Quantile balancing inverse probability weighting for non-probability samples

Maciej Beręsewicz*, Marcin Szymkowiak†, Piotr Chlebicki‡

Abstract

Usage of non-statistical data sources for statistical purposes have become increasingly popular in recent years, also in official statistics. However, statistical inference based on non-probability samples is made more difficult by nature of them being biased and not representative of the target population. In this paper we propose *quantile balancing inverse probability weighting estimator* (QBIPW) for non-probability samples. We use the idea of Harms and Duchesne (2006) which allows to include quantile information in the estimation process so known totals and distribution for auxiliary variables are being reproduced. 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 method to estimate the share of vacancies aimed at Ukrainian workers in Poland using an integrated set of administrative and survey data about job vacancies.

Key Words: data integration, calibration approach, job vacancy survey.

*Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland, E-mail: maciej.beresewicz@ue.poznan.pl; Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland.

†Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland; Statistical Office in Poznań, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland.

‡Stockholm University, Department of Mathematics, Matematiska institutionen, Albano hus 1, 106 91 Stockholm.

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 covers 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-statistical 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, & 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 & 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 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 $E(Y|\boldsymbol{X})$ 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 (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. 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 solution 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 $\boldsymbol{x}$ and the target variable y, with their corresponding values $\boldsymbol{x}_k$ and y_k, respectively. Let $\{(y_k, \boldsymbol{x}_k), k \in S_A\}$ be a dataset of a non-probability sample of size n_A and $\{(\boldsymbol{x}_k, d_k^B), k \in S_B\}$ – a dataset of a probability sample of size n_B. Only information about auxiliary variables $\boldsymbol{x}$ 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 $\boldsymbol{x}$	Study variable y
Non-probability sample (S_A)	1	?	✓	✓
	$\vdots$	?	$\vdots$	$\vdots$
	n_A	?	✓	✓
Probability sample (S_B)	1	✓	✓	?
	$\vdots$	$\vdots$	$\vdots$	?
	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 $\pi_k^A = P(R_k = 1|\boldsymbol{x}_k)$ for $k = 1, ..., N$ be their propensity scores. For π_k^A we can assume a parametric model $\pi(\boldsymbol{x}_k, \boldsymbol{\gamma})$ (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 $\boldsymbol{x}_k^{\circ}$ be a J_1-dimensional vector of auxiliary variables (benchmark variables) for which $\tau_{\boldsymbol{x}^{\circ}} = \sum_{k \in U} \boldsymbol{x}_k^{\circ} = \left(\sum_{k \in U} x_{k1}, \ldots, \sum_{k \in U} x_{kJ_1}\right)^T$ 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 $\tau_{\boldsymbol{x}^{\circ}}$ exactly. In other words, in order to find new calibration weights w_k^B we have to minimise a distance function $D(\boldsymbol{d}, \boldsymbol{v}) = \sum_{k \in S_B} d_k^B G\left(\frac{v_k^B}{d_k^B}\right) \to \min$ to fulfil calibration equations $\sum_{k \in S_B} v_k^B \boldsymbol{x}_k^{\circ} = \sum_{k \in U} \boldsymbol{x}_k^{\circ}$, where $\boldsymbol{d} = \left(d_1^B, \ldots, d_{n_B}^B\right)^T$, $\boldsymbol{v} = \left(v_1^B, \ldots, v_{n_B}^B\right)^T$ and $G(\cdot)$ is a function that must satisfy some regularity conditions: $G(\cdot)$ is strictly convex and twice continuously differentiable, $G(\cdot) \geq 0$, $G(1) = 0$, $G'(1) = 0$ and $G''(1) = 1$. Examples of $G(\cdot)$ functions are given by Deville and Särndal (1992). 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 $w_k^B = d_k^B + d_k^B \left(\tau_{\boldsymbol{x}} - \hat{\tau}_{\boldsymbol{x}\pi}\right)^T \left(\sum_{j \in S_B} d_j^B \boldsymbol{x}_j^{\circ} \boldsymbol{x}_j^{\circ T}\right)^{-1} \boldsymbol{x}_k^{\circ}$. It is worth adding that in order to avoid negative or large w_k^B weights in the process of minimising the $D(\cdot)$ function, one can consider some boundary constraints $L \leq \frac{w_k^B}{d_k^B} \leq U$, where $0 \leq L \leq 1 \leq U,\ k = 1, \ldots, n_B$. 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 $\hat{\tau}_{y\boldsymbol{x}^{\circ}} = \sum_{k \in S_B} w_k y_k$, where w_k^B are calibration weights obtained using a specific $G(\cdot)$ function.

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 (assuming for a moment that values of y_k in this sample are known).

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 $Q_{y,\alpha} = \inf\left\{t \,|\, F_y(t) \geq \alpha\right\}$, where $F_y(t) = N^{-1} \sum_{k \in U} H(t - y_k)$ and the

Heavyside function is given by

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

We assume that $\boldsymbol{Q}_{\boldsymbol{x}^*,\alpha}=\left(Q_{x_1,\alpha},\ldots,Q_{x_{J_2},\alpha}\right)^T$ is a vector of known population quantiles of order α for a vector of auxiliary variables $\boldsymbol{x}_k^*$, where $\alpha\in(0,1)$ and $\boldsymbol{x}_k^*$ is a J_2-dimensional vector of auxiliary variables. It is worth noting that, in general, the numbers J_1 and J_2 of auxiliary variables may be different. It may happen that for a specific auxiliary variable its population total and the corresponding quantile of order α will be known. However, in most cases quantiles will be known for continuous auxiliary variables, unlike totals, which will generally be known for categorical variables. In order to find new calibration weights w_k^B that reproduce known population quantiles in a vector $\boldsymbol{Q}_{\boldsymbol{x}^*,\alpha}$, an interpolated distribution function estimator of $F_y\left(t\right)$ is defined as $\hat{F}_{y,cal}(t)=\frac{\sum_{k\in S_B}w_k^B H_{y,S_B}(t,y_k)}{\sum_{k\in S_B}w_k^B}$, where the Heavyside function in formula (1) is replaced by the modified function $H_{y,S_B}\left(t,y_k\right)$ given by

$$H_{y,S_B}\left(t,y_k\right)=\begin{cases}1, & y_k\leqslant L_{y,S_B}(t),\\ \vartheta_{y,S_B}\left(t\right), & y_k=U_{y,S_B}\left(t\right),\\ 0, & y_k>U_{y,S_B}\left(t\right),\end{cases}\tag{2}$$

where appropriate parameters are defined as $L_{y,S_B}\left(t\right)=\max\left\{\left\{y_k,k\in S_B\mid y_k\leqslant t\right\}\cup\left\{-\infty\right\}\right\}$, $U_{y,S_B}\left(t\right)=\min\left\{\left\{y_k,k\in S_B\mid y_k>t\right\}\cup\left\{\infty\right\}\right\}$ and $\vartheta_{y,S_B}\left(t\right)=\frac{t-L_{y,S_B}(t)}{U_{y,S_B}(t)-L_{y,S_B}(t)}$ 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. $H\left(x\right)\approx\frac{1}{2}+\frac{1}{2}\tanh kx=(1+e^{-2kx})^{-1}$, 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 $\hat{Q}_{y,cal,\alpha}=\hat{F}_{y,cal}^{-1}(\alpha)$, where a vector $\boldsymbol{w}=\left(w_1^B,\ldots,w_{n_B}^B\right)^T$ is a solution of an optimization problem $D\left(\boldsymbol{d},\boldsymbol{v}\right)=\sum_{k\in S_B}d_k^B\,G\left(\frac{v_k^B}{d_k^B}\right)\to\min$ subject to the calibration constraints $\sum_{k\in S_B}v_k^B=N$ and $\hat{\boldsymbol{Q}}_{\boldsymbol{x}^*,cal,\alpha}=\left(\hat{Q}_{x_1,cal,\alpha},\ldots,\hat{Q}_{x_{J_2},cal,\alpha}\right)^T=\boldsymbol{Q}_{\boldsymbol{x}^*,\alpha}$ or equivalently $\hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right)=\alpha$, where $j=1,\ldots,J_2$.

