Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Quantile balancing inverse probability weighting for non-probability samples

by Maciej Beręsewicz, Marcin Szymkowiak and Piotr Chlebicki

Release date: December 23, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > “Standards of service to the public.”

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

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

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 data about job vacancies.

Key Words: Calibration approach; Data integration; Job vacancy survey.

1. Introduction

In official statistics, information about the target population and its characteristics is mainly collected through probability surveys, census or is obtained from administrative registers, which cover all (or nearly all) units of the population. However, owing to increasing non-response rates, particularly unit non-response and non-contact resulting from the growing respondent burden, as well as rising costs of surveys conducted by National Statistical Institutes (NSIs), non-probability data sources are becoming more popular (Beaumont, 2020; Beręsewicz, 2017). Non-probability surveys, such as opt-in web panels, social media, scanner data, mobile phone data or voluntary register data, are currently being explored for use in the production of official statistics (Citro, 2014; Daas, Puts, Buelens and Hurk, 2015). Since the selection mechanism in these sources is unknown, standard design-based inference methods cannot be directly applied.

To address this problem, several approaches based on inverse probability weighting (IPW), mass imputation (MI) and doubly robust (DR) estimators have been proposed for two main scenarios: 1) population-level data are available, either in the form of unit-level data (e.g. from a register covering the whole population) or known population totals/means, and 2) only survey data are available as a source of information about the target population (cf. Elliott and Valliant, 2017). Wu (2022) classified these approaches into three groups that require a joint randomization framework involving p (probability sampling design) and one of the outcome regression model ξ or propensity score model q. In this approach the IPW estimator

1. Maciej Beręsewicz, Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland and Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland. E-mail: maciej.beresewicz@ue.poznan.pl; Marcin Szymkowiak, Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland and Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland; Piotr Chlebicki, Stockholm University, Department of Mathematics, Albano hus 1, 106 91 Stockholm.

is under the qp framework, the MI estimator is under the ξp framework, DR is under the qp or ξp framework.

Most approaches assume that population data are used to reduce the bias of non-probability sampling by a proper reweighting to reproduce known population totals/means, by modelling $E(Y \mid \mathbf{X})$ using various techniques or combining both approaches (for instance doubly robust estimators, cf. Chen, Li and Wu (2020); multilevel regression and post-stratification (MRP) also called *Mister-P*, cf. Gelman and Little (1997)). Majority of these methods rely on a limited number of moments of continuous or count data, with some exceptions. For example, non-parametric approaches based on nearest neighbours (NN), such as those discussed by Yang, Kim and Hwang (2021) or kernel density estimation (KDE) described by Chen, Yang and Kim (2022) have also been proposed. In general, the standard approach, such as IPW, seems to be limited, although surveys and register data contain continuous or count data which can be used to account for distribution mismatches. This is because the standard IPW estimators do not allow to match quartiles, deciles or whole distribution between samples.

In this paper, we focus on the qp framework and extend existing IPW estimators to account for distribution differences through quantiles. We use the idea of survey calibration for quantiles described by Harms and Duchesne (2006) and propose a *quantile balancing* IPW (hereinafter QBIPW) estimator that reproduces quantiles and totals jointly for a set of auxiliary variables. Our contribution can be summarised as follows:

- we extend IPW estimator for non-probability samples to account not only for totals/means but also for the distribution of auxiliary variables in the form of a set of quantiles;
- we show that existing inference methods hold for the proposed techniques;
- we provide proofs for the existence of solutions for the proposed techniques;
- we show that adding information on quantiles is equivalent to using piecewise (constant) regression, which can be used to approximate a non-linear relationship through a linear model, making IPW more robust to model mis-specification, and consequently, decreasing the bias and improving efficiency of estimation.

In this paper we generally follow the notation used by Wu (2022). The paper has the following structure. In Section 2 we introduce the basic setup: the notation and a short description of calibration for totals and quantiles separately. In Section 3 we discuss the IPW estimator for the non-probability samples. In Section 4 we present the proposed estimator. Section 5 contains results of a simulation study and finally, Section 6 summarises results from a study in which the proposed approach was applied to combined data from a job vacancy survey and vacancies from administrative (public employment offices) sources. The paper ends with conclusions and identifies research problems that require further investigation.

2. Basic setup

2.1 Notation

Let $U = \{1,...,N\}$ denote the target population consisting of N labelled units. Each unit k has an associated vector of auxiliary variables $\mathbf{x}$ and the target variable y, with their corresponding values $\mathbf{x}_k$ and y_k, respectively. Let $\{(y_k, \mathbf{x}_k), k \in S_A\}$ be a dataset of a non-probability sample S_A of size n_A and $\{(\mathbf{x}_k, d_k^B), k \in S_B\}$ – a dataset of a probability sample S_B of size n_B. Only information about auxiliary variables $\mathbf{x}$ is found in both datasets and $d_k^B = 1/\pi_k^B$ are design-based weights assigned to each unit in the sample S_B. Table 2.1 summarises the setup, which contains no overlap.

Table 2.1
A setup with two data sources

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

Let $R_k = I(k \in S_A)$ be the indicator variable showing that unit k is included in the non-probability sample S_A, which is defined for all units in the population U. Let $\pi_k^A = P(R_k = 1 | \mathbf{x}_k)$ for $k = 1,\ldots,N$ be their propensity scores. For π_k^A we can assume a parametric model $\pi(\mathbf{x}_k; \boldsymbol{\gamma})$ (e.g. logistic regression) with unknown model parameters $\boldsymbol{\gamma}$.

The goal is to estimate a finite population total $\tau_y = \sum_{k \in U} y_k$ or the mean $\bar{\tau}_y = \tau_y / N$ of the variable of interest y. In probability surveys (assuming that y_k are known), the Horvitz-Thompson is the well-known estimator of a finite population total, which is expressed as $\hat{\tau}_{y\pi} = \sum_{k \in S_B} d_k^B y_k$. This estimator is unbiased for τ_y i.e. $E(\hat{\tau}_{y\pi}) = \tau_y$. However, a different approach should be applied for non-probability surveys, as π_k^A are unknown and have to be estimated (considering also that in our setup y_k are unknown in S_B and $\hat{\tau}_{y\pi}$ cannot be used).

In the paper we use the idea of Harms and Duchesne (2006) for quantiles which is an extension of the calibration approach proposed by Deville and Särndal (1992) for totals, thus in the next section we briefly discuss these two methods.

2.2 Calibration estimator for a total

Let $\mathbf{x}$ be a vector of auxiliary variables (benchmark variables) for which $\tau_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ is assumed to be known. The main idea of calibration for probability samples is to look for new calibration weights w_k^B that

are as close as possible to original weights d_k^B and reproduce known population totals $\tau_{\mathbf{x}}$ exactly. In other words, in order to find a new vector of calibration weights $\mathbf{w} = \text{argmin}_{\mathbf{v}} D(\mathbf{d}, \mathbf{v})$ we minimise a distance function $D(\mathbf{d}, \mathbf{v}) = \sum_{k \in S_B} d_k^B \, G\left(\frac{v_k^B}{d_k^B}\right) \to \min$ subject to calibration equations $\sum_{k \in S_B} v_k^B \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$, where $\mathbf{d} = (d_1^B, \ldots, d_{n_B}^B)^T$, $\mathbf{v} = (v_1^B, \ldots, v_{n_B}^B)^T$ and $G(\cdot)$ is a function that must satisfy some regularity conditions. For instance, if $G(x) = \frac{(x-1)^2}{2}$, then using the method of Lagrange multipliers the final calibration weights w_k^B can be expressed as $w_k^B = d_k^B + d_k^B \left(\tau_{\mathbf{x}} - \hat{\tau}_{\mathbf{x}\pi}\right)^T \left(\sum_{j \in S_B} d_j^B \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_k$. The final calibration estimator of a population total τ_y (assuming that y_k would be known in a sample S_B) can be expressed as $\hat{\tau}_{y\mathbf{x}} = \sum_{k \in S_B} w_k y_k$.

2.3 Calibration estimator for a quantile

Harms and Duchesne (2006) considered the estimation of quantiles using the calibration approach in a very similar way to what Deville and Särndal (1992) proposed for a finite population total τ_y. By analogy, in their approach it is not necessary to know values of all auxiliary variables for all units in the population. It is enough to know the corresponding quantiles for the benchmark variables. Below we briefly discuss the problem of finding calibration weights for probability sample S_B.

We want to estimate a quantile $Q_{y,\alpha}$ of order $\alpha \in (0,1)$ of the variable of interest y, which can be expressed as $Q_{y,\alpha} = \inf\left\{t \,\middle|\, F_y(t) \geq \alpha\right\}$, where $F_y(t) = N^{-1} \sum_{k \in U} H(t - y_k)$ and the Heaviside function is given by:

$$H(t - y_k) = \begin{cases} 1, & t \geq y_k, \\ 0, & t < y_k. \end{cases} \tag{2.1}$$

We assume that $\mathbf{Q}_{\mathbf{x},\alpha}$ is a vector of known population quantiles of order α for a vector of auxiliary variables $\mathbf{x}$, where $\alpha \in (0,1)$. Even though we use the same notation for a vector $\mathbf{x}$ its elements in calibration approach for totals and quantiles will be different respectively. In most cases quantiles will be known for continuous auxiliary variables, unlike totals, which will generally be known for categorical variables. It may, however, happen that for a specific auxiliary variable its population total and the corresponding quantile of order α will be known. In order to find a new vector of calibration weights $\mathbf{w}$ that reproduces known population quantiles in a vector $\mathbf{Q}_{\mathbf{x},\alpha}$, an interpolated distribution function estimator of $F_y(t)$ is defined as:

$$\hat{F}_{y,\text{cal}}(t) = \frac{\sum_{k \in S_B} w_k^B H_{y,S_B}(t, y_k)}{\sum_{k \in S_B} w_k^B},$$

where the Heaviside function in formula (2.1) is replaced by the modified function $H_{y,S_B}(t, y_k)$ given by

$$H_{y,S_B}(t, y_k) = \begin{cases} 1, & y_k \leqslant L_{y,S_B}(t), \\ \vartheta_{y,S_B}(t), & y_k = U_{y,S_B}(t), \\ 0, & y_k > U_{y,S_B}(t), \end{cases} \tag{2.2}$$

where appropriate parameters are defined as $L_{y,S_B}(t) = \max\{\{y_k, k \in S_B \mid y_k \leqslant t\} \cup \{-\infty\}\}$, $U_{y,S_B}(t) = \min\{\{y_k, k \in S_B \mid y_k > t\} \cup \{\infty\}\}$ and

$$\vartheta_{y,S_B}(t) = \frac{t - L_{y,S_B}(t)}{U_{y,S_B}(t) - L_{y,S_B}(t)}$$

for $k = 1,\ldots,n_B$, $t \in \mathbb{R}$. From a practical point of view a smooth approximation to the step function, based on the logistic function can be used i.e. $H(x) \approx \frac{1}{2} + \frac{1}{2}\tanh kx = (1 + e^{-2kx})^{-1}$, where a larger value of k corresponds to a sharper transition at $x = 0$.

