Lietuvos matematikos rinkinys
Proc. of the Lithuanian Mathematical Society, Ser. A
Vol. 64, 2023, pages 1–11
https://doi.org/10.15388/LMR.2023.33587

Vilnius University Press

On using a non-probability sample for the estimation of population parameters

Ieva Burakauskaitė, Andrius Čiginas

Institute of Data Science and Digital Technologies, Vilnius University
Akademijos str. 4, LT-08412 Vilnius, Lithuania
State Data Agency (Statistics Lithuania)
Gedimino aven. 29, LT-01500 Vilnius, Lithuania
E-mail: ieva.burakauskaite@mif.stud.vu.lt; andrius.ciginas@mif.vu.lt

Received July 10, 2023; published online November 20, 2023

Abstract. We aim to find a way to effectively integrate a non-probability (voluntary) sample under the data framework, where the study variable is also observed in a probability sample of some statistical survey. The selection bias that arises from voluntary participation in the survey is corrected by estimating the inclusion into the sample probabilities (propensity scores) for the units in the non-probability sample. The estimators for the propensity scores are constructed using a parametric logistic regression model. We consider two modeling scenarios: with an assumption that the willingness to participate in the voluntary survey does not depend on the survey variable itself and that such a variable does contribute to whether the individual responds or not. The maximum likelihood method is applied in both scenarios to estimate the propensity scores. The estimators of the population mean based on the estimated propensity scores are linearly combined with the unbiased estimator using the probability sample data. We compare the constructed estimators in the simulation study, where we estimate the population proportions using data from the Population and Housing Census surveys.

Keywords: data integration; not missing at random; propensity score adjustment; population census
AMS Subject Classification: 62D06

Introduction

The expanding access to alternative data sources and the decrease in response rates of probability surveys contribute to the continuing discussion on the statistical value

of these new data sources in the recent literature. It includes administrative data, Big Data and voluntary surveys, and may be described as non-probability samples. A look at the statistical analysis of non-probability samples and the challenges it might face was provided in [2], and more recently a critical review on the subject was given in [17]. As the selection mechanism for a unit to be included into a non-probability sample is unknown, such a sample does not represent the target population. Hence, selection bias is likely to occur if the estimators are constructed without a proper adjustment for the sampling process. A popular way to reduce the bias due to the use of a non-probability sample is to combine it with a probability sample that contains the same auxiliary variables. This approach can be viewed as data integration. Recently, a summarized introduction to data integration in survey sampling was provided in [7], as well as a comprehensive review of data integration for finite population inference was given in [9].

Probability sampling methods are an accepted approach to surveying finite populations in many areas of statistics, including the official statistics [15]. Probability samples are also being used in population censuses [1]. As it was recently demonstrated in [3] for the survey of the Lithuanian census, population censuses are also a potential area for an integration of non-probability and probability samples. State Data Agency (Statistics Lithuania) has conducted the Population and Housing Census 2021 primarily based on administrative data from state registers and information systems. However, additional information, that is not available in administrative sources, was collected through the voluntary statistical survey on population by ethnicity, native language and religion 2021. The latter data were combined with the probability sample data.

We consider the sampling framework where the survey sample consists of two parts: at first, data are collected as the non-probability (voluntary) sample, and the probability sample is drawn from the rest of the (census) population afterward. This framework is applied to the statistical survey of the Lithuanian census as described in Section 2.2. Such a scenario is similar to the one considered in [14, 6] but is different from that provided in [4], where the study variable is assumed to be unobserved in a probability sample. Additionally, we suppose that the complete auxiliary information is available, what is often the case in the official statistics. Then, assuming that the inclusion of the population unit into the non-probability sample does not depend on the study variable itself, the estimation procedures used in [4] can be simplified as in [3]. Differently from [3], we consider the situation when the selection mechanism is non-ignorable, that is, there is a dependence between the inclusion into the non-probability sample and the study variable. Modeling under this mechanism is typically complicated since the incomplete study variable has to be included as the covariate [16, 11]. However, in our case, the combination of non-probability and probability samples may be exploited. We estimate the population parameters by adapting the maximum likelihood approach from [11] to our framework.