As in the previous case, if $G\left(x\right)=\frac{(x-1)^2}{2}$, then, using the method of Lagrange multipliers, the final weights w_k^B can be expressed as $w_k^B=d_k^B+d_k^B\left(\mathbf{T_a}-\sum_{k\in S_B}d_k^B\boldsymbol{a}_k\right)^T\left(\sum_{j\in S_B}d_j^B\boldsymbol{a}_j\boldsymbol{a}_j^T\right)^{-1}\boldsymbol{a}_k$, where $\mathbf{T_a}=\left(N,\alpha,\ldots,\alpha\right)^T$ and the elements of $\boldsymbol{a}_k=\left(1,a_{k1},\ldots,a_{kJ_2}\right)^T$ are given by

$$
a_{kj} = \begin{cases} N^{-1}, & x_{kj} \leq L_{x_j,S_B}\left(Q_{x_j,\alpha}\right), \\ N^{-1}\vartheta_{x_j,S_B}\left(Q_{x_j,\alpha}\right), & x_{kj} = U_{x_j,S_B}\left(Q_{x_j,\alpha}\right), \\ 0, & x_{kj} > U_{x_j,S_B}\left(Q_{x_j,\alpha}\right), \end{cases} \tag{3}
$$

with $j = 1, \ldots, J_2$ and $\vartheta(\cdot)$ is defined as in (2).

Assuming that $y_1 \leq y_2 \ldots \leq y_{n_B}$, it can be shown that if there exists $l \in \{1, \ldots, n_B - 1\}$ such that $\hat{F}_{y,cal}(y_l) \leq \alpha$, $\hat{F}_{y,cal}(y_{l+1}) > \alpha$ and $\hat{F}_{y,cal}$ is invertible at point $\hat{Q}_{y,cal,\alpha}$, then the calibration estimator $\hat{Q}_{y,cal,\alpha}$ of quantile $Q_{y,\alpha}$ of order $\alpha \in (0,1)$ can be expressed as $\hat{Q}_{y,cal,\alpha} = y_l + \frac{N\alpha - \sum_{k=1}^{l} w_k^B}{w_{l+1}^B}(y_{l+1} - y_l)$.

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 $\boldsymbol{x}_k$;

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

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

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).

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 $\pi_k^A = P(R_k = 1|\boldsymbol{x}_k^\circ) = \pi(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}_0)$, where $\boldsymbol{\gamma}_0$ is the true value of unknown model parameters. The maximum likelihood estimator of π_k^A is computed as $\hat{\pi}_k^A = \pi(\boldsymbol{x}_k^\circ, \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ maximises the log-likelihood function under full information:

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

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

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

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

$$\boldsymbol{U}(\boldsymbol{\gamma}) = \sum_{k \in S_A} \boldsymbol{x}_k^\circ - \sum_{k \in S_B} d_k^B \pi\left(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}\right) \boldsymbol{x}_k^\circ. \tag{4}$$

Note that weights based on $\pi\left(\boldsymbol{x}_k^\circ, \hat{\boldsymbol{\gamma}}\right)$ where $\hat{\boldsymbol{\gamma}}$ are estimated based on the (4) do not reproduce the population size nor $\boldsymbol{x}^\circ$ totals.

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

$$\boldsymbol{G}(\boldsymbol{\gamma}) = \sum_{k \in S_A} \boldsymbol{h}\left(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}\right) - \sum_{k \in S_B} d_k^B \pi\left(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}\right) \boldsymbol{h}\left(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}\right), \tag{5}$$

and when we let $\boldsymbol{h}(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}) = \boldsymbol{x}_k^\circ/\pi(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma})$, then (5) reduces to calibration equations:

$$\boldsymbol{G}(\boldsymbol{\gamma}) = \sum_{k \in S_A} \frac{\boldsymbol{x}_k^\circ}{\pi\left(\boldsymbol{x}_k^\circ, \boldsymbol{\gamma}\right)} - \sum_{k \in S_B} d_k^B \boldsymbol{x}_k^\circ. \tag{6}$$

In this setting unit-level data are not required to get $\sum_{k \in S_B} d_k^B \boldsymbol{x}_k^\circ$. These sums can be replaced with known or estimated population totals from reference probability survey. Kim and Riddles (2012) showed that estimator (6) leads to optimal estimation when a linear regression model holds for y and $\boldsymbol{x}^\circ$.

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

$$\hat{\bar{\tau}}_{\text{IPW1}} = \frac{1}{N} \sum_{k \in S_A} \frac{y_k}{\hat{\pi}_k^A} \quad \text{and} \quad \hat{\bar{\tau}}_{\text{IPW2}} = \frac{1}{\hat{N}_A} \sum_{k \in S_A} \frac{y_k}{\hat{\pi}_k^A}, \tag{7}$$

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 distribution of auxiliary variables. To overcome this limitation we use the idea of 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 either $\boldsymbol{x}_k^*$ through $\boldsymbol{a}_k$ ($\pi_k^A = \pi(\boldsymbol{x}_k^*, \boldsymbol{\gamma}) = \pi(\boldsymbol{a}_k, \boldsymbol{\gamma})$) or both $\boldsymbol{x}_k^\circ$ and $\boldsymbol{x}_k^*$ ($\pi_k^A = \pi(\boldsymbol{x}_k^\circ, \boldsymbol{x}_k^*, \boldsymbol{\gamma})$. For simplicity we denote it as $\pi(\boldsymbol{x}_k, \boldsymbol{\gamma})$). Inclusion of $\boldsymbol{a}_k$ 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 offers arguments for the inclusion of higher moments for $\boldsymbol{x}^\circ$ auxiliary variables (Ai, Linton & Zhang, 2020; Imai & 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 et al., 2022).