A calibration estimator of quantile $Q_{y,\alpha}$ of order α for variable y is defined as $\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha)$, where a vector $\mathbf{w} = \text{argmin}_{\mathbf{v}}\, D(\mathbf{d},\mathbf{v})$ is a solution of an optimization problem $D(\mathbf{d},\mathbf{v}) = \sum_{k \in S_B} d_k^B\, G\left(\frac{v_k^B}{d_k^B}\right) \to \min$ subject to the calibration constraints $\sum_{k \in S_B} v_k^B = N$ and $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = \mathbf{Q}_{\mathbf{x},\alpha}$ or equivalently $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha$ for all auxiliary variables x_j from a vector $\mathbf{x}$.

As in the previous case, if $G(x) = \frac{(x-1)^2}{2}$, then, using the method of Lagrange multipliers, the final weights w_k^B can be expressed as $w_k^B = d_k^B + d_k^B\left(\mathbf{T_a} - \sum_{j \in S_B} d_j^B \mathbf{a}_j\right)^T \left(\sum_{j \in S_B} d_j^B \mathbf{a}_j \mathbf{a}_j^T\right)^{-1} \mathbf{a}_k$, where $\mathbf{T_a} = (N, \alpha, \ldots, \alpha)^T$ and the elements of $\mathbf{a}_k$ (the first element is 1) are given by:

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \le L_{x_j,S_B}(Q_{x_j,\alpha}), \\ N^{-1}\vartheta_{x_j,S_B}(Q_{x_j,\alpha}), & x_{kj} = U_{x_j,S_B}(Q_{x_j,\alpha}), \\ 0, & x_{kj} > U_{x_j,S_B}(Q_{x_j,\alpha}). \end{cases} \tag{2.3}$$

In the method described above it is assumed that a known population quantile is reproduced for a set of auxiliary variables, i.e. that the process of calibration is based on a particular quantile (of order α). For instance, it could be the median $\alpha = 0.5$. It is a straightforward process to find calibration weights that accurately reflect population quantiles (for instance, quartiles) for a selected set of auxiliary variables.

3. Inverse probability weighting estimator

3.1 Assumptions

Throughout the paper we follow standard assumptions for inference based on non-probability samples under the qp framework, i.e:

(A1) conditional independence of R_k and y_k given $\mathbf{x}_k$;

(A2) all units in the target population have non-zero propensity scores $\pi_k^A > 0$;

(A3) $R_1,\ldots,R_N$ are independent given auxiliary variables $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$.

To estimate the variance of the IPW estimator we follow the same regularity conditions as stated in the Supplementary Materials in the Chen, Li and Wu (2020) paper and Section 3.2 of Tsiatis (2006). For

simplicity we will use $\mathbf{x}_k$ (for which totals are known or estimated) and $\mathbf{a}_k$ (for which specific α -quantiles of selected variables in $\mathbf{x}_k$ are known or estimated) in the definitions of the IPW estimators.

3.2 The IPW estimator

Chen, Li and Wu (2020) discussed inverse probability weighting (IPW) and its asymptotic properties. Let us assume that propensity scores can be modelled parametrically as $\pi_k^A = P(R_k = 1 \mid \mathbf{x}_k) = \pi(\mathbf{x}_k; \boldsymbol{\gamma}_0)$, where $\boldsymbol{\gamma}_0$ is the true value of unknown model parameters. The maximum likelihood estimator of π_k^A is computed as $\hat{\pi}_k^A = \pi(\mathbf{x}_k; \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ maximises the log-likelihood function under full information:

$$\ell(\boldsymbol{\gamma}) = \log\left\{\prod_{k=1}^{N} (\pi_k^A)^{R_k} (1-\pi_k^A)^{1-R_k}\right\} = \sum_{k\in S_A} \log\left(\frac{\pi_k^A}{1-\pi_k^A}\right) + \sum_{k\in U} \log(1-\pi_k^A).$$

However, in practice, reference auxiliary variables $\mathbf{x}$ can be supplied by the probability sample S_B, so $\ell(\boldsymbol{\gamma})$ is replaced with pseudo-likelihood (PL) function:

$$\ell^*(\boldsymbol{\gamma}) = \sum_{k\in S_A} \log\left(\frac{\pi_k^A}{1-\pi_k^A}\right) + \sum_{k\in S_B} d_k^B \log(1-\pi_k^A).$$

The maximum PL estimator $\hat{\boldsymbol{\gamma}}$ can be obtained by solving pseudo-score equations given by $\mathbf{U}(\boldsymbol{\gamma}) = \partial\ell^*(\boldsymbol{\gamma}) / \partial\boldsymbol{\gamma} = \mathbf{0}$. If logistic regression is assumed for π_k^A, then $\mathbf{U}(\boldsymbol{\gamma})$ is given by:

$$\mathbf{U}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \mathbf{x}_k - \sum_{k\in S_B} d_k^B \pi\left(\mathbf{x}_k; \boldsymbol{\gamma}\right) \mathbf{x}_k. \tag{3.1}$$

Note that weights based on $\pi\left(\mathbf{x}_k; \hat{\boldsymbol{\gamma}}\right)$ where $\hat{\boldsymbol{\gamma}}$ are estimated based on (3.1) do not reproduce the population size nor totals for variables in a vector $\mathbf{x}$.

Alternatively, pseudo-score equations $\mathbf{U}(\boldsymbol{\gamma}) = \mathbf{0}$ can be replaced by a system of general estimating equations (Kim and Riddles, 2012; Wu, 2022). Let $\mathbf{h}(\mathbf{x}_k; \boldsymbol{\gamma})$ be a user-specified vector of functions with the same dimensions as $\boldsymbol{\gamma}$. The general estimating equations are given by:

$$\mathbf{G}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \mathbf{h}\left(\mathbf{x}_k; \boldsymbol{\gamma}\right) - \sum_{k\in S_B} d_k^B \pi(\mathbf{x}_k; \boldsymbol{\gamma}) \mathbf{h}(\mathbf{x}_k; \boldsymbol{\gamma}), \tag{3.2}$$

and when we let $\mathbf{h}(\mathbf{x}_k; \boldsymbol{\gamma}) = \mathbf{x}_k / \pi(\mathbf{x}_k; \boldsymbol{\gamma})$, then (3.2) reduces to calibration equations:

$$\mathbf{G}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \frac{\mathbf{x}_k}{\pi\left(\mathbf{x}_k; \boldsymbol{\gamma}\right)} - \sum_{k\in S_B} d_k^B \mathbf{x}_k. \tag{3.3}$$

In this setting unit-level data are not required to get $\sum_{k\in S_B} d_k^B \mathbf{x}_k$. These sums can be replaced with known or estimated population totals from the reference probability survey. Kim and Riddles (2012) showed that estimator (3.3) leads to optimal estimation when a linear regression model holds for y and $\mathbf{x}$.

After estimating $\boldsymbol{\gamma}$ we obtain propensity score $\hat{\pi}_k^A$ and two versions of the IPW estimator:

$$\hat{\bar{\tau}}_{\text{IPW1}} = \frac{1}{N}\sum_{k\in S_A}\frac{y_k}{\hat{\pi}_k^A} \quad \text{and} \quad \hat{\bar{\tau}}_{\text{IPW2}} = \frac{1}{\hat{N}_A}\sum_{k\in S_A}\frac{y_k}{\hat{\pi}_k^A}, \tag{3.4}$$

where $\hat{N}_A = \sum_{k\in S_A}\hat{\pi}_k^A$.

The standard IPW estimator takes only limited information into account and does not preserve the distribution of auxiliary variables. To overcome this limitation we use the idea of the calibration approach for quantiles described in Section 2.3 and modify the IPW estimator to balance distributions via specific quantiles (e.g. quartiles, deciles, etc.).

4. Quantile balancing inverse probability weighting

4.1 Proposed approach

In our proposal we modify the propensity scores π_k^A to take into account $\mathbf{a}_k$ in such way that the function will be given by $\pi_k^A = \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$. The proposed approach can be justified by the fact that the parametric form $\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ is an over-saturated representation that includes $\pi(\mathbf{x}_k; \boldsymbol{\gamma})$ as a special case. In addition, note that we distinguish $\boldsymbol{\gamma}$ from $\boldsymbol{\eta}$ as the latter contains additional parameters referring to $\mathbf{a}_k$. Inclusion of $\mathbf{a}_k$ allows to correct the discrepancies between quantiles of auxiliary variables of S_A with the reference to S_B.

The proposed approach is similar to those used in the literature on survey sampling or causal inference, which offer arguments for the inclusion of higher moments for $\mathbf{x}$ auxiliary variables (Ai, Linton and Zhang, 2020; Imai and Ratkovic, 2014). On the other hand, it is a simpler version of the propensity score that balances distributions in comparison to methods using kernel density (Hazlett, 2020) or numerical integration (Sant'Anna, Song and Xu, 2022).

Our approach is not limited to a specific estimation method of the parameters of the propensity score π_k^A. Thus, we start with the use of the full or pseudo-likelihood function with the proposed modification and then solve it using (3.1), which takes the following form if both $\mathbf{x}_k$ and $\mathbf{a}_k$ are used:

$$\mathbf{U}(\boldsymbol{\gamma}) = \begin{pmatrix} \sum_{k\in S_A}\mathbf{x}_k - \sum_{k\in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})\,\mathbf{x}_k \\ \sum_{k\in S_A}\mathbf{a}_k - \sum_{k\in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})\,\mathbf{a}_k \end{pmatrix}. \tag{4.1}$$

Note that this procedure does not allow to reproduce quantiles for the selected $\mathbf{x}$ variables but it reduces the discrepancies between distributions of S_A and S_B. In the simulation study we show that this is particularly significant for the non-linear selection models. On the other hand, including more variables may lead to instability of the estimation procedure and large weights.

To overcome the limitation of maximum likelihood estimation (MLE) we can use the generalised estimating equations in (3.2) which allow to reproduce totals and quantiles from the probability sample S_B. In this situation $\mathbf{G}(\boldsymbol{\eta})$ changes to:

$$\mathbf{G}(\boldsymbol{\eta}) = \sum_{k \in S_A} \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}) - \sum_{k \in S_B} d_k^B \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}) \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}).$$

If we assume logistic regression for $\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ and $\mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta}) = \binom{\mathbf{x}_k}{\mathbf{a}_k} \Big/ \pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ then $\mathbf{G}(\boldsymbol{\eta})$ reduces to:

$$\mathbf{G}(\boldsymbol{\gamma}) = \begin{pmatrix} \sum_{k \in S_A} \frac{\mathbf{x}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{x}_k \\ \sum_{k \in S_A} \frac{\mathbf{a}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{a}_k \end{pmatrix}. \tag{4.2}$$

This allows to estimate the $\boldsymbol{\eta}$ parameters so that the population size N, known population totals (or estimated population totals) for some auxiliary variables in a vector $\mathbf{x}$ or/and known population quantiles for other variables from this vector are all being reproduced. In the Section A in the Appendix we state conditions which guarantee existence and uniqueness of solutions for $\mathbf{U}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathbf{G}(\boldsymbol{\eta}) = \mathbf{0}$.