A basic setup is presented in Section 1.1. In Section 1.2, a post-stratified generalized regression (calibrated) estimator of the population mean, combining the non-probability and probability samples, is applied similarly as in [6]. In Section 1.3, the alternative inverse probability weighted (hereinafter referred to as IPW) estimator, based on estimated inclusion probabilities (propensity scores) for the non-probability sample, is presented. The propensity scores are modeled parametrically, assuming the selection mechanism of the non-probability sample to be non-ignorable. Both the

post-stratified and IPW estimators are linearly combined into the composite estimator in Section 1.4. In the simulation study using Lithuanian census data, the proposed estimators are compared with the one considered in [4, 3], where the stronger assumption is imposed on the propensity score model, see Section 2. In Section 3, the most relevant findings and some future insights are outlined.

1 Methods

1.1 Sampling framework and auxiliary information

We consider the study variable y that might be continuous or binary with the fixed values $y_1, \dots, y_N$ in a finite population $\mathcal{U} = \{1, \dots, N\}$ of size N. Our goal is to estimate the population mean or proportion

$$\mu = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k. \tag{1}$$

In order to estimate parameter (1), we assume a sample s of size n to be collected in two steps.

1. At first, a non-probability sample s_A of size n_A is obtained from the survey population $\mathcal{U}$.
2. Then, a sample s_B of size n_B is drawn according to any probability sampling design without replacement from the rest of the survey population $\mathcal{U} \backslash \{s_A\}$.

We interpret that the sample $s = s_A \cup s_B$, $n = n_A + n_B$, is drawn according to the probability sampling design $p(\cdot)$ with the inclusion into the sample probabilities $\pi_k = P_p\{k \in s\} > 0$, $k \in \mathcal{U}$, where we set $\pi_k = 1$ if $k \in s_A$. Hereinafter P_p, E_p, and V_p denote the probability, expectation and variance, calculated according to the randomness induced by $p(\cdot)$, respectively.

We assume that the complete auxiliary information is available, that is, the values $\mathbf{x}_k$ of the auxiliary variables $\mathbf{x}$ are known for all units $k \in \mathcal{U}$. The relationship between the variables y and $\mathbf{x}$ is supposed to be described by a semiparametric outcome regression model ξ:

$$E_\xi(y_k|\mathbf{x}_k) = m(\mathbf{x}_k, \boldsymbol{\beta}) \quad \text{and} \quad V_\xi(y_k|\mathbf{x}_k) = v_k^2 \sigma^2, \quad k \in \mathcal{U}, \tag{2}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters, $v_k = v(\mathbf{x}_k)$ is a known function of $\mathbf{x}_k$, and $m(\mathbf{x}_k, \boldsymbol{\beta})$ has a known form as well. We take $m(\mathbf{x}_k, \boldsymbol{\beta}) = \mathbf{x}_k' \boldsymbol{\beta}$, where 1 is the first component of the vector $\mathbf{x}_k$ for all $k \in \mathcal{U}$, and, therefore, get a linear regression model. Here E_ξ and V_ξ denote the expectation and variance with respect to the model ξ.

1.2 Post-stratified generalized regression estimator

Let us consider the combined sample s with the accompanying probability sampling design $p(\cdot)$. Linear regression model (2) is used to build the generalized regression estimator [13]

$$\hat{\mu}^{\mathrm{GR}} = \frac{1}{N} \sum_{k \in s} d_k y_k + \left(\frac{1}{N} \sum_{k \in \mathcal{U}} \mathbf{x}_k - \frac{1}{N} \sum_{k \in s} d_k \mathbf{x}_k \right)' \widehat{\mathbf{B}} \tag{3}$$

of (1), where $d_k = 1/\pi_k$ denote the sampling weights and

$$\widehat{\mathbf{B}} = \left(\sum_{k\in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \sum_{k\in s} \frac{d_k \mathbf{x}_k y_k}{c_k}$$

with positive constants c_k, for instance, $c_k = v_k^2$.

Estimator (3) is equivalent to the calibrated estimator [5]

$$\hat{\mu}^{\mathrm{GR}} = \frac{1}{N} \sum_{k\in s} w_k y_k, \tag{4}$$

where the calibrated weights w_k, $k \in s$, are chosen to minimize the distance function

$$\sum_{k\in s} \frac{c_k (w_k - d_k)^2}{d_k}$$

subject to the calibration equations

$$\sum_{k\in s} w_k \mathbf{x}_k = \sum_{k\in \mathcal{U}} \mathbf{x}_k.$$

Estimator (3) is approximately design unbiased, that is, $E_p(\hat{\mu}^{\mathrm{GR}}) \approx \mu$. Generalized regression estimator (3) is also referred to as the post-stratified estimator in [6], with two post-strata, that is, s_A and $\mathcal{U}\backslash\{s_A\}$. According to [6], such estimation could prove to be efficient with a large non-probability sample.

