮	?	⋮	⋮
	n_A	?	✓	✓
Probability sample (S_B)	1	✓	✓	?
	⋮	⋮	⋮	?
	n_B	✓	✓	?

Basic setup

- Let $U = \{1, ..., N\}$ denote the target population consisting of N labelled units.
- Each unit i has an associated vector of auxiliary variables $\boldsymbol{x}_i$ (a realisation of the random vector $\boldsymbol{X}$ in the super-population) and the study variable y_i (a realisation of the random variable Y in the super-population).
- Let $\{(y_i, \boldsymbol{x}_i), i \in S_A\}$ be a dataset of a non-probability sample of size n_A and let $\{(\boldsymbol{x}_i, \pi_i), i \in S_B\}$ be a dataset of a probability sample of size n_B, where only information about variables $\boldsymbol{X}$ and inclusion probabilities π are available.
- Let δ be an indicator of inclusion into non-probability sample. Each unit in the sample S_B has been assigned a design-based weight given by $d_i = 1/\pi_i$.
- Let $\mathbb{P}(R_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i, \theta_0)$ be the probability of inclusion into non-probability sample.
- There is no overlap between the two samples.

Implemented approaches

The package implements the following approaches:

- Inverse probability weighting (with possible calibration constraints).
- Mass imputation estimators (nearest neighbours, prediction via GLM and predictive mean matching, non-parametric).
- Doubly robust estimators.

Inverse probability weighting

- Maximum likelihood approach

$$\ell^*(\theta) = \sum_{i \in S_A} \log\left\{\frac{\pi(\boldsymbol{x}_i, \theta)}{1 - \pi(\boldsymbol{x}_i, \theta)}\right\} + \sum_{i \in S_B} d_i^B \log\{1 - \pi(\boldsymbol{x}_i, \theta)\} \tag{1}$$

- Generalized estimation equations approach

$$\mathbf{U}(\theta) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \theta) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \theta) \mathbf{h}(\mathbf{x}_i, \theta) \tag{2}$$

for $h(\boldsymbol{x}_i) = \boldsymbol{x}_i / \pi(\boldsymbol{x}_i, \theta_0)$ we get

$$\mathbf{U}(\theta) = \sum_{i \in S_A} \boldsymbol{x}_i \pi_i \left(\boldsymbol{x}_i^{\mathrm{T}} \theta\right)^{-1} - \sum_{i \in S_B} d_i^B \boldsymbol{x}_i \tag{3}$$

Inverse probability weighting

With the estimated propensity scores we can consider two approaches for population mean estimation, depending on whether population size is known or not.

$$\begin{aligned} \hat{\mu}_{IPW1} &= \frac{1}{N} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A} \\ \hat{\mu}_{IPW2} &= \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A}, \end{aligned} \tag{4}$$

where $\hat{N}^A = \sum_{i \in S_A} \hat{d}_i^A$.

Mass imputation

The following approaches are implemented:

- Semi-parametric prediction approach (i.e. imputation are some model $m(.)$ predictions),
- Nearest neighbours imputation (i.e. imputation is based on looking for nearest neighbours based on $\boldsymbol{x}$),
- Predictive mean matching imputation (i.e. imputation is based on matching based on $\hat{y}$; two variants, non-parametric approach via loess).

The following estimator is used

$$\hat{\mu}_{MI} = \frac{1}{N} \sum_{i \in B} w_i \widehat{y}_i. \tag{5}$$

Doubly robust estimators

- When the population size is known

$$\hat{\mu}_{\mathrm{DR1}} = \frac{1}{N} \sum_{i \in S_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N} \sum_{i \in S_B} d_i^B \hat{m}_i \qquad (6)$$

- When the population size is unknown

$$\hat{\mu}_{\mathrm{DR2}} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i^B \hat{m}_i, \qquad (7)$$

where $\hat{N}^A = \sum_{i \in S_A} \left(\hat{\pi}_i^A\right)^{-1}$ and $\hat{N}^B = \sum_{i \in S_B} d_i^B$.

Variable selection, variance estimators

- We have implemented variable selection for all proposed estimators using LASSO, SCAD or MCP penalty.
- Variance estimation: analytical and bootstrap approaches (with known or estimated population size).

Contents

1. Introduction

2. Methods implemented in the package

3. Case study

4. Exercises

Case study

- **Data**: We focus on the integration of the Job Vacancy survey with Online/Admin data.
- **Population**: entities with at least one vacancy.
- **Target variable**: whether the entity has at least one vacancy in a one shift.
- **Target quantity**: share of companies that has at least one vacancy on a single shift.
- **Assumptions**:
  - There is an overlap between sources which we ignore.
  - We assume that no measurement errors are present in both sources.
  - Admin population was aligned with the JVS definitions so over-coverage is present at the unit-level.
  - Units in admin data were linked using a business identifier so we assume no linkage errors.

Thank you! Now, we can proceed to the practical part

Contents

1. Introduction
2. Methods implemented in the package
3. Case study
4. Exercises

Thank you! Now, we can proceed to the practical part