Remark 1. *Harms and Duchesne (2006, page 41) note that if the relationship between y and a scalar auxiliary variable x is exactly linear then calibration to an α-quantile of the auxiliary variable yields the exact population α-quantile of the target variable y.*

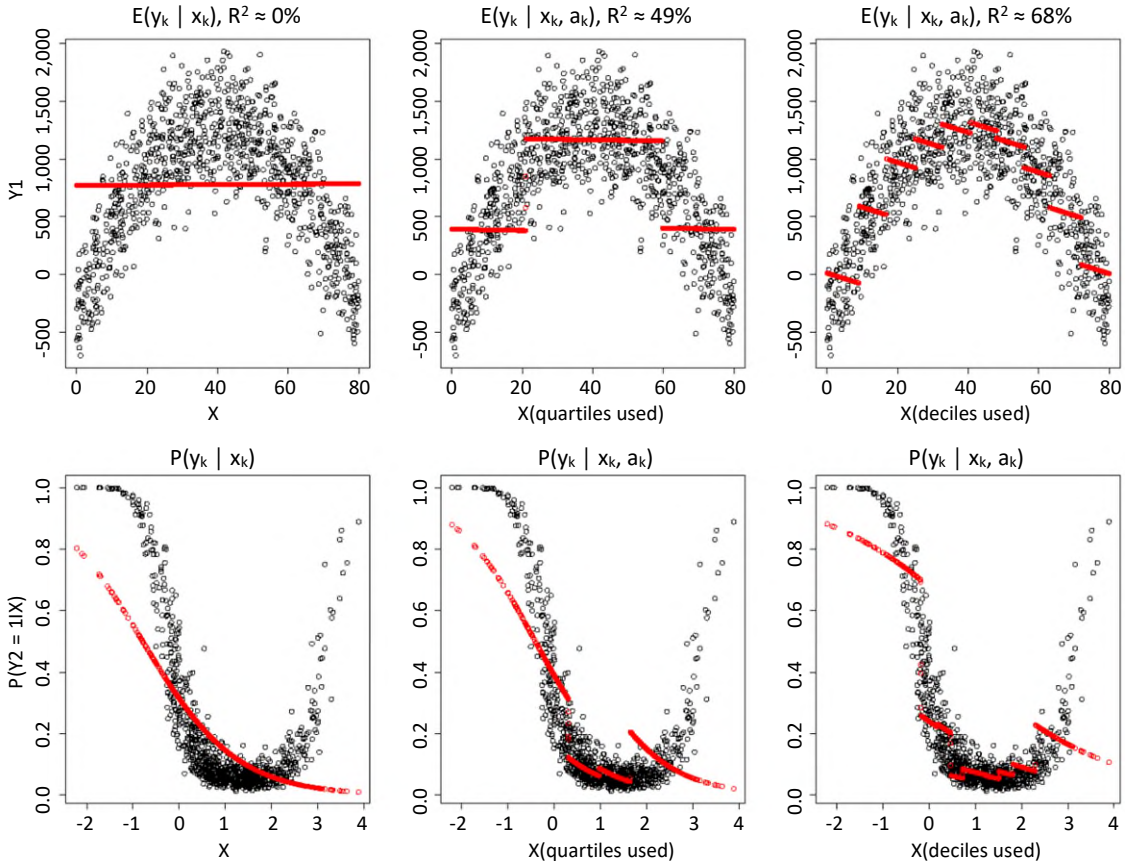
Remark 2. *The use of $\boldsymbol{a}_k$ in (4.2) is related to the use of piecewise (constant) regression where variables $\boldsymbol{x}_k$ are split into breaks (e.g. by quartiles or deciles) and then $\boldsymbol{a}_k$ are included in linear regression as given below*

$$\hat{y}_k = \underbrace{\boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}}_{\text{linear part}} + \underbrace{\boldsymbol{a}_k^T \hat{\boldsymbol{\beta}}^*}_{\text{piecewise part}} .$$

Thus, we use the piecewise method to approximate the non-linear relationship between y and $\boldsymbol{x}$. Consider the following example: generate $n = 1{,}000$ observations of $X \sim$ Uniform(0, 80), $Y_1 \sim N(1{,}300 - (X - 40)^2, 300)$ and $p = P(Y_2 = 1) = \text{logit}(-3 + (X - 1.5)^2 + N(0, 0.5))$ to yield (y_k, p_k, x_k) triplet for $k = 1,\ldots,n$. Then, let x_k be an auxiliary variable in the following settings: 1) x_k is used as it is and 2) x_k is used along with a_k specified for quartiles, and 3) x_k is used along with a_k specified for deciles. Figure 4.1 presents how the prediction of Y_1 (first row) and $P(Y_2 = 1)$ (second row) changes when x_k or x_k and a_k are used. For Y_1 we used linear regression and for Y_2 probabilities we modelled using logistic regression. As expected including quantiles through $\boldsymbol{a}_k$ improves prediction of both models.

Remark 3. *The QBIPW estimator obtained by solving (4.2) is actually doubly robust i.e. it is asymptotically unbiased if either the inverse probability model or outcome model is correct. Usage of quantiles through $\boldsymbol{a}_k$ allows to approximate non-linear relationships in both models. In the simulation study we will show that the proposed approach significantly reduces bias when the two models are non-linear.*

Figure 4.1 Prediction of Y_1 (top row) and $P(Y_2 = 1)$ (bottom row) based on x_k or x_k and a_k (quartiles or deciles) from the example in Remark 2



Note: Black circles denote observed y_k, red circles denote predictions $\hat{m}_k$ based on the linear model and $\hat{p}_k$ based on logistic regression.

After estimation of the $\boldsymbol{\eta}$ parameters we can estimate the population mean using either $\hat{\bar{\tau}}_{\text{IPW1}}$ or $\hat{\bar{\tau}}_{\text{IPW2}}$ given in equation (3.4) which we denote as $\hat{\bar{\tau}}_{\text{QBIPW1}}$ or $\hat{\bar{\tau}}_{\text{QBIPW2}}$ respectively. To summarize our proposed approach:

- estimated propensity scores will reproduce either known or estimated moments as well as specified α-quantiles (e.g. median, quartiles or deciles, etc.). We use more information regarding count or continuous variables if they are available in both datasets;
- inclusion of α-quantiles in the π_k^A approximates the relationship between inclusion R and $\mathbf{x}$. Thus, the inclusion of α-quantiles makes estimates more robust to model mis-specification for the propensity and outcome model when calibrated IPW is used.

Inference based on the IPW estimator does not change as we add new variables to the propensity score model. We do not assume that quantiles are based on, let's say, instrumental variables or are collected through paradata (Park, Kim and Kim, 2019) but treat them as if we are adding higher moments. Estimation of the the variance of the QBIPW estimators is provided in the next section.

4.2 Variance estimation

In order to estimate the variance of the QBIPW estimators one can modify the equation $\mathbf{\Phi}_n(\boldsymbol{\eta}) = \mathbf{0}$ (Wu, 2022, equation 6.1) by imposing constraints on quantiles. In the case when both $\mathbf{x}_k$ and $\mathbf{a}_k$ are used and $\mathbf{h}(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})$ is defined as previously, the equation is given as:

$$\mathbf{\Phi}_n(\boldsymbol{\eta}) = N^{-1} \begin{pmatrix} \sum_{k \in U} R_k \dfrac{y_k - \bar{\tau}_y}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} \\ \sum_{k \in U} R_k \dfrac{\mathbf{x}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{x}_k \\ \sum_{k \in U} R_k \dfrac{\mathbf{a}_k}{\pi(\mathbf{x}_k, \mathbf{a}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \mathbf{a}_k \end{pmatrix}. \tag{4.3}$$

The asymptotic variance for the QBIPW estimator can be derived from the standard sandwich form given by:

$$\mathrm{AV}(\hat{\boldsymbol{\eta}}) = \left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}\right]^{-1} \mathrm{Var}\{\mathbf{\Phi}_n(\boldsymbol{\eta}_0)\} \left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}'\right]^{-1},$$

where $\boldsymbol{\eta}_0$ is the true value of unknown model parameters and $\boldsymbol{\phi}_n(\boldsymbol{\eta}) = \partial \mathbf{\Phi}_n(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ depends on the form of π_k^A and $\mathbf{h}$. Chen, Li and Wu (2020) provided the analytical form of the variance estimator when logistic regression is used for π_k.

Our approach, based on $\mathbf{a}_k$ or $\mathbf{x}_k$, may lead to higher variance as the number of variables can be significantly larger than $\mathbf{x}_k$ alone as adding new variables may inflate the variance of $\pi_k^A(1-\pi_k^A)$.

5. Simulation study

Our simulation study to compare the proposed estimators follows the procedure described in Kim and Wang (2019) and Yang, Kim and Hwang (2021). We generate a finite population $\mathcal{F}_N = \{\mathbf{x}_k = (x_{k1}, x_{k2}), \mathbf{y}_k = (y_{k1}, y_{k2}) : k = 1, \ldots, N\}$ with size $N =$ 100,000, where y_{k1} is the continuous outcome, y_{k2} is the binary outcome, $x_{k1} \sim N(1,1)$ and $x_{k2} \sim \mathrm{Exp}(1)$.

The finite population target variables (Y_1 and Y_2) and the big sample S_A were generated using the outcome variables (denoted as outcome model; OM) and the inclusion probability (denoted as probability model; PM) presented in Table 5.1, where $\alpha_k \sim N(0,1)$, $\epsilon_k \sim N(0,1)$ and x_{k1}, x_{k2}, α_k and ϵ_k are mutually independent. The variable α_k induces dependence of y_{k1} and y_{k2} even after adjusting for x_{k1} and x_{k2}.

From this population, we select a big sample S_A of size approximately 70,000 (for PM1) and 50,000 (for PM2), depending on the selection mechanism, assuming the inclusion indicator $\delta_{Bk} \sim \mathrm{Bernoulli}(p_k)$ with p_k denoting the inclusion probability for unit k. Then we select a simple random sample S_B of size $n =$ 1,000 from $\mathcal{F}_N$. The goal is to estimate the population mean $\bar{\tau}_j = N^{-1} \sum_{k=1}^{N} y_{kj}$, $j = 1, 2$.