According to [13], a design-consistent estimator of the variance $V_p(\hat{\mu}^{\mathrm{GR}})$ is

$$\hat{\psi}^{\mathrm{GR}} = \frac{1}{N^2} \sum_{k\in s} \sum_{l\in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{(y_k - \mathbf{x}_k' \widehat{\mathbf{B}})(y_l - \mathbf{x}_l' \widehat{\mathbf{B}})}{\pi_k \pi_l}, \tag{5}$$

where $\pi_{kl} = P_p\{k, l \in s\} > 0$ are the second-order inclusion into the sample probabilities.

1.3 Inverse probability weighted estimator

Let us consider only the non-probability sample s_A. As the selection mechanism for a unit to be included into the non-probability sample is unknown, such a sample itself does not represent the target population. Therefore, any naive estimator based on it is typically biased [8].

Let R_k be the selection indicator for a unit $k \in \mathcal{U}$, that is, $R_k = 1$ if $k \in s_A$, and $R_k = 0$ otherwise. The inclusion into the sample s_A can be described by the probabilities

$$\pi_k^A = E_q(R_k \,|\, \mathbf{x}_k, y_k) = P_q(R_k = 1 \,|\, \mathbf{x}_k, y_k), \quad k \in \mathcal{U}, \tag{6}$$

that are analogous to the inclusion into the sample probabilities π_k for probability samples, and are called the propensity scores. Here the subscript q refers to the propensity score model.

For the estimation of the propensity scores π_k^A, $k \in s_A$, the following assumptions are considered.

A1. The indicator R_k and the study variable y_k are not independent given the covariates $\mathbf{x}_k$.

A2. Every unit has a nonzero propensity score: $\pi_k^A > 0$ for all $k \in \mathcal{U}$.

A3. The indicators R_k and R_l are independent given the respective covariates $\mathbf{x}_k$ and $\mathbf{x}_l$, $k \neq l$.

Due to assumption A1, we have the non-ignorable selection mechanism, which is similar to the notion of "not missing at random" (hereinafter referred to as NMAR) used in missing data literature [12].

The propensity scores are often modeled using the parametric logistic regression model

$$\pi_k^A = \pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta}) = \frac{\exp(\tilde{\mathbf{x}}_k'\boldsymbol{\theta})}{1 + \exp(\tilde{\mathbf{x}}_k'\boldsymbol{\theta})}, \tag{7}$$

with the model parameter $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_m, \theta_{m+1})'$, and $\tilde{\mathbf{x}}_k = (1, x_{k1}, \dots, x_{km}, y_k)'$. Then, the propensity score estimates $\hat{\pi}_k^A$ are obtained from the maximum likelihood estimator $\hat{\pi}_k^A = \pi(\tilde{\mathbf{x}}_k, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ maximizes the estimated log-likelihood function

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{k \in s} w_k \left(R_k \log \left\{ \frac{\pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta})}{1 - \pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta})} \right\} + \log \left\{ 1 - \pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta}) \right\} \right) \\ &= \sum_{k \in s_A} w_k \tilde{\mathbf{x}}_k'\boldsymbol{\theta} - \sum_{k \in s} w_k \log \left\{ 1 + \exp(\tilde{\mathbf{x}}_k'\boldsymbol{\theta}) \right\}, \end{aligned}$$

analogous to [11]. Here w_k, $k \in s = s_A \cup s_B$, are the calibrated weights from (4). The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by applying the Newton–Raphson or some other iterative procedure to solve the score equations

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k \in s} w_k \left\{ R_k - \pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta}) \right\} \tilde{\mathbf{x}}_k = \mathbf{0}.$$

The estimated propensity scores $\hat{\pi}_k^A = \pi(\tilde{\mathbf{x}}_k, \hat{\boldsymbol{\theta}})$, $k \in s_A$, are then used to compute the IPW estimator

$$\hat{\mu}^{\mathrm{IPW}} = \frac{1}{\widehat{N}^A} \sum_{k \in s_A} \frac{y_k}{\hat{\pi}_k^A}, \quad \text{where } \widehat{N}^A = \sum_{k \in s_A} \frac{1}{\hat{\pi}_k^A}, \tag{8}$$

of population parameter (1). In order for estimator (8) to efficiently correct the selection bias, the propensity score model has to be well-specified.