Our approach is not limited to 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 (4), which takes the following form when only $\boldsymbol{x}^*$ is used

$$\boldsymbol{U}(\boldsymbol{\gamma})=\sum_{k\in S_A}\boldsymbol{a}_k-\sum_{k\in S_B}d_k^B\pi\left(\boldsymbol{a}_k,\boldsymbol{\gamma}\right)\boldsymbol{a}_k,$$

or is calculated as follows if both $\boldsymbol{x}_k^\circ$ and $\boldsymbol{a}_k$ are used

$$\boldsymbol{U}(\boldsymbol{\gamma})=\begin{pmatrix}\sum_{k\in S_A}\boldsymbol{x}_k^\circ-\sum_{k\in S_B}d_k^B\pi\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)\boldsymbol{x}_k^\circ\\ \sum_{k\in S_A}\boldsymbol{a}_k-\sum_{k\in S_B}d_k^B\pi\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)\boldsymbol{a}_k\end{pmatrix}.$$

Note that this procedure does not allow to reproduce quantiles for the $\boldsymbol{x}^\circ$ variables but it reduces the discrepancies between distributions of S_A and S_B. In the simulation study we show that this is in particular 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 (5) which allow to reproduce totals and quantiles from the probability sample S_B. In this situation $\boldsymbol{G}(\boldsymbol{\gamma})$ changes either to

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

or to

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\gamma})&=\sum_{k\in S_A}\boldsymbol{h}\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)-\sum_{k\in S_B}d_k^B\pi\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)\boldsymbol{h}\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)\\&=\sum_{k\in S_A}\boldsymbol{h}\left(\boldsymbol{x}_k,\boldsymbol{\gamma}\right)-\sum_{k\in S_B}d_k^B\pi\left(\boldsymbol{x}_k,\boldsymbol{\gamma}\right)\boldsymbol{h}\left(\boldsymbol{x}_k,\boldsymbol{\gamma}\right).\end{aligned}$$

If we assume logistic regression for $\pi\left(\boldsymbol{a}_k,\boldsymbol{\gamma}\right)$ or $\pi\left(\boldsymbol{x}_k,\boldsymbol{\gamma}\right)$ and $\boldsymbol{h}(\boldsymbol{a}_k,\boldsymbol{\gamma})=\boldsymbol{a}_k/\pi(\boldsymbol{a}_k,\boldsymbol{\gamma})$ or $\boldsymbol{h}(\boldsymbol{x}_k,\boldsymbol{\gamma})=\boldsymbol{x}_k/\pi(\boldsymbol{x}_k,\boldsymbol{\gamma})$ then $\boldsymbol{G}(\boldsymbol{\gamma})$ reduces to

$$\boldsymbol{G}(\boldsymbol{\gamma})=\begin{pmatrix}\sum_{k\in S_A}\frac{\boldsymbol{x}_k^\circ}{\pi\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)}-\sum_{k\in S_B}d_k^B\boldsymbol{x}_k^\circ\\ \sum_{k\in S_A}\frac{\boldsymbol{a}_k}{\pi\left(\boldsymbol{x}_k^\circ,\boldsymbol{a}_k,\boldsymbol{\gamma}\right)}-\sum_{k\in S_B}d_k^B\boldsymbol{a}_k\end{pmatrix}=\sum_{k\in S_A}\frac{\boldsymbol{x}_k}{\pi\left(\boldsymbol{x}_k,\boldsymbol{\gamma}\right)}-\sum_{k\in S_B}d_k^B\boldsymbol{x}_k,\tag{8}$$

where $\boldsymbol{x}_k=\binom{\boldsymbol{x}_k^\circ}{\boldsymbol{a}_k}$. This allows to estimate $\boldsymbol{\gamma}$ parameters so the population size N, estimated totals $\hat{\tau}_{\boldsymbol{x}^\circ}$ (or totals $\tau_{\boldsymbol{x}^\circ}$ if are known) or/and $\boldsymbol{Q}_{x^*,\alpha}$ are reproduced. In the Section A in the Appendix we state conditions which guarantee existence and uniqueness of solutions for $\boldsymbol{U}(\boldsymbol{\gamma})=\mathbf{0}$ and $\boldsymbol{G}(\boldsymbol{\gamma})=\mathbf{0}$. The interpretation of result from the Appendix is that long as the matrices $(\boldsymbol{a}_k)_{k\in S_A}$, $(\boldsymbol{a}_k)_{k\in S_B}$, $(\boldsymbol{x}_k^\circ)_{k\in S_A}$ and $(\boldsymbol{x}_k^\circ)_{k\in S_B}$ are full rank and $S_A \neq U$ than a unique solution will exist.

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

Remark 2. *The use of $\boldsymbol{a}_k$ in (8) is related to the use of piecewise (constant) 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*

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

Thus, we use the piecewise method to approximate the non-linear relationship between y and $\boldsymbol{x}$. Consider the following example: generate $n = 1000$ observations of $X \sim \text{Uniform}(0, 80)$, $Y_1 \sim N(1300 - (X-40)^2, 300)$ and $p = P(Y_2 = 1) = \text{logit}(-3 + (X-1.5)^2 + N(0, 0.5))$ so we get (y_k, p_k, x_k) pairs for $k = 1, ..., n$. Then, let x_k be an auxiliary variable in the following settings: 1) x_k is used as x_k° and 2) x_k is used as x_k^* where quartiles or deciles are specified, and 3) x_k is used as x_k^* and x_k°. Figure 1 presents how the prediction of Y_1 (first row) and $P(Y_2 = 1)$ (second row) changes when x_k^* or x_k° or x_k^* and x_k° together are included. 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 (8) is actually doubly robust i.e. it is asymptotically unbiased if either 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 reduce bias when two models are non-linear.*

After estimation of $\boldsymbol{\gamma}$ parameters we can estimate the population mean using either $\hat{\bar{\tau}}_{IPW1}$ or $\hat{\bar{\tau}}_{IPW2}$ given in equation (7) which we denote as $\hat{\bar{\tau}}_{QBIPW1}$ or $\hat{\bar{\tau}}_{QBIPW2}$ respectively. To summarize our proposed approach:

- estimated propensity scores will reproduce either known or estimated moments as well as