Table 5.1
Outcome models and the probability of inclusion in sample S_A used in the simulation study

Type	Form	Formulae
Continuous	linear (OM1)	$y_{k1} = 1 + x_{k1} + x_{k2} + \alpha_k + \epsilon_k$
	non-linear (OM2)	$y_{k2} = 0.5(x_{k1} - 1.5)^2 + x_{k2}^2 + \alpha_k + \epsilon_k$
Binary	linear (OM3)	$P(y_{k1} = 1 \mid x_{k1}, x_{k2}; \alpha_k) = \text{logit}(1 + x_{k1} + x_{k2} + \alpha_k)$
	non-linear (OM4)	$P(y_{k2} = 1 \mid x_{k1}, x_{k2}; \alpha_k) = \text{logit}\left\{0.5(x_{k1} - 1.5)^2 + x_{k2}^2 + \alpha_k\right\}$
Selection (S_A)	linear (PM1)	$\text{logit}(p_k) = x_{k2}$
	non-linear (PM2)	$\text{logit}(p_k) = -3 + (x_{k1} - 1.5)^2 + (x_{k2} - 2)^2$

Note: OM = outcome model; PM = probability model.

In the results, we report Monte Carlo bias (B), standard error (SE) and root mean square error (RMSE) based on $R = 500$ simulations for each y variable:

$$\text{B} = \bar{\hat{\tau}} - \bar{\tau},\ \text{SE} = \sqrt{\frac{\sum_{r=1}^{R} (\hat{\tau}^{(r)} - \bar{\hat{\tau}})^2}{R-1}} \quad \text{and} \quad \text{RMSE} = \sqrt{\text{B}^2 + \text{SE}^2}, \quad \text{where} \quad \bar{\hat{\tau}} = \sum_{r=1}^{R} \hat{\tau}^{(r)} / R \quad \text{and} \quad \hat{\tau}^{(r)}$$

is an estimate of the mean in the r-th replication. The simulation was conducted in R (R Core Team, 2023) using the `nonprobsvy` (Chrostowski and Beręsewicz, 2024) and the `jointCalib` (Beręsewicz, 2023) packages. To solve calibration constraints in the QBIPW estimator we used Newton's method with a trust region global strategy *double deglog* proposed by Dennis Jr and Schnabel (1996) and implemented in the `nleqslv` (Hasselman, 2023) package.

The simulation results are presented in Tables 5.2 and 5.3. The latter shows the empirical coverage (CR) of 95% confidence intervals (CI) based on analytical variance estimators for the IPW estimators. The CR is defined as $\frac{1}{R}\sum_{r=1}^{R} I\left(\bar{\tau} \in \text{CI}^{(r)}\right) \times 100$, where $\text{CI}^{(r)}$ is the CI calculated in the r-th iteration, and we expect it to be close to the nominal 95% level. Table B.1 in the appendix also provides information on the quality of the sums and quantiles reproduced.

The following estimators were considered with a set of calibration equations including totals or quantiles for auxiliary variables estimated from sample S_B:

- Naïve calculated from sample S_A only,
- Mass imputation nearest neighbour with 5 neighbours using x_1, x_2 only (NN),
- Mass imputation based on linear regression using x_1, x_2 only (GLM),
- Doubly robust estimator using x_1, x_2 only with two variants of IPW: MLE (DR MLE; where IPW MLE is based on (3.1)) and GEE (calibrated; DR GEE; where IPW GEE is based on (3.3)),
- IPW MLE and IPW GEE with the following variants: estimated totals for x_1, x_2 only,
- QBIPW MLE and QBIPW GEE based on (4.1) and (4.2) respectively with the following variants:
  - estimated quartiles and totals for x_1, x_2 (QBIPW1),
  - estimated deciles and totals for x_1, x_2 (QBIPW2).

Table 5.2 contains the results for four scenarios, in which combinations of OM and PM for continuous and binary outcome models are considered, as defined in Table 5.1. In Scenario I, for continuous OM, the majority of estimators demonstrate satisfactory performance, exhibiting minimal bias. The proposed estimators, which employ quantiles and totals, demonstrate comparable root mean square error (RMSE) to that of the nearest neighbour (NN), generalised linear model (GLM) or double robust (DR) estimators. As anticipated, the QBIPW with GEE is distinguished by a lower standard error compared to MLE. For binary outcomes, the performance of the QBIPW estimators, particularly the GEE-based ones, is even more pronounced. The RMSE is comparable to that of GLM and DR estimators. Furthermore, the proposed QBIPW under GEE demonstrates an improvement in RMSE compared to the IPW GEE estimator.

Table 5.2
Results for continuous and binary Y (B, SE and RMSE are multiplied by 100)

OM PM	Scenario I linear (1) linear (1)			Scenario II linear (1) non-linear (2)			Scenario III non-linear (2) linear (1)			Scenario IV non-linear (2) non-linear (2)		
	B	SE	RMSE	B	SE	RMSE	B	SE	RMSE	B	SE	RMSE
Continuous Y (OM1 and OM2)												
Naive	18.71	0.40	18.70	-41.84	0.50	41.80	60.67	0.60	60.70	31.11	0.80	31.10
NN	0.20	6.10	6.10	0.66	6.30	6.40	0.43	15.20	15.20	0.90	14.90	14.90
GLM	0.43	4.50	4.60	-0.19	4.50	4.50	-19.92	13.60	24.10	110.72	15.00	111.70
DR (MLE)	0.32	4.60	4.60	0.01	4.60	4.60	1.72	5.80	6.00	227.42	77.00	240.10
DR (GEE)	0.33	4.60	4.60	-0.19	4.50	4.50	0.96	7.10	7.10	123.79	18.10	125.10
Inverse probability weighting (MLE)												
IPW	-0.01	5.70	5.70	69.73	27.50	75.00	0.42	9.30	9.40	467.97	175.10	499.60
QBIPW1	-0.02	5.70	5.70	6.58	11.70	13.50	0.44	9.60	9.60	29.93	22.20	37.30
QBIPW2	-0.50	5.80	5.80	3.12	11.90	12.30	0.73	12.70	12.70	4.04	21.50	21.80
Inverse probability weighting (calibrated; GEE)												
IPW	0.33	4.60	4.60	-0.19	4.50	4.50	0.96	7.10	7.10	123.79	18.10	125.10
QBIPW1	0.33	4.50	4.50	0.38	4.60	4.60	2.61	9.70	10.00	27.84	12.50	30.50
QBIPW2	0.97	4.20	4.30	0.11	4.50	4.50	4.78	10.70	11.70	8.24	12.50	14.90
Binary Y (OM3 and OM4)												
Naive	1.04	0.07	1.04	-4.19	0.07	4.19	2.85	0.09	2.85	-1.51	0.10	1.52
NN	0.11	0.99	0.99	0.10	1.00	1.00	0.03	1.43	1.43	0.02	1.37	1.37
GLM	0.02	0.34	0.34	0.02	0.33	0.33	-0.19	0.50	0.54	2.40	0.45	2.44
DR (MLE)	0.03	0.34	0.34	0.05	0.34	0.34	0.06	0.41	0.41	3.31	0.43	3.33
DR (GEE)	0.03	0.34	0.34	0.03	0.33	0.33	0.04	0.42	0.42	2.78	0.42	2.82
Inverse probability weighting (MLE)												
IPW	-0.02	0.47	0.47	0.37	0.70	0.79	-0.02	0.77	0.77	2.32	1.31	2.66
QBIPW1	-0.02	0.47	0.47	-0.54	0.60	0.81	-0.02	0.75	0.75	-0.06	0.75	0.76
QBIPW2	-0.08	0.50	0.51	0.07	0.81	0.81	-0.14	0.79	0.80	-0.09	1.19	1.19
Inverse probability weighting (calibrated; GEE)												
IPW	0.01	0.39	0.39	-1.86	0.23	1.87	0.03	0.61	0.61	-0.88	0.28	0.92
QBIPW1	-0.01	0.37	0.37	-0.54	0.25	0.60	-0.02	0.58	0.58	0.09	0.52	0.52
QBIPW2	0.03	0.35	0.35	-0.11	0.27	0.29	0.05	0.52	0.53	-0.05	0.57	0.57

Note: DR = double robust; GEE = generalised estimating equations; GLM = generalised linear model; IPW = inverse probability weighting; MLE = maximum likelihood estimation; NN = nearest neighbour; OM = outcome model; PM = probability model; QBIPW = quantile balancing inverse probability weighting estimator; RMSE = root mean square error; SE = standard error.

In Scenario II, where the OM is linear and the PM is non-linear, it is evident that there are varying performance patterns across estimators. In the case of continuous outcomes, the NN, GLM, and DR

estimators demonstrate consistent and reliable performance, exhibiting minimal bias and low RMSE. Notwithstanding, the IPW MLE estimator exhibits a pronounced increase in both bias and RMSE, underscoring its sensitivity to non-linear probability models. The proposed QBIPW estimators demonstrate superior performance compared to the standard IPW, with QBIPW2 (GEE) exhibiting the lowest bias and a reasonable SE among IPW estimators. In the case of binary outcomes, the impact of the non-linear PM is less pronounced. The majority of estimators demonstrate low bias, with the QBIPW estimators under GEE exhibiting particularly favourable performance. It is noteworthy that QBIPW2 (GEE) achieves the lowest RMSE (0.29) among all IPW estimators, outperforming even the standard IPW (GEE) estimator.

The third scenario introduces a non-linear OM with a linear PM, which presents particular challenges for parametric methods. In the case of continuous outcomes, the GLM estimator demonstrates an elevated level of bias and root mean square error, which is indicative of its inherent limitation in accurately capturing non-linear relationships. The NN estimator demonstrates a tendency to maintain low bias, although this is accompanied by an increase in SE. The performance of DR methods varies. The DR (MLE) method demonstrates higher bias (1.72) in comparison to the DR (GEE) method (0.96). Among IPW estimators, those belonging to the QBIPW category demonstrate improved performance, with QBIPW2 (GEE) achieving the lowest RMSE within this category. In the case of binary outcomes, the impact of the non-linear OM is less pronounced. The majority of estimators demonstrate low bias, with QBIPW estimators under GEE once again exhibiting robust performance. QBIPW2 (GEE) achieves the lowest RMSE (0.53) among IPW estimators, exhibiting a performance level comparable to that of GLM and DR estimators.