The variance estimator $\widehat{V}^{\mathrm{IPW}}$ for (8) may be obtained by using resampling methods, for example, the bootstrap procedure from [10].

Remark 1. If assumption A1 is altered into a stricter one, that is, the variables R_k and y_k are assumed to be independent given the covariates $\mathbf{x}_k$, we have $\pi_k^A = P_q(R_k = 1 \,|\, \mathbf{x}_k, y_k) = P_q(R_k = 1 \,|\, \mathbf{x}_k)$ for all $k \in \mathcal{U}$. Then, the selection mechanism is called ignorable. Such a case is analyzed in [4, 3]. The methodology on the derivation of the respective IPW estimator $\hat{\mu}^{\mathrm{IPWi}}$ and its asymptotic variance estimator $\widehat{V}^{\mathrm{IPWi}}$ under the ignorable selection mechanism is provided in [3]. The latter study is an adaptation of [4] methodology to the framework with complete auxiliary information.

1.4 Composite estimators

In order to reduce the variance of estimators, design-based post-stratified estimator (3) is linearly combined with the model-based IPW estimator under the non-ignorable, and the ignorable selection mechanisms. That is, two composite estimators

$$\hat{\mu}^{\mathrm{C}} = \hat{\lambda}_1 \hat{\mu}^{\mathrm{GR}} + (1 - \hat{\lambda}_1)\hat{\mu}^{\mathrm{IPW}} \quad \text{with } \hat{\lambda}_1 = \frac{\widehat{V}^{\mathrm{IPW}}}{\hat{\psi}^{\mathrm{GR}} + \widehat{V}^{\mathrm{IPW}}}, \tag{9}$$

and

$$\hat{\mu}^{\mathrm{Ci}} = \hat{\lambda}_2 \hat{\mu}^{\mathrm{GR}} + (1 - \hat{\lambda}_2)\hat{\mu}^{\mathrm{IPWi}} \quad \text{with } \hat{\lambda}_2 = \frac{\widehat{V}^{\mathrm{IPWi}}}{\hat{\psi}^{\mathrm{GR}} + \widehat{V}^{\mathrm{IPWi}}} \tag{10}$$

of population parameter (1) are considered. Similar combinations are investigated in [14, 3]. Compositions (9) and (10) give more weight to the estimators with a smaller variance.

The respective variance estimators are

$$\widehat{V}^{\mathrm{C}} = \hat{\lambda}_1 \hat{\psi}^{\mathrm{GR}} \tag{11}$$

and

$$\widehat{V}^{\mathrm{Ci}} = \hat{\lambda}_2 \hat{\psi}^{\mathrm{GR}}, \tag{12}$$

and it is interpreted that the variances of the design-based estimators are reduced by the factors $\hat{\lambda}_1$ and $\hat{\lambda}_2$, respectively.

2 Application to the survey of the Lithuanian census

2.1 Motivation

With a new role as the governing organization for state data and the access to the unified database of the main state registers and information systems, Statistics Lithuania carried out the Population and Housing Census 2021 based on administrative data from these registers and information systems: Residents, Real Estate, Address registers, and the State Social Insurance Fund Board (SoDra) database, among others. However, as some variables of interest could not be obtained from any administrative source, the statistical survey on population by ethnicity, native language and religion was conducted in 2021. It aimed to evaluate population proportions for such variables as, for example, the 16 major religions professed in Lithuania. Statistics Lithuania had the complete data from previous censuses and additional auxiliary information, which led to the efficient estimation of the proportions of interest [3].

2.2 The union of voluntary and probability samples

The survey sample $s \subset \mathcal{U}$ was drawn in a few steps.

1. A voluntary online survey was launched at the beginning of 2021 and continued from the 15th of January to the 28th of February. This led to the collection of statistical data from approximately 2% of the census population (54 thousand respondents). It comprised the non-probability sample s_A.