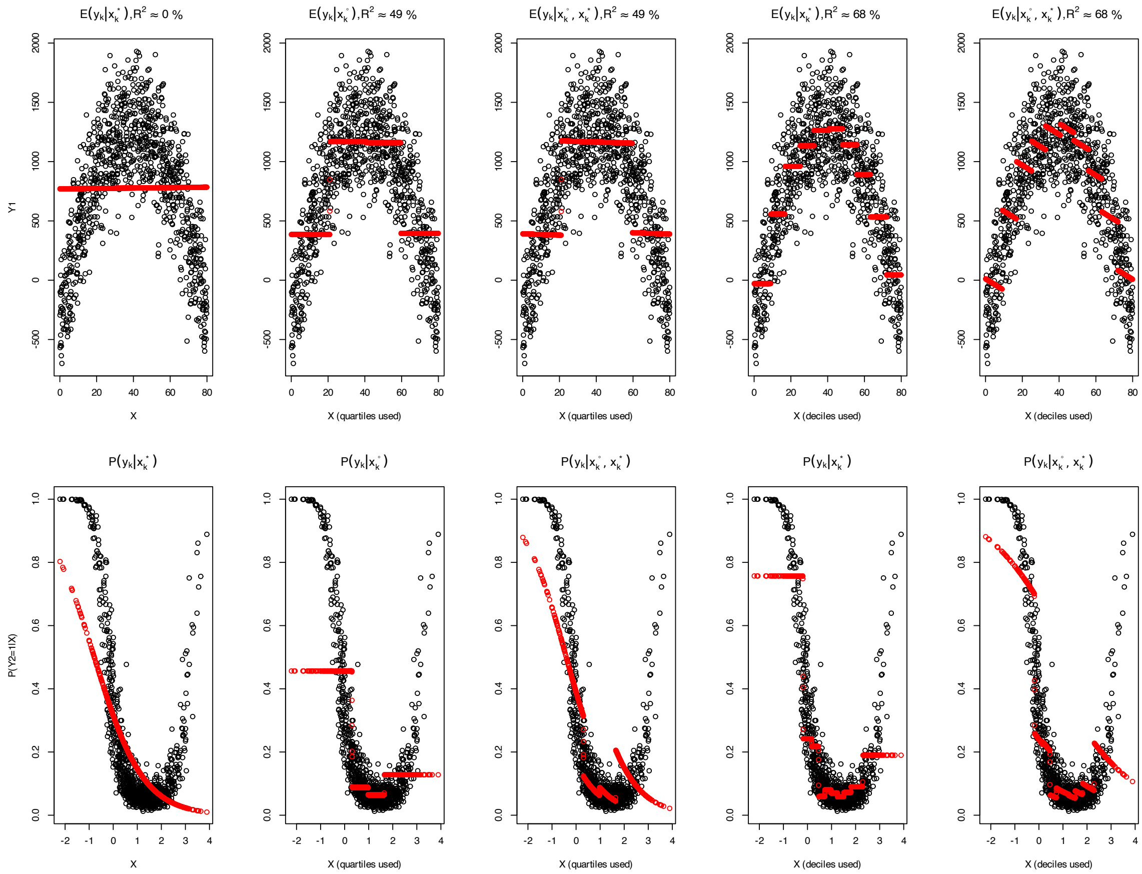

Figure 1: Prediction of Y_1 (top row) and $\mathrm{P}(Y_2 = 1)$ (bottom row) based on x_k^* or x_k° or x_k^* and x_k° together from the example in Remark 2. Black circles denote observed y_k, red circles denote predictions $\hat{m}_k$ based on the linear model and $\hat{p}_k$ based on logistic regression

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 approximate relationship between inclusion R and $\boldsymbol{x}^*$. 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 & 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 $\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = \mathbf{0}$ (Wu, 2022, eq. 6.1) by imposing constraints on quantiles. In the case when both $\boldsymbol{x}_k^\circ$ and $\boldsymbol{x}_k^*$ are used and $\boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{\gamma}) = \frac{\boldsymbol{x}_k}{\pi_k(\boldsymbol{x}_k, \boldsymbol{\gamma})}$, the equation is given as:

$$\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = N^{-1} \begin{pmatrix} \sum_{k \in U} R_k \frac{y_i - \mu}{\pi(\boldsymbol{x}_k^\circ, \boldsymbol{a}_k, \boldsymbol{\gamma})} \\ \sum_{k \in U} R_k \frac{\boldsymbol{x}_k^\circ}{\pi(\boldsymbol{x}_k^\circ, \boldsymbol{a}_k, \boldsymbol{\gamma})} - \sum_{k \in S_B} d_k^B \boldsymbol{x}_k^\circ \\ \sum_{k \in U} R_k \frac{\boldsymbol{a}_k}{\pi(\boldsymbol{x}_k^\circ, \boldsymbol{a}_k, \boldsymbol{\gamma})} - \sum_{k \in S_B} d_k^B \boldsymbol{a}_k \end{pmatrix}. \tag{9}$$

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

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

where $\phi_n(\boldsymbol{\eta}) = \partial \boldsymbol{\Phi}_n(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ depends on the form of π_k^A and $\boldsymbol{h}$. 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 $\boldsymbol{a}_k$ or $\boldsymbol{x}_k$, will lead to higher variance as the number of variables can be significantly larger than $\boldsymbol{x}_k^\circ$ as adding new variables inflates 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 $\mathcal{F}_N = \{\boldsymbol{x}_k = (x_{k1}, x_{k2}), \boldsymbol{y}_k = (y_{k1}, y_{k2}) : k = 1, \ldots, N\}$ 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 $x_{k2} \sim \mathrm{Exp}(1)$. 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 $N^{-1} \sum_{k=1}^{N} y_{kj}$, $j = 1, 2$.

The finite population $\mathcal{F}_N$ 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}.

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)	$P(y_{k1} = 1 \vert x_{k1}, x_{k2}; \alpha_k) = \text{logit}(1 + x_{k1} + x_{k2} + \alpha_k)$
	non-linear (OM4)	$P(y_{k2} = 1 \vert x_{k1}, x_{k2}; \alpha_k) = \text{logit}\left\{0.5(x_{k1} - 1.5)^2 + x_{k2}^2 + \alpha_k\right\}$
Selection (S_A)	linear (PM1)	$\text{logit}(p_k) = x_{k2}$
	non-linear (PM2)	$\text{logit}(p_k) = -3 + (x_{k1} - 1.5)^2 + (x_{k2} - 2)^2$

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

Simulation results are presented in Tables 5.2–5.3 and show also the empirical coverage (CR) of confidence intervals based on analytic variance estimators for the IPW estimators. In Table B.1 in the Appendix information about the quality of reproducing totals and quantiles is also presented.

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 as $\boldsymbol{x}^\circ$ only (GLM),
- Doubly robust estimator using x_1, x_2 as $\boldsymbol{x}^\circ$ only with two variants of IPW: MLE (DR MLE) and GEE (calibrated; DR GEE),

- IPW MLE and IPW GEE with the following variants: estimated totals for x_1, x_2 only,
- QBIPW MLE and GEE with the following variants:
  - estimated quartiles for x_1, x_2 only (QBIPW1A),
  - estimated quartiles and totals for x_1, x_2 (QBIPW1B),
  - estimated deciles for x_1, x_2 only (QBIPW2A),
  - estimated deciles and totals for x_1, x_2 (QBIPW2B).

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. However, the QBIPW estimators utilising quartiles or deciles alone exhibit a slight deviation from the linear relationship. 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.

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 QBIPW2B (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 QBIPW2B (GEE) achieves the

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

	Scenario I			Scenario II			Scenario III			Scenario IV		
OM	linear (1)			linear (1)			non-linear (2)			non-linear (2)		
PM	linear (1)			non-linear (2)			linear (1)			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 MLE	-0.01	5.70	5.70	69.73	27.50	75.00	0.42	9.30	9.40	467.97	175.10	499.60
QBIPW1A	1.81	4.40	4.80	24.23	10.20	26.30	11.82	7.70	14.10	169.68	31.30	172.50
QBIPW1B	-0.02	5.70	5.70	6.58	11.70	13.50	0.44	9.60	9.60	29.93	22.20	37.30
QBIPW2A	0.97	5.20	5.30	18.98	8.90	21.00	6.86	9.50	11.70	122.39	25.30	125.00
QBIPW2B	-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 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
QBIPW1A	2.16	3.90	4.40	-30.56	0.80	30.60	9.73	8.50	12.90	52.81	2.10	52.90
QBIPW1B	0.33	4.50	4.50	0.38	4.60	4.60	2.61	9.70	10.00	27.84	12.50	30.50
QBIPW2A	1.22	3.90	4.10	-32.20	1.10	32.20	4.76	9.10	10.20	37.13	2.00	37.20
QBIPW2B	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 MLE	-0.02	0.47	0.47	0.37	0.70	0.79	-0.02	0.77	0.77	2.32	1.31	2.66
QBIPW1A	0.01	0.37	0.37	-0.32	0.40	0.51	0.08	0.53	0.54	1.17	0.98	1.53
QBIPW1B	-0.02	0.47	0.47	-0.54	0.60	0.81	-0.02	0.75	0.75	-0.06	0.75	0.76
QBIPW2A	-0.00	0.45	0.45	0.11	0.45	0.46	-0.00	0.59	0.59	-0.17	0.69	0.71
QBIPW2B	-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 GEE	0.01	0.39	0.39	-1.86	0.23	1.87	0.03	0.61	0.61	-0.88	0.28	0.92
QBIPW1A	0.08	0.32	0.33	-3.35	0.08	3.35	0.16	0.50	0.53	-1.20	0.13	1.21
QBIPW1B	-0.01	0.37	0.37	-0.54	0.25	0.60	-0.02	0.58	0.58	0.09	0.52	0.52
QBIPW2A	0.06	0.33	0.34	-3.25	0.11	3.25	0.10	0.50	0.51	-1.39	0.15	1.40
QBIPW2B	0.03	0.35	0.35	-0.11	0.27	0.29	0.05	0.52	0.53	-0.05	0.57	0.57

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 lower 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 QBIPW2B (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. QBIPW2B (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 QBIPW2B 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. QBIPW2B 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, with the exception of instances where only quartiles or deciles were employed (e.g., QBIPW1A, QBIPW2A). 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 QBIPW2B (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.