In Scenario IV, which combines non-linear OM and PM, all estimators are confronted with the most challenging conditions. In the case of continuous outcomes, the majority of estimators are unable to perform satisfactorily. The GLM and DR estimators demonstrate a considerable degree of bias and error. The NN estimator demonstrates a tendency to maintain low bias, although it exhibits a high standard error. Among IPW estimators, those based on maximum likelihood estimation (MLE) exhibit considerable bias and root-mean-square error. The proposed QBIPW estimators under MLE and GEE demonstrate markedly improved performance, with QBIPW2 achieving the lowest bias and RMSE among all estimators. In the case of binary outcomes, although the overall performance is superior to that observed for continuous outcomes, challenges remain. The GLM and DR estimators exhibited an increased bias, with values ranging from 2.40 to 3.31. The proposed QBIPW estimators under GEE once again demonstrate superior performance compared to other IPW estimators. QBIPW2 achieves the lowest RMSE (0.57) in this category, comparable to the NN estimator (RMSE: 1.37). This scenario highlights the effectiveness of the proposed QBIPW estimators, particularly under GEE, in addressing complex non-linear relationships in both outcome and probability models.

Table 5.3 presents the empirical coverage rates (CR) for the proposed estimators. In Scenario I, the proposed estimators exhibited a CR that was close to the nominal 95% level. In Scenario II, this phenomenon is particularly evident in the IPW GEE, where a non-linear propensity score is employed and the assumed model is mis-specified. Nevertheless, the proposed QBIPW2 (with deciles and totals) exhibits CRs that are

nearly at the nominal level for both continuous and binary outcomes. Notably, the proposed approach significantly enhances the performance of the standard IPW when MLE is utilized as the estimation technique. The CI has the shortest width for the DR and IPW GEE estimators, which are one third (for continuous) or half (binary) the width of the NN estimator.

Table 5.3
Empirical coverage rate and average length of confidence interval based on analytical variance estimator

	OM	**Scenario I**		**Scenario II**		**Scenario III**		**Scenario IV**	
	PM	**linear (1)**		**linear (1)**		**non-linear (2)**		**non-linear (2)**	
		linear (1)		**non-linear (2)**		**linear (1)**		**non-linear (2)**	
Type	**Estimator**	**CR**	**Length**	**CR**	**Length**	**CR**	**Length**	**CR**	**Length**
				Continuous					
Basic	NN	95.60	24.73	94.80	24.78	94.80	59.82	96.60	59.84
	GLM	93.60	17.63	94.60	17.69	69.00	54.66	0.00	60.57
	DR (MLE)	93.60	17.69	95.40	18.40	95.40	24.00	6.80	363.34
	DR (GEE)	93.60	17.65	94.60	17.61	96.00	28.21	0.00	72.78
IPW MLE	IPW	95.80	22.36	10.40	122.99	99.20	38.23	2.40	787.33
	QBIPW1	94.80	21.67	98.20	47.09	100.00	45.92	83.40	92.19
	QBIPW2	95.60	23.55	97.59	55.69	99.60	56.49	96.98	102.93
IPW GEE	IPW	93.80	17.77	96.00	18.76	96.40	28.40	0.00	79.91
	QBIPW1	93.80	17.69	94.80	18.35	95.40	39.57	45.60	51.73
	QBIPW2	94.80	17.80	95.80	18.43	98.00	46.65	92.20	49.54
				Binary					
Basic	NN	94.20	3.82	95.00	3.82	93.20	5.34	95.00	5.34
	GLM	94.80	1.36	94.60	1.35	95.00	2.11	0.20	1.91
	DR (MLE)	92.80	1.30	94.00	1.32	95.80	1.72	0.00	1.97
	DR (GEE)	92.60	1.30	94.00	1.31	94.80	1.71	0.00	1.77
IPW MLE	IPW	94.60	1.81	98.40	3.15	95.20	3.12	84.00	5.91
	QBIPW1	94.40	1.74	75.60	2.29	93.00	2.84	95.80	3.02
	QBIPW2	95.20	1.90	96.39	3.57	96.60	3.05	92.77	5.30
IPW GEE	IPW	93.20	1.46	0.00	0.95	94.80	2.43	18.00	1.22
	QBIPW1	94.20	1.43	49.60	1.08	96.20	2.37	96.40	2.24
	QBIPW2	94.20	1.44	93.80	1.07	96.80	2.40	96.40	2.36

Note: CR = coverage rates; DR = double robust; GEE = generalised estimating equations; GLM = generalised linear model; IPW = inverse probability weighting; MLE = maximum likelihood estimation; NN = nearest neighbour; OM = outcome model; PM = probability model; QBIPW = quantile balancing inverse probability weighting estimator.

In Scenario III, the proposed approaches yielded results that were close to the nominal rates which exhibited excellent performance. Finally, with respect to the Scenario IV, it can be observed that when both models are mis-specified, only those incorporating deciles and totals (along with the NN) in the case of continuous data and QBIPW1 (under MLE and GEE) for binary data, display a coverage rate that is close to the nominal 95%. In Section B of the Appendix, we present a detailed analysis of the quality of the reproduced totals and quantiles for the proposed methods.

The simulation study illustrates the substantial advantages of integrating quantiles into IPW estimators. This approach markedly improves the robustness and accuracy of estimates, particularly in instances where both the outcome and probability models are mis-specified. QBIPW estimators demonstrate superior performance in complex, non-linear scenarios involving both continuous and binary outcomes, with GEE-based versions

generally outperforming their MLE counterparts. It seems that the width of CI is comparable with the GLM and DR estimators, which suggests that using more constraints doesn't increase the variance. The proposed approach also provides narrower CI than the NN estimator. Consequently, researchers engaged in the analysis of continuous or count variables from multiple sources should consider employing quantile-based IPW methods to enhance the robustness of their results, especially when confronted with uncertain model specifications or complex relationships in real-world data analysis.

6. Real data application

6.1 Data description

In this section we present an attempt to integrate administrative and survey data about job vacancies for the end of 2022Q2 in Poland. The aim was to estimate the share of vacancies aimed at Ukrainian workers. After the Russian invasion of Ukraine on the 24th February 2022, around 3.5 million persons (mainly women and children) arrived in Poland between 24th February and mid-May 2022 (Duszczyk and Kaczmarczyk, 2022). Some of them went to other European countries, but about 1 million stayed in Poland (as of 2023, cf. Statistics Poland (2023)).

The first source we used is the Job Vacancy Survey (JVS, known in Poland as the Labour Demand Survey), which is a stratified sample of 100,000 units, with a response rate of about 60% (S_B). The survey population consists of companies and their local units with 1 or more employees. The sampling frame includes information about the industry code NACE (19 levels), region (16 levels), sector (2 levels), size (3 levels) and the number of employees according to administrative data integrated by Statistics Poland (RE). The JVS sample contains 304 strata created separately for enterprises with up to 9 employees and those with more than 10 employees (cf. Statistics Poland, 2021).

The survey measures whether an enterprise has a job vacancy at the end of the quarter (on June 30, 2022) according to the following definition: *vacancies are positions or jobs that are unoccupied as a result of employee turnover or newly created positions or jobs that simultaneously meet the following three conditions*: (1) the jobs were actually vacant on the day of the survey, (2) the employer made efforts to find people willing to take on the job, (3) if suitable candidates were found to fill the vacancies, the employer would be willing to hire them. In addition, the JVS restricts the definition of vacancies by excluding traineeships, mandate contracts, contracts for specific work and business-to-business (B2B) contracts. Of the 60,000 responding units, around 7,000 reported at least one vacancy. Our target population included units with at least one vacancy, which according to the survey was between 38,000 and 43,000 at the end of 2022Q2.

The second source is the Central Job Offers Database (CBOP), which is a register of all vacancies submitted to Public Employment Offices (PEOs – S_A). CBOP is available online and can be accessed programmatically (via the API). CBOP includes all types of contracts and jobs outside Poland, so data cleaning was carried out to align the definition of a vacancy with that used in the JVS. CBOP data collected

via API include information about whether a vacancy is outdated (e.g. 17% of vacancies were outdated when downloaded at the end of 2022Q2). CBOP also contains information about unit identifiers (REGON and NIP), so we were able to link units to the sampling frame to obtain auxiliary variables with the same definitions as in the survey (24% of records did not contain an identifier because the employer can withhold this information). The final CBOP dataset contained about 8,500 units included in the sampling frame.

The overlap between JVS and CBOP was around 2,600 entities (around 4% of the JVS sample and 30% of CBOP), but only 30% of these reported at least one vacancy in the JVS survey. This suggests significant under-reporting in the JVS, which, however, is a problem beyond the scope of this paper. For the empirical study we treated both sources as separate and their correct treatment with the proposed methods will be investigated in the future.

6.2 Analysis and results

We defined our target variable as follows: Y is *the share of vacancies that have been translated into Ukrainian* (denoted as UKR), calculated separately for each unit. After the Russian aggression, many online job advertising services introduced information on whether the employer is particularly interested in employing Ukrainians (e.g. the Ukrainian version of the website was available, and job ads featured the following annotation: "We invite people from Ukraine").

Table 6.1 shows the distribution of registered employment according to the JVS survey, for companies observed in the CBOP and companies willing to employ Ukrainians (with at least one vacancy translated into Ukrainian; denoted as CBOP-UKR). Companies observed in the CBOP are on average larger than those responding to the JVS. The median employment for JVS was 8, while for CBOP and CBOP-UKR – 20 and 18 employees, respectively. This suggests that companies are selected into the CBOP depending on their size.

Table 6.1
Quantiles of registered employment (RE) for JVS, CBOP and the target variable UKR at the end of 2022Q2

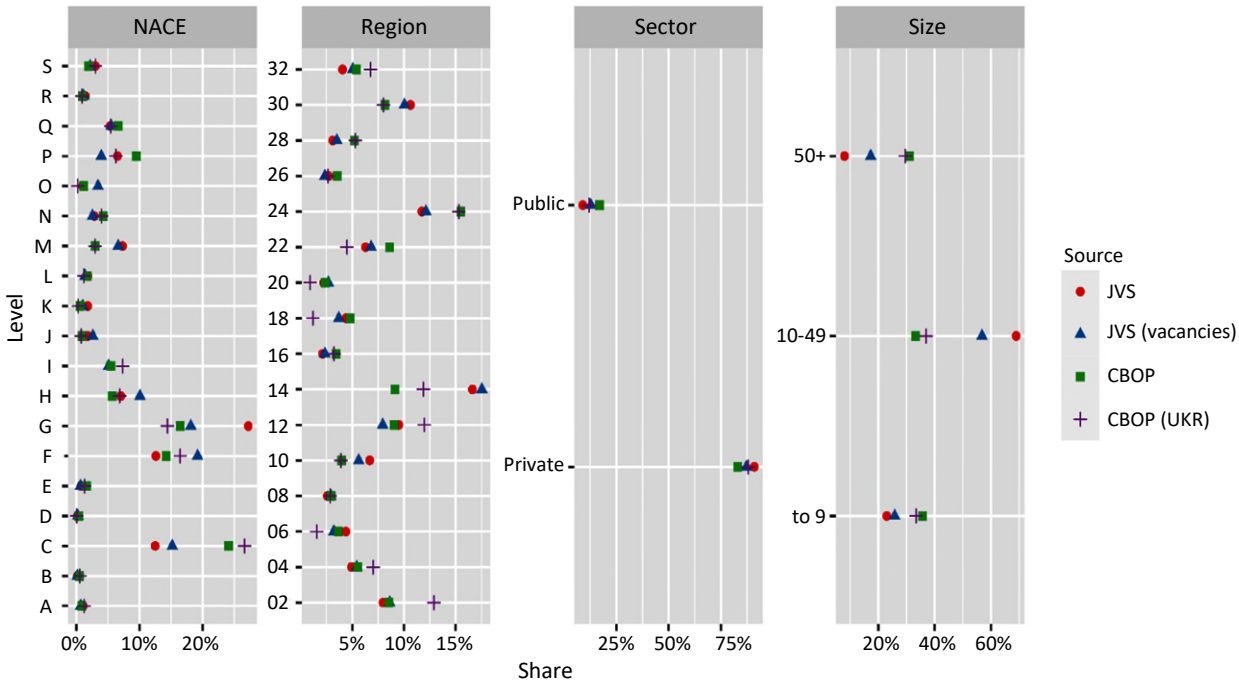
Decile of the RE	JVS	CBOP	CBOP-UKR
10%	3	4	4
20%	4	6	5
30%	5	8	8
40%	6	13	11
50%	8	20	18
60%	10	32	28
70%	17	51	48
80%	37	90	91
90%	100	211	215

Note: CBOP = Central Job Offers Database; JVS = Job Vacancy Survey; UKR = Ukrainian.