2. After the end of the online survey, a sampling frame for probability sampling was constructed. It excluded such addresses as, for example, institutions, with at least one individual that already participated in the online survey, with more than 15 permanent residents, etc. Let s_O denote units that were not included in the sampling frame.
3. The probability sample s_B was drawn from the sampling frame $\mathcal{U}\backslash\{s_A \cup s_O\}$, which was divided into $H = 113$ strata according to the municipality intersected with the area of residence, that is, urban or rural. The number of addresses sampled from a particular stratum was proportional to the size of the stratum. During this step, a total of around 40 thousand addresses were sampled from the Population Register, what resulted in approximately 6% of the census population interviewed through the telephone survey (171 thousand respondents).

The working sampling design $p(\cdot)$ is characterized by the inclusion probabilities

$$\pi_k = \begin{cases} 1, & \text{if } k \in s_A \cup s_O, \\ m_k n'_h / N'_h, & \text{if } k \in s_B, \end{cases}$$

where N'_h denotes the size of the hth stratum, n'_h is the number of addresses selected, and m_k is the number of individuals in the corresponding address. The sample part s_O is treated as a separate post-stratum.

The response rate in the probability sample s_B was approximately 88%. Missing values in the sample $s = s_A \cup s_O \cup s_B$ were filled in using historical, deductive, and k-nearest neighbor imputation methods consecutively.

2.3 Simulation study

2.3.1 Framework

A few proportions of interest are of Roman Catholics and Evangelical Reformed Believers religious groups in 2021. In the simulation study, we focus on these particular religions.

Table 1. Comparison of proportions of some sociodemographic characteristics.

		Voluntary sample	Population	Difference in %
Ethnicity	Pole	0.35	0.07	441
Education	higher	0.48	0.20	134
County	Vilnius	0.64	0.29	121
Employment	employed	0.63	0.45	41
Age group	$\geqslant 30, < 50$	0.37	0.27	37
Marital status	married	0.52	0.42	25
Gender	male	0.41	0.46	−11

The analysis of the collected voluntary sample using completely observed sociodemographic characteristics leads to the finding that certain auxiliary variables can explain the chance of participating in the voluntary survey. A comparison of proportions of these sociodemographic characteristics between the voluntary sample and the 2021 census population is provided in Table 1. The results here portray the biased nature of the non-probability sample and suggest that some groups of individuals

are more likely to participate in such online surveys, for instance, people with higher education, employed, married, etc.

The sociodemographic characteristics provided in Table 1 are then used as covariates $x_{k1}, \ldots, x_{k7}$ in the propensity score model:

$$\pi_k^A = \pi(\tilde{\mathbf{x}}_k, \boldsymbol{\theta}) = \frac{\exp(\theta_0 + \theta_1 x_{k1} + \ldots + \theta_8 x_{k8} + \theta_9 y_k)}{1 + \exp(\theta_0 + \theta_1 x_{k1} + \ldots + \theta_8 x_{k8} + \theta_9 y_k)},$$

where x_{k8} and y_k, $k \in s$, denote the same binary variable of interest, that is, is the respondent Roman Catholic or not / Evangelical Reformed Believer or not, in 2011 and 2021 census populations, respectively. The estimator $\hat{\boldsymbol{\theta}}$ of the model parameter $\boldsymbol{\theta}$ is found using the maximum likelihood method from Section 1.3.

In order to accelerate the acquisition of the simulation results, we draw a simple random sample from the 2011 census population and analyze a population of around 609 thousand instead of around 3.04 million Lithuania residents. The 2021 estimate of the propensity score model parameter $\boldsymbol{\theta}$ is used to generate 1000 Monte Carlo samples of voluntary participation in the artificial 2011 survey under two scenarios.

1. When the size of the non-probability sample s_A in the 2011 census population is of the same magnitude as in the 2021 census population, – approximately 2%.
2. When the size of the non-probability sample s_A in the 2011 census population is much bigger than that in the 2021 census population, – 20%.

We set the size of the probability sample s_B in the 2011 census population to remain of the same magnitude as in the 2021 census population, – approximately 6%, – and generate 1000 Monte Carlo probability samples in addition to the non-probability samples.

2.3.2 Main findings

Figures 1 and 2 illustrate the distributions of generalized regression estimator (3), as well as IPW and composite estimators under the ignorable (MAR assumption), and the non-ignorable (NMAR assumption) selection mechanisms.