In Scenario III, the proposed approaches yielded results that were close to the nominal rates, with the exception of the QBIPW1A and QBIPW2A, which exhibited excellent performance. Finally, with respect to the Scenario IV, it can be observed that when both models

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

	OM	Scenario I linear (1)		Scenario II linear (1)		Scenario III non-linear (2)		Scenario IV non-linear (2)	
	PM	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
	QBIPW1A	92.80	17.59	18.20	34.75	84.20	38.47	0.00	100.83
	QBIPW1B	94.80	21.67	98.20	47.09	100.00	45.92	83.40	92.19
	QBIPW2A	92.00	19.01	72.00	43.19	98.40	47.94	1.20	170.79
	QBIPW2B	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
	QBIPW1A	92.80	16.01	0.00	19.98	86.20	37.13	0.00	63.37
	QBIPW1B	93.80	17.69	94.80	18.35	95.40	39.57	45.60	51.73
	QBIPW2A	70.80	12.09	0.00	19.15	71.00	32.24	0.00	63.74
	QBIPW2B	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
	QBIPW1A	95.00	1.43	83.40	1.59	96.80	2.21	60.00	2.58
	QBIPW1B	94.40	1.74	75.60	2.29	93.00	2.84	95.80	3.02
	QBIPW2A	93.00	1.55	98.40	1.96	95.40	2.36	98.60	3.65
	QBIPW2B	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
	QBIPW1A	93.80	1.26	0.00	1.32	94.40	2.09	3.60	1.99
	QBIPW1B	94.20	1.43	49.60	1.08	96.20	2.37	96.40	2.24
	QBIPW2A	75.80	1.00	0.00	1.32	73.80	1.62	0.60	2.01
	QBIPW2B	94.20	1.44	93.80	1.07	96.80	2.40	96.40	2.36

are mis-specified, only those incorporating deciles and totals (along with the NN) in the case of continuous data and QBIPW1B (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. 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 24 February 2022, around 3.5 million persons (mainly women and children) arrived in Poland between 24 February and mid-May 2022 (Duszczyk & 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 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 31, 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 via 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: 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

Table 6.1 shows the distribution of registered employment according to the JVS survey, 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 6, while for CBOP and CBOP-UKR – 20 and 18 employees, respectively. This suggests that companies are selected into the CBOP depending on their size.

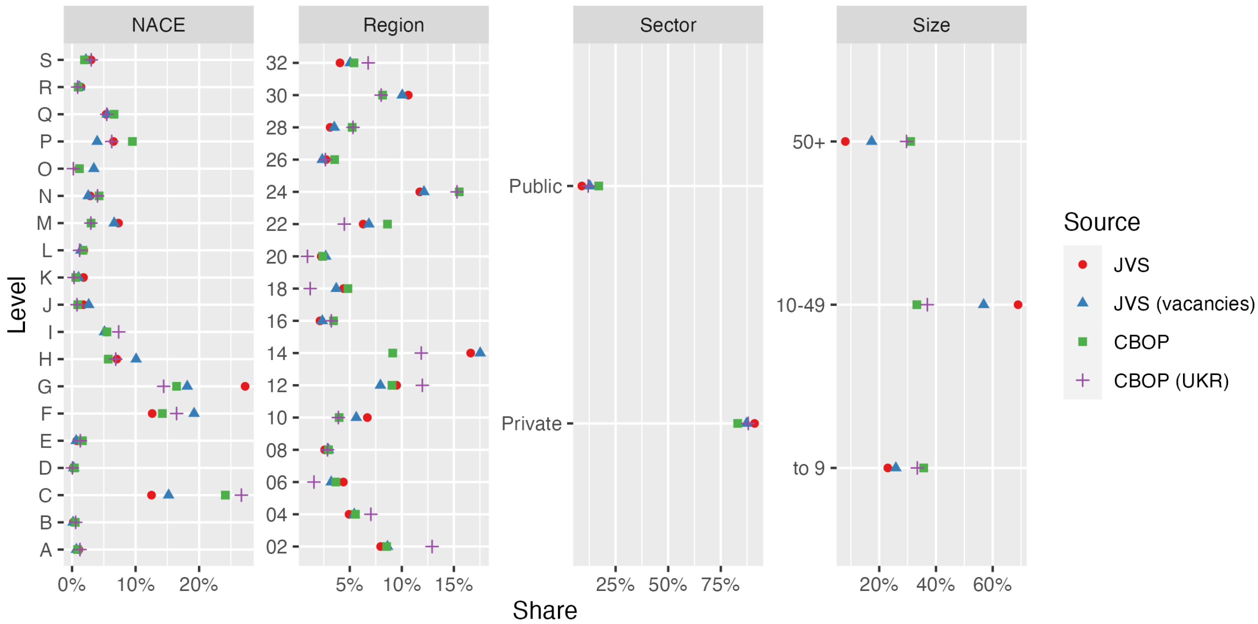

Figure 2: The distribution of 4 auxiliary variables in the sources at the end of 2022Q2. JVS: the total sample, JVS (vacancies): companies with at least one vacancy, CBOP: register data set, CBOP-UKR: units with at least one vacancy translated into Ukrainian

The distributions of four categorical variables in the two sources are shown in Figure 2. 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 3 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).