The distributions of four categorical variables in the two sources are shown in Figure 6.1. The largest discrepancies between the sources can be seen in the case of company size, where the shares in CBOP are almost equal. There are some differences for selected regions (e.g. 14 Mazowieckie with Warsaw, the capital of Poland; 02 Dolnośląskie with Wrocław and 12 Małopolskie with Cracow). Regions from the eastern part

of Poland (06 Lubelskie, 18 Podkarpackie, 20 Podlaskie), but also Pomorskie (22) with Tricity are characterised by the lowest share of vacancies for Ukrainians. The largest differences in terms of NACE are found for manufacturing (C), wholesale and retail trade (G) and hotels and restaurants (I).

Figure 6.1 The distribution of 4 auxiliary variables in the sources at the end of 2022Q2. JVS: the total sample, JVS (vacancies): companies with at least one vacancy, CBOP: register data set, CBOP-UKR: units with at least one vacancy translated into Ukrainian



Note: CBOP = Central Job Offers Database; JVS = Job Vacancy Survey; UKR = Ukrainian.

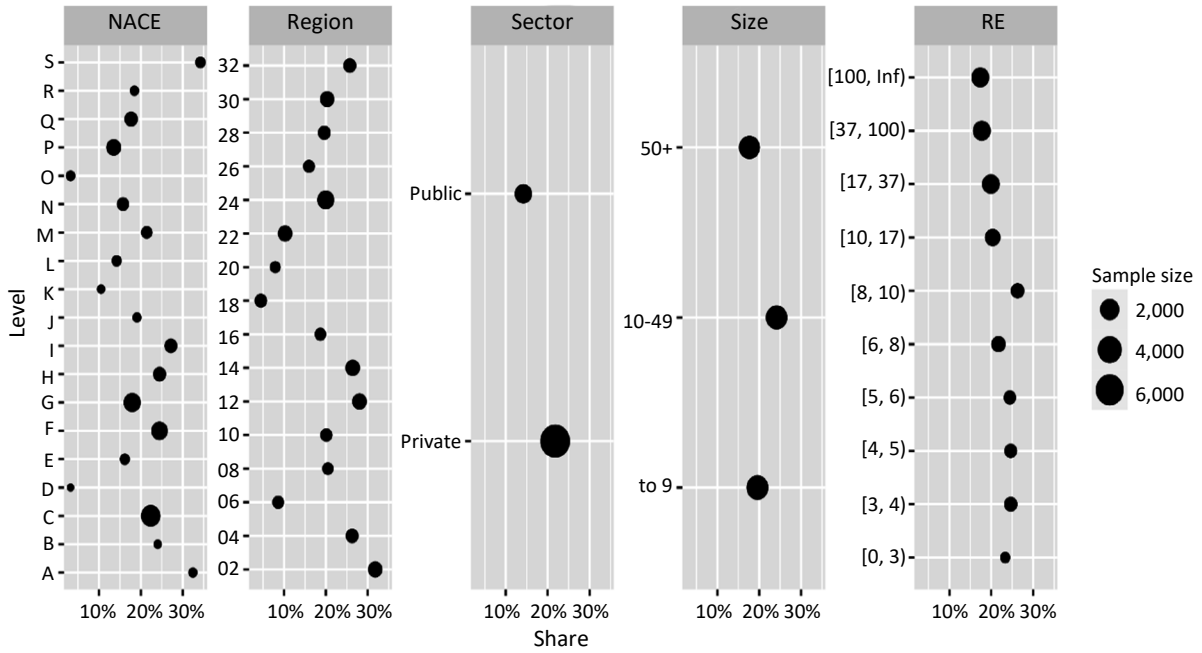
Figure 6.2 shows the share of vacancies targeted at Ukrainians according to four categorical variables and deciles of registered employment (RE) based on deciles estimated from the JVS survey. In general, the shares range between about 10% and 30%, especially across regions. The shares for the RE deciles range between 17% for the largest units and 26% for the median RE (8 employees).

The following combinations of variables were considered:

- Set 0: Region (16 levels), NACE (19 levels), Sector (2 levels), Size (3 levels), log (RE), log(# vacancies) (the number of vacancies), I (# vacancies = 1) (whether employer seeks only one person),
- Set 1A: Set 0 (without log (RE)) + Quartiles of the RE (estimated from the JVS),
- Set 1B: Set 0 + Quartiles of the RE,
- Set 2A: Set 0 (without log (RE)) + Deciles of the RE (without 10%),
- Set 2B: Set 0 + Deciles of RE (without 10%).

We decided not to include the number of vacancies for quantile balancing as almost 45% of vacancies were equal to 1 for CBOP and almost 30% for JVS. We consider the following estimators, assuming linear and logistic regression: GLM with Set 0, NN with set 0 (denoted as NN1), NN with $\log(\text{RE})$ and log(# vacancies) only (denoted as NN2), DR GEE with Set 0, IPW estimators under MLE and GEE with the following sets: IPW with Set 0, IPW with Set 1A (QBIPW1A), IPW with Set 1B (QBIPW1B), IPW with Set 2A (QBIPW2A) and IPW with Set 2B (QBIPW1B).

Figure 6.2 The target variable (the share of ads directed at Ukrainians) depending on 5 auxiliary variables at the end of 2022Q2



Variance was estimated using the following bootstrap approach: 1) JVS sample was resampled using a stratified bootstrap approach, 2) CBOP was resampled using simple random sampling with replacement. This procedure was repeated $B = 500$ times. Table 6.2 shows point estimates (denoted as Point), bootstrap standard errors (denoted as SE), the coefficients of variation (CV) and 95% confidence intervals.

The naive estimator suggests a 20.51% share, while other methods provide slightly higher estimates. GLM and DR GEE estimators show similar results (22.68% and 21.75% respectively) with low standard errors. NN estimators show higher variability, with NN1 having the highest point estimate (25.88%) but also the largest standard error. IPW and QBIPW estimators, both MLE and GEE-based, provide results ranging from 21.53% to 22.88%. Generally, GEE-based methods show lower standard errors and narrower confidence intervals compared to their MLE counterparts. Overall, most methods suggest the share of job vacancies for Ukrainians is between 21% and 23%, with GEE-based estimators showing the most precise estimates.

Table 6.2
Estimates of the share of job vacancies aimed at Ukrainians at the end of 2022Q2 in Poland

Estimator	Point	SE	CV	2.5%	97.5%
		Naïve			
Naïve	20.51	–	–	–	–
		Estimators			
GLM	22.68	0.59	2.61	21.52	23.84
NN1	25.88	6.41	24.78	13.31	38.45
NN2	23.38	1.91	8.18	19.63	27.13
DR GEE	21.75	0.60	2.78	20.57	22.94
		Inverse probability weighting (MLE)			
IPW	22.88	0.80	3.52	21.31	24.46
QBIPW1A	22.24	0.76	3.43	20.74	23.73
QBIPW1B	22.88	0.80	3.52	21.31	24.46
QBIPW2A	22.26	0.78	3.48	20.75	23.78
QBIPW2B	22.88	0.81	3.53	21.30	24.46
		Inverse probability weighting (GEE)			
IPW	21.75	0.60	2.78	20.57	22.94
QBIPW1A	21.53	0.60	2.78	20.36	22.70
QBIPW1B	21.69	0.60	2.78	20.51	22.87
QBIPW2A	21.69	0.61	2.83	20.49	22.89
QBIPW2B	21.74	0.61	2.83	20.53	22.94

Note: CV = coefficients of variation; DR = double robust; GEE = generalised estimating equations; GLM = generalised linear model; IPW = inverse probability weighting; MLE = maximum likelihood estimation; NN = nearest neighbour; QBIPW = quantile balancing inverse probability weighting estimator; SE = standard error.

7. Summary

Recent years have seen a revolution in survey research with regard to the use and integration of new data sources for statistical inference. Efforts in this area have focused on the integration of survey data and inference based on non-probability samples, with applications in official statistics (Salvatore, 2023). This is mainly because probability samples tend to be very costly and are associated with respondent burden. On the other hand, statistical inference based on non-probabilities samples is a challenge, mainly owing to bias inherent in the estimation process. Consequently, it is necessary to develop new methods of statistical inference based on non-probability samples, which could be be useful in real applications (Wu, 2022).

In this paper we have described quantile balancing inverse probability weighting for non-probability surveys. Based on our approach, for some of the methods considered in the article (IPW and DR) it is possible to reproduce not only totals for a set of auxiliary variables but also for a set of quantiles (or estimated quantiles). Such a solution improves robustness against model mis-specification, helps to decrease bias and improves the efficiency of estimation. These gains have been confirmed by our simulation study, in which we have shown that the inclusion of quantiles for inverse probability weighting estimators improves the quality of estimates.

The estimators proposed in this paper have the potential for real-life applications, provided that continuous variables of sufficient quality can be made available for use as auxiliary variables. The example of our empirical study, in which we estimated the share of job vacancies aimed at Ukrainian workers, demonstrates the effectiveness of the proposed methods in producing estimates of high precision.

Regarding aspects not addressed in this paper, one issue worth investigating further is the problem of under-coverage in non-probability samples, which is related to the violation of the positivity assumption (A2). The approach proposed in the paper equates the distributions of variables between the probability (or the population) and the non-probability samples and can serve as an alternative to other approaches proposed in the literature, which consist, for example, of using an additional sample of individuals from the population not covered by the non-probability sample (Chen, Li and Wu, 2023).