As the box plots of proportions of the Roman Catholics religious group in Fig. 1 and the mean squared error comparison in Table 2 indicate, the most precise results, compared to the red line that denotes the true proportion, are obtained using the generalized regression estimator and the composite estimator under the NMAR assumption. Since the "true" propensity score model represents the non-ignorable selection mechanism, estimators based on the incorrectly specified model tend to produce estimates with a significant bias. That is, the IPW estimator under the MAR assumption underestimates the proportion of interest, and even its combination with the generalized regression estimator does not reduce the selection bias sufficiently.

Another property observed while comparing the scenarios when the non-probability sample comprises approximately 2% and 20% of the 2011 census population (Fig. 1) is that the increase in the non-probability sample size leads to the reduced variance of estimators, but biases remain.

As of estimating the proportion of the Evangelical Reformed Believers, the findings are similar. The box plots in Fig. 2 and the mean squared error comparison in Table 3 indicate that the most precise estimators are the generalized regression, as well as IPW

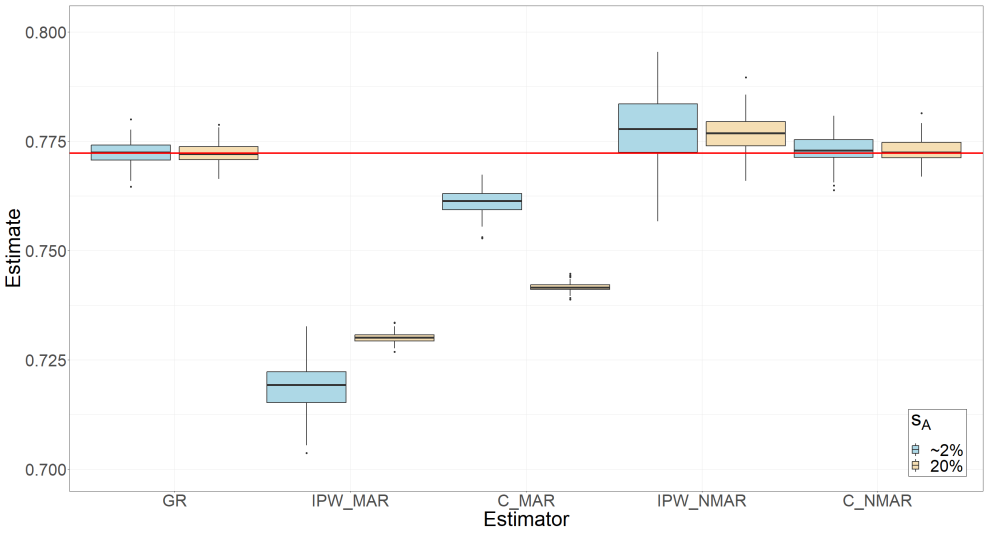

Fig. 1. Comparison of estimators under scenarios when the non-probability sample s_A comprises ∼2% and 20% of the 2011 census population. The red line illustrates the true Roman Catholics proportion (0.772).

Table 2. Comparison of mean squared errors of estimators $\hat{\mu}^{\mathrm{GR}}$, $\hat{\mu}^{\mathrm{IPWi}}$, $\hat{\mu}^{\mathrm{Ci}}$, $\hat{\mu}^{\mathrm{IPW}}$ and $\hat{\mu}^{\mathrm{C}}$ for the Roman catholics proportion (common multiplier is 10^{-4}), and their variance estimators $\hat{\psi}^{\mathrm{GR}}$, $\widehat{V}^{\mathrm{IPWi}}$, $\widehat{V}^{\mathrm{Ci}}$, $\widehat{V}^{\mathrm{IPW}}$ and $\widehat{V}^{\mathrm{C}}$ (common multiplier is 10^{-9}).

s_A comprises	$\hat{\mu}^{\mathrm{GR}}$	$\hat{\mu}^{\mathrm{IPWi}}$	$\hat{\mu}^{\mathrm{Ci}}$	$\hat{\mu}^{\mathrm{IPW}}$	$\hat{\mu}^{\mathrm{C}}$	$\hat{\psi}^{\mathrm{GR}}$	$\widehat{V}^{\mathrm{IPWi}}$	$\widehat{V}^{\mathrm{Ci}}$	$\widehat{V}^{\mathrm{IPW}}$	$\widehat{V}^{\mathrm{C}}$
∼2%	0.0811	29.0046	1.3279	0.9649	0.1083	0.0028	0.0090	0.0069	3.1661	0.0004
20%	0.0488	17.8551	9.4021	0.3650	0.0801	0.0001	0.0002	0.0170	0.2529	0.0054

and composite estimators under the NMAR assumption. Similarly as it was observed in Fig. 1, when the propensity score model is incorrectly specified, the IPW estimator is significantly biased. However, in this case the composite estimator seems to correct the selection bias better.