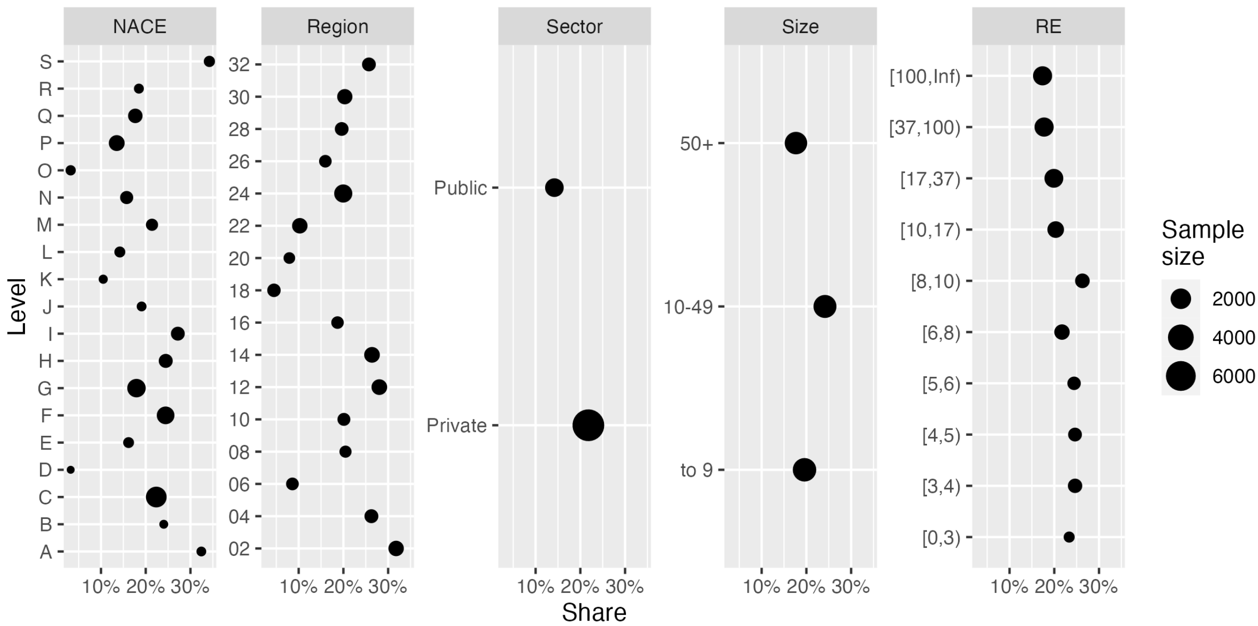

Figure 3: The target variable (the share of ads directed at Ukrainians) depending on 5 auxiliary variables at the end of 2022Q2

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), $I(\# \text{ vacancies} = 1)$ (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 as the $\boldsymbol{x}^*$ variable, 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(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).

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.

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

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.

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 vacancies aimed at Ukrainian workers, demonstrates the effectiveness of the proposed methods in producing estimates of high precision.

As regards 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 sample 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 & 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.

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 at https://github.com/ncn-foreigners/nonprobsvy. We also thank Natalie Shlomo for 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.

References

Ai, C., Linton, O., & 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.

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., & 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., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021.

Chen, Y., Li, P., & Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Survey Methodology*, *49*(2), 497–515.

Chrostowski, Ł., & 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), 137–162.

Daas, P. J., Puts, M. J., Buelens, B., & 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., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM.

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

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

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

Gelman, A. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, *23*, 127.

Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, *32*(1), 37–52.

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., & 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., & Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, *38*(2), 157–165.

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

Park, S., Kim, J. K., & Kim, K. (2019). A note on propensity score weighting method using paradata in survey sampling. *Survey Methodology*, *45*(3), 451–463.

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., & 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*, 283–311.

Yang, S., Kim, J.-K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, *47*, 29–58.

Appendix for the paper *Quantile balancing inverse probability weighting for non-probability samples*

A Assumptions that guarantee unique solution to estimating equations

A.1 Only for quantiles

We prove the result only for logistic regression. In the cases of cloglog and probit regression the proof is similar. We consider the following assumptions:

(A1) The matrix $\sum_{k\in S_A} \boldsymbol{a}_k\boldsymbol{a}_k^T$ is positive define.

(A2) The matrix $\sum_{k\in S_B} \boldsymbol{a}_k\boldsymbol{a}_k^T$ is positive define.

(A3) We require the inequality $\sum_{k\in S_B} d_k^B\boldsymbol{a}_k > \sum_{k\in S_A} \boldsymbol{a}_k$, with $>$ being understood as pairwise comparison, to be satisfied.

Conditions (A1) and (A2) 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. (A3) will be satisfied if $n_A < N$ since elements of $\boldsymbol{a}$ are non-negative and usage d_k^B allow for unbiased estimation of population totals and quantiles. Notice that (A1) implies $n_A \geq J_2$ where J_2 is the dimension of $\boldsymbol{a}$ and that there are at least J_2 linearly independent $\boldsymbol{a}$'s in S_A. (A2) implies analogous properties of S_B.

As a slight abuse of notation we write $\pi\left(\boldsymbol{\gamma}^T\boldsymbol{a}_k\right) := \pi(\boldsymbol{\gamma},\boldsymbol{a}_k)$.

Theorem 1. *Under (A1), (A2), (A3) the solution to the GEE with only quantiles:*

$$\boldsymbol{G}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \frac{\boldsymbol{a}_k}{\pi\left(\boldsymbol{\gamma}^T\boldsymbol{a}_k\right)} - \sum_{k\in S_B} d_k^B\boldsymbol{a}_k = \boldsymbol{0},$$

exists and is unique, and under (A1), (A2) the same holds for:

$$\boldsymbol{U}(\boldsymbol{\gamma}) = \sum_{k\in S_A} \boldsymbol{a}_k - \sum_{k\in S_B} d_k^B\boldsymbol{a}_k\pi\left(\boldsymbol{\gamma}^T\boldsymbol{a}_k\right) = \boldsymbol{0}.$$

Proof. We will prove that ℓ^* (the anti-derivative of $\boldsymbol{U}$) has a unique global maximum. Notice that if (A2) holds the hessian of ℓ^*:

$$\frac{\partial \ell^*}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}(\boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \boldsymbol{U}(\boldsymbol{\gamma}) = -\sum_{k \in S_B} d_k^B \pi' \left(\boldsymbol{\gamma}^T \boldsymbol{a}_k\right) \boldsymbol{a}_k \boldsymbol{a}_k^T,$$

is positive define for any $\boldsymbol{\gamma} \in \mathbb{R}^{J_2}$ 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 $\boldsymbol{U}(\boldsymbol{\gamma})$ has at most one root.

Strict concavity of ℓ^* implies that a sufficient condition for existence of a root of $\boldsymbol{U}$ is:

$$\lim_{\|\gamma\| \to \infty} |\ell^*(\boldsymbol{\gamma})| = \infty, \text{ which is equivalent to: } \lim_{\|\gamma\| \to \infty} \ell^*(\boldsymbol{\gamma}) = -\infty,$$

where $\|\cdot\|$ denotes the euclidean norm $\boldsymbol{\zeta} \overset{\|\cdot\|}{\mapsto} \sqrt{\left|\boldsymbol{\zeta}^T \boldsymbol{\zeta}\right|}$. To prove this fact it is sufficient to prove that $\lim_{m \to \infty} |\ell^*(\boldsymbol{\gamma}_m)| = \infty$ for sequences $(\boldsymbol{\gamma}_m)_{m \in \mathbb{N}} \subseteq \mathbb{R}^{J_2}$ with $\lim_{m \to \infty} \|\boldsymbol{\gamma}_m\| = \infty$, such that the limit $\lim_{m \to \infty} \operatorname{sgn}(\boldsymbol{\gamma}_m) = \lim_{m \to \infty} \frac{\boldsymbol{\gamma}_m}{\|\boldsymbol{\gamma}_m\|}$ exists.

For sequences defined above it is clear that for any $\boldsymbol{\zeta}$ we have either that $\boldsymbol{\zeta}^T \boldsymbol{\gamma}_m$ is bounded, $\boldsymbol{\zeta}^T \boldsymbol{\gamma}_m \to \infty$ or $\boldsymbol{\zeta}^T \boldsymbol{\gamma}_m \to -\infty$ as $m \to \infty$.