Acknowledgements

The authors' work has been financed by the National Science Centre in Poland, OPUS 20, grant no. 2020/39/B/HS4/00941.

We thank reviewers and Associate Editor for the comments that improved the paper. In addition, authors would like to thank Łukasz Chrostowski who developed the `nonprobsvy` package which implements state-of-the-art estimators proposed in the literature. The package is available through `CRAN` and at https://github.com/ncn-foreigners/nonprobsvy. We also thank Natalie Shlomo and Tymon Świtalski for their valuable comments on the draft version of this paper.

Codes to reproduce the simulation study are freely available from the github repository: https://github.com/ncn-foreigners/paper-nonprob-qcal. An R package that implements joint calibration is available at https://github.com/ncn-foreigners/jointCalib. The package is based on calibration implemented in `survey`, `sampling` or `laeken` packages.

Appendix

A. Assumptions that guarantee unique solution to estimating equations

In the appendix we use notation $\mathbf{z}_k = \begin{pmatrix} \mathbf{x}_k \\ \mathbf{a}_k \end{pmatrix}$.

(B1) The matrix $\sum_{k \in S_A} \mathbf{z}_k \mathbf{z}_k^T$ is positive definite.

(B2) The matrix $\sum_{k \in S_B} \mathbf{z}_k \mathbf{z}_k^T$ is positive definite.

In the appendix we use the following notation:

$$\begin{aligned} \pi(\mathbf{z}; \boldsymbol{\eta}) &:= \frac{\exp(\mathbf{z}^T \boldsymbol{\eta})}{1 + \exp(\mathbf{z}^T \boldsymbol{\eta})}, \\ \pi(t) &:= \frac{\exp(t)}{1 + \exp(t)}, \\ \pi'(\mathbf{z}; \boldsymbol{\eta}) &:= \pi(\mathbf{z}; \boldsymbol{\eta}) \big(1 - \pi(\mathbf{z}; \boldsymbol{\eta})\big). \end{aligned}$$

Theorem 1. *Assumption (B1) is a necessary and sufficient condition for existence of a unique solution to:*

$$\boldsymbol{G}(\boldsymbol{\eta}) = \sum_{k \in S_A} \frac{\boldsymbol{z}_k}{\pi(\boldsymbol{z}_k; \boldsymbol{\eta})} - \sum_{k \in S_B} d_k^B \boldsymbol{z}_k = \boldsymbol{0},$$

and assumption (B2) is a necessary and sufficient condition for existence of a unique solution to:

$$\boldsymbol{U}(\boldsymbol{\eta}) = \sum_{k \in S_A} \boldsymbol{z}_k - \sum_{k \in S_B} d_k^B \pi(\boldsymbol{z}_k; \boldsymbol{\eta}) \boldsymbol{z}_k = \boldsymbol{0}.$$

Conditions (B1) and (B2) are easy to check through the process of data analysis. If either of these is violated a potential remedy is to reduce the number of quantiles used. Notice that (B1) implies $n_A \geq J$ where J is the dimension of $\mathbf{z}$ and that there are at least J linearly independent $\mathbf{z}$'s in S_A. (B2) implies analogous properties of S_B.

Proof. We will first prove that ℓ^* (the anti-derivative of $\mathbf{U}$) has a unique global maximum. Notice that if (B2) holds the hessian of ℓ^*:

$$\frac{\partial \ell^*}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \mathbf{U}(\boldsymbol{\eta}) = -\sum_{k \in S_B} d_k^B \pi'(\mathbf{z}; \boldsymbol{\eta}) \mathbf{z}_k \mathbf{z}_k^T,$$

is positive definite for any $\boldsymbol{\eta} \in \mathbb{R}^J$ so ℓ^* is strictly concave. Therefore, it has at most one global maximum and since hessian is always invertible the second derivative test tells us that $\mathbf{U}(\boldsymbol{\eta})$ has at most one root.

Now notice that (B2) implies, by rank-nullity theorem, that we have:

$$\text{Span}(\mathbf{z}_1, \ldots, \mathbf{z}_{n_A})^\perp = (\mathbb{R}^J)^\perp = \{\mathbf{0}\},$$

where $(\cdot)^\perp$ denotes the orthogonal component, which in turn gives us:

$$\exists j \in S_A : |\mathbf{z}_j^T \boldsymbol{\eta}_m| \to \infty, \tag{A.1}$$

so π_j^A tends to either 0 or 1. This implies that the PL function:

$$\ell^*(\boldsymbol{\eta}) = \sum_{k \in S_A} \left(\log(\pi_k^A) - \log(1 - \pi_k^A) \right) + \sum_{k \in S_B} d_k^B \log(1 - \pi_k^A),$$

which is bounded from above must tend to $-\infty$.

Similarly an analogous reasoning for any anti-derivative of $\mathbf{G}$, for example:

$$H(\boldsymbol{\eta}) = \sum_{k \in S_A} \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{1}{\pi(t)} dt - \sum_{k \in S_B} d_k^B \boldsymbol{\eta}^T \mathbf{z}_k,$$

for some $q \in \mathbb{R}$, the hessian of which is:

$$\frac{\partial}{\partial \boldsymbol{\eta}} \mathbf{G}(\boldsymbol{\eta}) = -\sum_{k \in S_A} \frac{\mathbf{z}_k \mathbf{z}_k^T}{\pi(\mathbf{z}_k; \boldsymbol{\eta})^2} \pi'(\mathbf{z}_k; \boldsymbol{\eta}),$$

guarantees that $\mathbf{G}$ has at most one root. For reasons that will become clearer down the line we choose $q < 0$ such that:

$$\left\| \sum_{k \in S_B} d_k^B \mathbf{z}_k - \sum_{k \in S_A} \mathbf{z}_k \right\| < n_A(e^{-q} - q),$$

which is possible since $q \mapsto e^{-q} - q$ is a continuous bijection from $\mathbb{R}$ to $\mathbb{R}$.

Strict concavity of H implies that a sufficient condition for existence of a root of $\mathbf{G}$ is:

$$\lim_{\|\boldsymbol{\eta}\| \to \infty} H(\boldsymbol{\eta}) = -\infty, \tag{A.2}$$

where $\|\cdot\|$ denotes the euclidean norm $\zeta \overset{\|\cdot\|}{\mapsto} \sqrt{\left|\zeta^T \zeta\right|}$. To prove (A.2) it is sufficient to prove that $\lim_{m\to\infty} |H(\boldsymbol{\eta}_m)| = \infty$ for sequences $(\boldsymbol{\eta}_m)_{m\in\mathbb{N}} \subseteq \mathbb{R}^J$ with $\lim_{m\to\infty} \|\boldsymbol{\eta}_m\| = \infty$, such that the sequence $\operatorname{sgn}(\boldsymbol{\eta}_m) = \frac{\boldsymbol{\eta}_m}{\|\boldsymbol{\eta}_m\|}$ is constant.

To see that the claim above is true first observe that for any vector $\boldsymbol{\eta}$ and $s \geq 1$:

$$\begin{aligned}
H(s\boldsymbol{\eta}) - sH(\boldsymbol{\eta}) &= \sum_{k\in S_A} \left(\int_q^{s\boldsymbol{\eta}^T \mathbf{z}_k} \frac{1}{\pi(t)}\, dt - \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(t)}\, dt \right) \\
&\leq \sum_{k\in S_A} \left(\int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(st)}\, dt + \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt - \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{s}{\pi(t)}\, dt \right) \\
&= \sum_{k\in S_A} \int_q^{\boldsymbol{\eta}^T \mathbf{z}_k} \frac{e^{-s}-1}{e^t}\, dt + n_A \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt \\
&= \sum_{k\in S_A} \left(e^{-s}-1\right)\left(-e^{-\boldsymbol{\eta}^T \mathbf{z}_k} + e^{-q}\right) + n_A \int_{q/s}^{q} \frac{s}{\pi(st)}\, dt \\
&= \left(1-e^{-s}\right)\left(\sum_{k\in S_A} e^{-\boldsymbol{\eta}^T \mathbf{z}_k} - n_A e^{-q} \right) + n_A \left(sq - q + e^{-q} - e^{-qs} \right),
\end{aligned}$$

where the second line follows by substitution $us = t, sdu = dt,$ and the last term is bounded from above in s. By local boundedness of H there exists a non-negative constant R such that:

$$\forall s > 0 : H\left(s\boldsymbol{\eta}\right) \leq sH\left(\boldsymbol{\eta}\right) + R.$$

Thus, we have for any sequence of normed vectors $\left(\boldsymbol{\omega}_m\right)_{m\in\mathbb{N}} \subseteq \mathbb{R}^J$:

$$\begin{aligned}
\sup\{t > 0 : H(t\boldsymbol{\omega}_m) > -M\} &\leq \sup\{t > 0 : tH(\boldsymbol{\omega}_m) > -M - R\} \\
&\leq \frac{-M-R}{H(\boldsymbol{\omega}_m)} \to \infty \Leftrightarrow H(\boldsymbol{\omega}_m) \to 0.
\end{aligned}$$

However, by continuity the last implies that for every $\varepsilon > 0$ there exists $\mathbf{u}$ with $\|\mathbf{u}\| = 1$ and $H(\mathbf{u}) > -\varepsilon$, but this is impossible since then:

$$\begin{aligned}
\sum_{k\in S_B} d_k^B \mathbf{u}^T \mathbf{z}_k - \varepsilon &> \sum_{k\in S_A} \int_q^{\mathbf{u}^T \mathbf{z}_k} \frac{1}{\pi(t)}\, dt = n_A(e^{-q} - q) + \sum_{k\in S_A} (\mathbf{u}^T \mathbf{z}_k + e^{-\mathbf{u}^T \mathbf{z}_k}) \\
\Rightarrow \left\| \sum_{k\in S_B} d_k^B \mathbf{z}_k - \sum_{k\in S_A} \mathbf{u}^T \mathbf{z}_k \right\| &\geq \left| \mathbf{u}^T \left(\sum_{k\in S_B} d_k^B \mathbf{z}_k - \sum_{k\in S_A} \mathbf{u}^T \mathbf{z}_k \right) \right| \\
&\geq n_A(e^{-q} - q) + \sum_{k\in S_A} e^{-\mathbf{u}^T \mathbf{z}_k} + \varepsilon,
\end{aligned}$$

which is impossible by virtue of how constant q was chosen. Therefore we have that:

$$\forall M > 0: \sup_{\|\boldsymbol{\omega}\|=1} \inf\{t > 0: \forall s > t: H(s\boldsymbol{\omega}) < -M\} = \sup_{\|\boldsymbol{\omega}\|=1} \sup\{t > 0: H(t\boldsymbol{\omega}) \geq -M\} < \infty.$$

Now assume that:

$$\forall M > 0, \|\boldsymbol{\omega}\| = 1 \; \exists N > 0 \; \forall r > N: H(r\boldsymbol{\omega}) < -M.$$

Choose any $M > 0$ and let:

$$T := \sup_{\|\boldsymbol{\omega}\|=1} \sup\{t > 0: H(t\boldsymbol{\omega}) \geq -M\} < \infty.$$

Now for any $r > T$ we have:

$$\forall \|z\| = 1: H(rz) < -M.$$

Putting it altogether yields:

$$\forall M > 0 \; \exists T > 0 \; \forall r > T \; \forall \|\boldsymbol{\omega}\| = 1: H(rz) < -M.$$

It is sufficient to check that:

$$H(\boldsymbol{\eta}_m) \to \infty,$$

for sequences of $\boldsymbol{\eta}_m$ with constant sign and increasing to infinity in norm as was claimed previously.