Nevertheless, even though the increase in non-probability sample size reduces the variances of estimators, when the propensity score model is incorrectly specified, both the IPW and composite estimators remain biased.

3 Conclusions

Making statistical inferences from non-probability samples might be a faster and cheaper approach, however, it also comes with some obstacles such as the unknown sample selection mechanism. This uncertainty often leads to the sample selection bias of the estimators based only on the non-probability sample. To improve the estimation, we construct the composition of the model-based IPW estimator with the design-based post-stratified estimator as a way to integrate both voluntary and probability samples. According to the simulation study, such a combination can correct the sample selection bias.

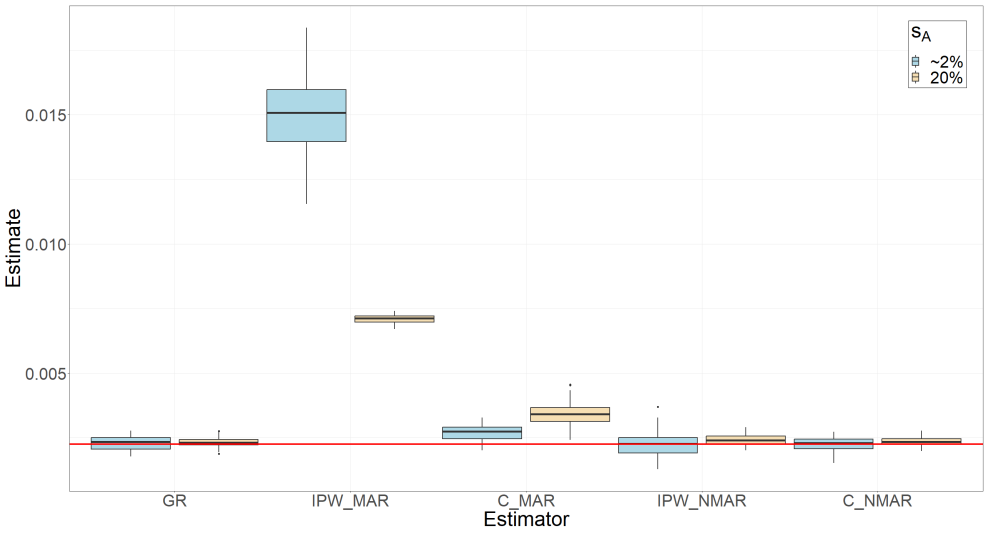

Fig. 2. Comparison of estimators under scenarios when the non-probability sample s_A comprises ∼2% and 20% of the 2011 census population. The red line illustrates the true Evangelical Reformed Believers proportion (0.002).

Table 3. Comparison of mean squared errors of estimators $\hat{\mu}^{\mathrm{GR}}$, $\hat{\mu}^{\mathrm{IPWi}}$, $\hat{\mu}^{\mathrm{Ci}}$, $\hat{\mu}^{\mathrm{IPW}}$ and $\hat{\mu}^{\mathrm{C}}$ for the Evangelical Reformed Believers proportion (common multiplier is 10^{-5}) and their variance estimators $\hat{\psi}^{\mathrm{GR}}$, $\widehat{V}^{\mathrm{IPWi}}$, $\widehat{V}^{\mathrm{Ci}}$, $\widehat{V}^{\mathrm{IPW}}$ and $\widehat{V}^{\mathrm{C}}$ (common multiplier is 10^{-14}).

s_A comprises	$\hat{\mu}^{\mathrm{GR}}$	$\hat{\mu}^{\mathrm{IPWi}}$	$\hat{\mu}^{\mathrm{Ci}}$	$\hat{\mu}^{\mathrm{IPW}}$	$\hat{\mu}^{\mathrm{C}}$	$\hat{\psi}^{\mathrm{GR}}$	$\widehat{V}^{\mathrm{IPWi}}$	$\widehat{V}^{\mathrm{Ci}}$	$\widehat{V}^{\mathrm{IPW}}$	$\widehat{V}^{\mathrm{C}}$
∼2%	0.0068	16.5450	0.0286	0.0200	0.0072	0.0753	23.1086	0.0735	3.9709	0.0365
20%	0.0035	2.3593	0.1618	0.0075	0.0037	0.0142	0.2047	4.3639	0.2055	0.0186