Now notice that (A1) implies, by rank-nullity theorem, that with probability 1 we have:

$$\operatorname{Span}\left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n_A}\right)^{\perp} = \left(\mathbb{R}^{J_2}\right)^{\perp} = \{\mathbf{0}\},$$

where $(\cdot)^{\perp}$ denotes the orthogonal component, which implies that:

$$\exists j \in S_A : |\boldsymbol{a}_j^T \boldsymbol{\gamma}_m| \to \infty,$$

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

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

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

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

$$H(\boldsymbol{\gamma}) = \sum_{k \in S_A} \int_0^{\boldsymbol{\gamma}^T \boldsymbol{a}_k} \frac{1}{\pi(t)} dt - \sum_{k \in S_B} d_k^B \boldsymbol{\gamma}^T \boldsymbol{a}_k,$$

the hessian of which is:

$$\frac{\partial}{\partial \boldsymbol{\gamma}} \boldsymbol{G}(\boldsymbol{\gamma}) = - \sum_{k \in S_A} \frac{\boldsymbol{a}_k \boldsymbol{a}_k^T}{\pi(\boldsymbol{a}_k, \boldsymbol{\gamma})^2} \pi'(\boldsymbol{a}_k, \boldsymbol{\gamma}),$$

guarantees that $\boldsymbol{G}$ has at most one root. Now at least one of the two following cases must happen:

- $\exists j \in S_A : \boldsymbol{\gamma}_m^T \boldsymbol{a}_j \to -\infty$ we have that the dominating term $\int_0^{\boldsymbol{\gamma}_m^T \boldsymbol{a}_j} \frac{1}{\pi(t)} dt$ decays to $-\infty$ exactly as fast as $-\exp\left(-\boldsymbol{\gamma}_m^T \boldsymbol{a}_j\right)$ so $H(\boldsymbol{\gamma}_m) \to -\infty$.
- $\forall j \in S_A : \boldsymbol{\gamma}_m^T \boldsymbol{a}_j \to \infty$ or $\boldsymbol{\gamma}_m^T \boldsymbol{a}_j$ is bounded and in this case by (A3) and $\int_0^x \frac{1}{\pi(t)} dt = O(x+1)$ as $x \to \infty$ (which is easily checked by L'Hôpital's rule), therefore we have:

  $$H(\boldsymbol{\gamma}_m) \leq \boldsymbol{\gamma}_m^T \left(\sum_{k \in S_A} \boldsymbol{a}_k - \sum_{k \in S_B} d_k^B \boldsymbol{a}_k \right) + n_A \to -\infty,$$

  where divergence to $-\infty$ follows from $\boldsymbol{\gamma}_m^T \boldsymbol{a}_k \geq 0$ for those units.

Lastly since H is clearly locally bounded and diverges to $-\infty$ at $\|\cdot\| = \infty$ it follows that H is bounded from above. $\square$

A.2 For quantiles and variables

(B1) The matrix $\sum_{k \in S_A} \boldsymbol{x}_k \boldsymbol{x}_k^T$ is positive define.

(B2) The matrix $\sum_{k \in S_B} \boldsymbol{x}_k \boldsymbol{x}_k^T$ is positive define.

(B3) The function H defined only in terms of $\boldsymbol{x}^\circ$ i.e:

$$H^\circ(\boldsymbol{\gamma}) = \sum_{k \in S_A} \int_0^{\boldsymbol{\gamma}^T \boldsymbol{x}_k^\circ} \frac{1}{\pi(t)} dt - \sum_{k \in S_B} d_k^B \boldsymbol{\gamma}^T \boldsymbol{x}_k^\circ,$$

satisfies $\limsup_{\|\boldsymbol{\gamma}\| \to \infty} H^\circ(\boldsymbol{\gamma}) = -\infty$.

The assumption (B3) is also a sufficient condition for existence of a unique solution for $\boldsymbol{G} = \boldsymbol{0}$ when considering only $\boldsymbol{x}^\circ$ variables.

Theorem 2. *Under assumptions (B1), (B2), (B3), (A3) the solutions to:*

$$\boldsymbol{G}(\boldsymbol{\gamma}) = \sum_{k \in S_A} \frac{\boldsymbol{x}_k}{\pi\left(\boldsymbol{x}_k, \boldsymbol{\gamma}\right)} - \sum_{k \in S_B} d_k^B \boldsymbol{x}_k = \boldsymbol{0},$$

and under (B1), (B2) to:

$$\boldsymbol{U}(\boldsymbol{\gamma}) = \sum_{k \in S_A} \boldsymbol{x}_k - \sum_{k \in S_B} d_k^B \pi\left(\boldsymbol{x}_k, \boldsymbol{\gamma}\right) \boldsymbol{x}_k = \boldsymbol{0},$$

for defined vectors $\boldsymbol{x}_k = \left(\begin{smallmatrix} \boldsymbol{x}_k^\circ \\ \boldsymbol{a}_k \end{smallmatrix}\right)$ *exist and are unique.*

Proof. Under assumed conditions it is clear that if a solution exists then it is unique. The proof of existence of solution to $\boldsymbol{U} = \boldsymbol{0}$ is exactly the same as in the case of only $\boldsymbol{a}$.

Now for H function the case where $\exists j \in S_A : \boldsymbol{\gamma}_m^T \boldsymbol{x}_j \to -\infty$ is exactly analogous to the proof of 1 (for the sequences $\boldsymbol{\gamma}_m$ defined earlier). If on the other-hand we have $\forall j \in S_A : \boldsymbol{\gamma}_m^T \boldsymbol{x}_j \to \infty$ or $\boldsymbol{\gamma}_m^T \boldsymbol{x}_j$ is bounded then:

- If $\lim_{m \to \infty} \|(\boldsymbol{\gamma}_2)_m\| = \infty$ and $\limsup_{m \to \infty} \|(\boldsymbol{\gamma}_1)_m\| < \infty$ then for large m and all $k \in S_A$ we have either of the two possibilities (depending on whether $\boldsymbol{\gamma}_m^T \boldsymbol{x}_j$ is bounded):

$$\left| \int_0^{\boldsymbol{\gamma}_m^T \boldsymbol{x}_k} \frac{1}{\pi(t)} dt \right| < C + \boldsymbol{\gamma}_m^T \boldsymbol{x}_k, \quad \int_0^{\boldsymbol{\gamma}_m^T \boldsymbol{x}_k} \frac{1}{\pi(t)} dt \leq \boldsymbol{\gamma}_m^T \boldsymbol{x}_k + 1,$$

  for some absolute positive constant C, so:

$$\begin{aligned} H(\boldsymbol{\gamma}_m) &\leq \boldsymbol{\gamma}_m^T \left(\sum_{k \in S_A} \boldsymbol{x}_k - \sum_{k \in S_B} d_k^B \boldsymbol{x}_k \right) + n_A(1 + C) \\ &= (\boldsymbol{\gamma}_2)_m^T \left(\sum_{k \in S_A} \boldsymbol{a}_k - \sum_{k \in S_B} d_k^B \boldsymbol{a}_k \right) + (\boldsymbol{\gamma}_1)_m^T \left(\sum_{k \in S_A} \boldsymbol{x}_k^\circ - \sum_{k \in S_B} d_k^B \boldsymbol{x}_k^\circ \right) + n_A(1 + C), \end{aligned}$$

  where the first term diverges to $-\infty$ and the rest is bounded.