Now we proceed to prove the central claim about H. We have already proved that:

$$\forall \boldsymbol{\eta} \in \mathbb{R}^J \; \exists R > 0 \; \forall s > 0: H(s\boldsymbol{\eta}) \leq sH(\boldsymbol{\eta}) + R,$$

and chosen q such that:

$$\forall \|\mathbf{z}\| = 1: H(\mathbf{z}) < 0.$$

The convergence to $-\infty$ for sequences $(\boldsymbol{\eta}_m)_{m\in\mathbb{N}}$ defined previously is straightforwardly attained.

It is not difficult to see that (B2) is in fact necessary for the existence of a unique solution to $\mathbf{G} = \mathbf{0}$. If (B2) is violated then, since the matrix in (B2) is clearly non-negative define, we can pick some $\boldsymbol{\xi}$ such that:

$$\mathbf{0} \neq \boldsymbol{\xi} \in \operatorname{Span}(\mathbf{z}_1, \ldots, \mathbf{z}_{n_A})^{\perp}.$$

Now if $\mathbf{G} = \mathbf{0}$ has a solution $\boldsymbol{\eta}$ then:

$$\mathbf{0} = \mathbf{G}(\boldsymbol{\eta}) = \sum_{k\in S_A} \frac{\mathbf{z}_k}{\pi(\boldsymbol{\eta}^T\mathbf{z}_k)} - \sum_{k\in S_B} d_k^B \mathbf{z}_k = \sum_{k\in S_A} \frac{\mathbf{z}_k}{\pi(\boldsymbol{\eta}^T\mathbf{z}_k + \boldsymbol{\xi}^T\mathbf{z}_k)} - \sum_{k\in S_B} d_k^B \mathbf{z}_k = \mathbf{G}(\boldsymbol{\eta}+\boldsymbol{\xi}).$$

So $\boldsymbol{\xi}+\boldsymbol{\eta}$ is also a solution to $\mathbf{G} = \mathbf{0}$ and therefore the space of solutions to the GEE equation has exactly dimension equal to $J - \operatorname{Rank}\left(\sum_{k\in S_A} \mathbf{z}_k\mathbf{z}_k^T\right)$ if such a solution exists. An analogous arguments shows that assumption (B1) is not only sufficient but also necessary for the existence of a unique solution to $\mathbf{U} = \mathbf{0}$.

B. Quality of reproduced totals and quantiles

This section presents the results of the quality assessment of the total and quantile reproduction for all IPW estimators employed in the simulation study. For the following quantities and each simulation run $(r = 1, \ldots, 500)$ the euclidean norm of the difference between them and their estimates respectively have been calculated as:

- Population size:

$$v_{r,\hat{N}} = |\hat{N}_r - N|.$$

- Quantiles (for x_1 and x_2 jointly, considering both quartiles and deciles, where $\alpha \in \{0.25, 0.50, 0.75, 0.1, \ldots, 0.9\}$):

$$v_{r,\hat{Q}} = \sqrt{\sum_{\alpha}\sum_{p=1}^{2}\left(\hat{Q}_{r,x_p,\alpha} - Q_{r,x_p,\alpha}\right)^2}.$$

- Totals (for x_1 and x_2 jointly):

$$v_{r,\hat{\tau}} = \sqrt{\sum_{p=1}^{2}\left(\hat{\tau}_{r,x_p} - \tau_{r,x_p}\right)^2}.$$

The reference values for $\hat{\tau}_{x_p,r}$ and $Q_{r,x_p,\alpha}$ are estimated from the probability sample S_B. These may vary in the simulation as we do not calibrate the S_B sample.

Table B.1 reports both the mean and median values of $v_{r,\hat{N}}$, $v_{r,\hat{\tau}}$, and $v_{r,\hat{Q}}$. The inclusion of both measures is intended to account for instances where the algorithm did not converge, resulting in imperfect reproduction of totals and quantiles for the GEE estimator.

As anticipated based on the design, GEE-based estimators demonstrate superior performance compared to MLE-based methods across all measures. It is noteworthy that GEE-based methods demonstrate superior performance with regard to quantile reproduction, with QBIPW2 achieving the lowest error rates. The proposed QBIPW1 and QBIPW2 methods demonstrate exceptional performance, particularly in the context of the more challenging PM2 scenario. This is due to their design, which accounts for both quantiles and totals. QBIPW1 accounts for quartiles, while QBIPW2 utilises deciles. In the interpretation we focus on the *median* results.

The results demonstrate a pronounced discrepancy in performance between the PM1 and PM2 scenarios, with all estimators exhibiting better outcomes under PM1. This discrepancy is apparent not only in the reproduction of population size and totals, but also in the reproduction of quantiles. For example, under PM1, the majority of methods achieve mean quantile errors below 0.2, whereas under PM2, these errors frequently exceed 1.0 for MLE-based methods. Nevertheless, the proposed GEE-based QBIPW1 and QBIPW2 methods demonstrate remarkable resilience in quantile estimation, even under PM2 (with mean quantile errors as low as 0.05 for QBIPW2), thereby illustrating their capacity to capture the population's

distribution across diverse probability models. The consistent performance in quantile reproduction, coupled with the accurate estimation of population size and totals, serves to reinforce the efficacy of the proposed methods in incorporating both quantile and total information in IPW estimation. The superior performance of QBIPW1 and QBIPW2 under PM2 serves to demonstrate their capacity to handle complex probability models by leveraging a more comprehensive set of population characteristics.

Table B.1
Quality of reproduced population size, totals and quantiles

Estimator	PM1			PM2		
	$v_{r,\hat{N}}$	$v_{r,\hat{Q}}$	$v_{r,\hat{\tau}}$	$v_{r,\hat{N}}$	$v_{r,\hat{Q}}$	$v_{r,\hat{\tau}}$
Mean from simulations						
Naive	–	0.52	–	–	2.28	–
MLE						
IPW	658.81	0.16	1,049.47	11,779.12	2.98	78,050.24
QBIPW1	726.83	0.16	1,094.00	3,403.08	0.90	13,198.45
QBIPW2	1,334.84	0.15	1,884.13	7,310.56	0.46	20,054.55
GEE						
IPW	0.00	0.15	0.00	0.00	1.37	0.00
QBIPW1	0.11	0.10	2.74	0.00	0.36	0.00
QBIPW2	133.95	0.05	406.16	41.10	0.05	36.12
Medians from simulations						
Naive	–	0.51	–	–	2.28	–
MLE						
IPW	358.00	0.15	760.42	10,617.02	2.76	70,068.24
QBIPW1	353.72	0.15	732.87	3,081.35	0.86	10,515.02
QBIPW2	442.59	0.12	901.13	3,242.52	0.41	8,157.91
GEE						
IPW	0.00	0.14	0.00	0.00	1.37	0.00
QBIPW1	0.00	0.10	0.00	0.00	0.36	0.00
QBIPW2	0.00	0.04	0.00	0.00	0.05	0.00

Note: GEE = generalised estimating equations; IPW = inverse probability weighting; MLE = maximum likelihood estimation; PM = probability model; QBIPW = quantile balancing inverse probability weighting estimator.

References

Ai, C., Linton, O. and Zhang, Z. (2020). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*, 30, 1949-1970. https://doi.org/10.5705/ss.202018.0107.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46(1), 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Beręsewicz, M. (2017). A two-step procedure to measure representativeness of internet data sources. *International Statistical Review*, 85(3), 473-493.

Beręsewicz, M. (2023). *jointCalib: A Joint Calibration of Totals and Quantiles* [R package version 0.1.0]. https://CRAN.R-project.org/package=jointCalib.

Chen, S., Yang, S. and Kim, J.K. (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, 10(1), 1-24. https://doi.org/10.1093/jssam/smaa036.

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.

Chen, Y., Li, P. and Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Survey Methodology*, 49(2), 497-515. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00005-eng.pdf.

Chrostowski, Ł. and Beręsewicz, M. (2024). *nonprobsvy: Package for Inference Based on Non-Probability Samples* [R package version 0.1.0]. https://github.com/ncn-foreigners/nonprobsvy.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2), 137-161. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf.

Daas, P.J., Puts, M.J., Buelens, B. and Hurk, P.A.v.d. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2), 249-262.

Dennis Jr, J.E. and Schnabel, R.B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Duszczyk, M. and Kaczmarczyk, P. (2022). The war in Ukraine and migration to Poland: Outlook and challenges. *Intereconomics*, 57(3), 164-170.

Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2). https://doi.org/10.1214/16-STS598.

Gelman, A. and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2), 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32(1), 37-52. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9255-eng.pdf.

Hasselman, B. (2023). *Nleqslv: Solve Systems of Nonlinear Equations* [R package version 3.3.5]. https://CRAN.R-project.org/package=nleqslv.

Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *SSRN Electronic Journal*, 30(3). https://doi.org/10.2139/ssrn.2746753.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1), 243-263. https://doi.org/10.1111/rssb.12027.

Kim, J.K. and Riddles, M.K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38(2), 157-165. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2012002/article/11754-eng.pdf.

Kim, J.-K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87, S177-S191.

Park, S., Kim, J.K. and Kim, K. (2019). A note on propensity score weighting method using paradata in survey sampling. *Survey Methodology*, 45(3), 451-463. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00002-eng.pdf.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Salvatore, C. (2023). Inference with non-probability samples and survey data integration: A science mapping study. *Metron*, 1-25.

Sant'Anna, P.H., Song, X. and Xu, Q. (2022). Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6), 1093-1120.

Statistics Poland (2021). *Methodological Report the Demand for Labour*. https://stat.gov.pl/obszary-tematyczne/rynek-pracy/popyt-na-prace/zeszyt-metodologiczny-popyt-na-prace,3,1.html.

Statistics Poland (2023). *Residents of Ukraine Under Temporary Protection*. https://stat.gov.pl/en/topics/population/internationa-migration/residents-of-ukraine-under-temporary-protection,9,1.html.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data* (Vol. 4). Springer.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

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. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.