- If $\lim_{m\to\infty} \|(\boldsymbol{\gamma}_1)_m\| = \infty$ and $\limsup_{m\to\infty} \|(\boldsymbol{\gamma}_2)_m\| < \infty$ then we have:

$$
\begin{aligned}
H(\boldsymbol{\gamma}_m) &= \sum_{k\in S_A} \left(\int_0^{(\gamma_1)_m^T \boldsymbol{x}_k^\circ} \frac{1}{\pi(t)} dt + \int_{(\gamma_1)_m^T \boldsymbol{x}_k^\circ}^{(\gamma_1)_m^T \boldsymbol{x}_k^\circ + (\gamma_2)_m^T \boldsymbol{a}_k} \frac{1}{\pi(t)} dt \right) \\
&\quad - \sum_{k\in S_B} d_k^B (\boldsymbol{\gamma}_2)_m^T \boldsymbol{a}_k - \sum_{k\in S_B} d_k^B (\boldsymbol{\gamma}_1)_m^T \boldsymbol{x}_k^\circ \\
&= H^\circ((\boldsymbol{\gamma}_1)_m) - \sum_{k\in S_B} d_k^B (\boldsymbol{\gamma}_2)_m^T \boldsymbol{a}_k + \sum_{k\in S_A} \int_{(\gamma_1)_m^T \boldsymbol{x}_k^\circ}^{(\gamma_1)_m^T \boldsymbol{x}_k^\circ + (\gamma_2)_m^T \boldsymbol{a}_k^\circ} \frac{1}{\pi(t)} dt \\
&\sim H^\circ((\boldsymbol{\gamma}_1)_m) + \sum_{k\in S_A} (\boldsymbol{\gamma}_2)_m^T \boldsymbol{a}_k^\circ - \sum_{k\in S_B} d_k^B (\boldsymbol{\gamma}_2)_m^T \boldsymbol{a}_k,
\end{aligned}
$$

  which follows from $\int_x^{x+z} \frac{1}{\pi(t)} dt \sim z$ as $x \to \infty$, and the expression tends to $-\infty$.

- If $\lim_{m\to\infty} \|(\boldsymbol{\gamma}_1)_m\| = \infty$ and $\lim_{m\to\infty} \|(\boldsymbol{\gamma}_2)_m\| = \infty$ then:

$$
\begin{aligned}
H(\boldsymbol{\gamma}_m) &\le \boldsymbol{\gamma}_m^T \left(\sum_{k\in S_A} \boldsymbol{x}_k - \sum_{k\in S_B} d_k^B \boldsymbol{x}_k \right) + \sum_{\substack{k\in S_A \\ \boldsymbol{\gamma}_m^T \boldsymbol{x}_k \to \infty}} \frac{\log\left(\boldsymbol{\gamma}_m^T \boldsymbol{x}_k\right) + 1}{\sqrt{\boldsymbol{\gamma}_m^T \boldsymbol{x}_k}} \\
&= (\boldsymbol{\gamma}_2)_m^T \left(\sum_{k\in S_A} \boldsymbol{a}_k - \sum_{k\in S_B} d_k^B \boldsymbol{a}_k \right) + (\boldsymbol{\gamma}_1)_m^T \left(\sum_{k\in S_A} \boldsymbol{x}_k^\circ - \sum_{k\in S_B} d_k^B \boldsymbol{x}_k^\circ \right) \\
&\quad + \sum_{\substack{k\in S_A \\ \boldsymbol{\gamma}_m^T \boldsymbol{x}_k \to \infty}} \frac{\log\left(\boldsymbol{\gamma}_m^T \boldsymbol{x}_k\right) + 1}{\sqrt{\boldsymbol{\gamma}_m^T \boldsymbol{x}_k}} \\
&\le (\boldsymbol{\gamma}_2)_m^T \left(\sum_{k\in S_A} \boldsymbol{a}_k - \sum_{k\in S_B} d_k^B \boldsymbol{a}_k \right) \\
&\quad + \sum_{k\in S_A} \int_0^{(\gamma_1)_m^T \boldsymbol{x}_k^\circ} \frac{1}{\pi(t)} dt - \sum_{k\in S_B} d_k^B \, (\boldsymbol{\gamma}_1)_m^T \, \boldsymbol{x}_k^\circ + \sum_{\substack{k\in S_A \\ \boldsymbol{\gamma}_m^T \boldsymbol{x}_k \to \infty}} \frac{\log\left(\boldsymbol{\gamma}_m^T \boldsymbol{x}_k\right) + 1}{\sqrt{\boldsymbol{\gamma}_m^T \boldsymbol{x}_k}},
\end{aligned}
$$

  where the first two terms diverge to $-\infty$ by virtue of (A3), (B3) respectively and the last term converges to 0 which completed the proof.

$\square$

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:

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

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

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

- Totals (for x_1 and x_2 jointly):

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

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 $\nu_{r,\hat{N}}$, $\nu_{r,\hat{\tau}}$, and $\nu_{r,\hat{Q}}$. 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 QBIPW1B and QBIPW2B achieving the lowest error rates. The proposed QBIPW1B and QBIPW2B 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. QBIPW1B accounts for quartiles, while QBIPW2B utilises deciles.

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

Estimator	PM1			PM2		
	$\nu_{r,\hat{N}}$	$\nu_{r,\hat{Q}}$	$\nu_{r,\hat{\tau}}$	$\nu_{r,\hat{N}}$	$\nu_{r,\hat{Q}}$	$\nu_{r,\hat{\tau}}$
Mean from simulations						
Naive	–	0.52	–	–	2.28	–
MLE						
IPW	658.81	0.16	1049.47	11779.12	2.98	78050.24
QBIPW1A	449.44	0.17	2720.53	2742.00	1.45	18665.68
QBIPW1B	726.83	0.16	1094.00	3403.08	0.90	13198.45
QBIPW2A	389.22	0.15	2587.69	2393.39	1.10	17879.56
QBIPW2B	1334.84	0.15	1884.13	7310.56	0.46	20054.55
GEE						
IPW	0.00	0.15	0.00	0.00	1.37	0.00
QBIPW1A	0.00	0.14	2410.60	0.01	1.25	22525.29
QBIPW1B	0.11	0.10	2.74	0.00	0.36	0.00
QBIPW2A	0.33	0.05	1459.76	0.01	1.19	23332.56
QBIPW2B	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	10617.02	2.76	70068.24
QBIPW1A	341.71	0.17	2648.90	2775.17	1.41	16464.82
QBIPW1B	353.72	0.15	732.87	3081.35	0.86	10515.02
QBIPW2A	249.74	0.14	2088.06	1812.43	1.12	16543.72
QBIPW2B	442.59	0.12	901.13	3242.52	0.41	8157.91
GEE						
IPW0	0.00	0.14	0.00	0.00	1.37	0.00
QBIPW1A	0.00	0.13	2281.27	0.01	1.25	22461.34
QBIPW1B	0.00	0.10	0.00	0.00	0.36	0.00
QBIPW2A	0.00	0.03	1157.90	0.01	1.19	23263.38
QBIPW2B	0.00	0.04	0.00	0.00	0.05	0.00

The results demonstrate a pronounced discrepancy in performance between the PM1 and PM2 scenarios, with all estimators exhibiting superior 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 QBIPW1B and QBIPW2B methods demonstrate remarkable resilience in quantile estimation, even under PM2 (with mean quantile errors as low as 0.05 for QBIPW2B), 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 QBIPW1B and QBIPW2B under PM2 serves to demonstrate their capacity to handle complex probability models by leveraging a more comprehensive set of population characteristics.