However, the selection bias remains significant if the propensity scores are modeled under incorrect assumptions. As observed in the simulation study, if the propensity score model is built under the MAR assumption when the actual inclusion into the non-probability sample mechanism is non-ignorable, the composite estimator barely corrects the selection bias of the IPW estimator, even with a much larger non-probability sample.

The simulation study also suggests that, under a well-specified propensity score model, it is possible to benefit from the voluntary sample, especially if the estimators based on it are combined with those using the probability sample, and good auxiliary information is available. Therefore, it might prove useful to collect a larger non-probability sample in the future by promoting the survey more, as it lets us to reduce the variances of estimators. Additionally, a much smaller probability sample might be sufficient to evaluate the parameters of an even more complex propensity score model under the NMAR assumption.

References

[1] A. Argüeso, J.L. Vega. A population census based on registers and a "10% survey" methodological challenges and conclusions. *Stat. J. IAOS*, **30**(1):35–39, 2014.

[2] R. Baker, M. Battaglia M.P. Couper J.A. Dever K.J. Gile R. Tourangeau J.M. Brick, N.A. Bates. Summary report of the AAPOR task force on non-probability sampling. *J. Surv. Stat. Method.*, **1**(2):90–143, 2013.

[3] I. Burakauskaitė, A. Čiginas. An approach to integrating a non-probability sample in the population census. *Mathematics*, **11**(8):1782–1795, 2023.

[4] Y. Chen, P. Li, C. Wu. Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.*, **115**(532):2011–2021, 2020.

[5] J.C. Deville, C.-E Särndal. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, **87**(418):376–382, 1992.

[6] J.-K. Kim, S.-M. Tam. Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.*, **89**(1):382–401, 2021.

[7] J.K. Kim. A gentle introduction to data integration in survey sampling. *Stat. Surv.*, **85**(1):19–29, 2022.

[8] X.-L. Meng. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 us presidential election. *Ann. Appl. Stat.*, **12**(2):685–726, 2018.

[9] J.N.K. Rao. On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, **83**(1):242–272, 2021.

[10] J.N.K. Rao, C.F.J. Wu, K. Yue. Some recent work on resampling methods for complex surveys. *Surv. Methodol.*, **18**(2):209–217, 1992.

[11] M.K. Riddles, J.K. Kim, J. Im. A propensity-score-adjustment method for nonignorable nonresponse. *J. Surv. Stat. Method.*, **4**(2):215–245, 2016.

[12] D.B. Rubin. Inference and missing data. *Biometrika*, **63**(3):581–592, 1976.

[13] C.-E. Särndal, B. Swensson, J. Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer, New York, NY, USA, 1992.

[14] S.-M. Tam, J.-K. Kim. Big data ethics and selection-bias: an official statistician's perspective. *Stat. J. IAOS*, **34**(4):577–588, 2018.

[15] Y. Tille. *Sampling and Estimation from Finite Populations*. Wiley Series in Survey Methodology. Wiley, Hoboken, NJ, USA, 2020.

[16] S. Wang, J. Shao, J.K. Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Stat. Sin.*, **24**(3):1097–1116, 2014.

[17] C. Wu. Statistical inference with non-probability survey samples. *Surv. Methodol.*, **48**(2):283–311, 2013.

REZIUMĖ

Savanoriškosios imties panaudojimas populiacijos parametrams vertinti

I. Burakauskaitė and A. Čiginas

Siekiame rasti būdą efektyviai integruoti netikimybinę (savanoriškąją) imtį atveju, kai tyrimo kintamasis stebimas ir tikimybinėje imtyje. Dėl savanoriško dalyvavimo apklausoje atsirandantį parametrų vertinimo poslinkį koreguojame vertindami netikimybinės imties respondentų tikimybes (polinkius) dalyvauti apklausoje. Šių tikimybių įvertinius sudarome naudodami parametrinį logistinės regresijos modelį. Modeliavimą atliekame dviem būdais: esant prielaidai, kad polinkiai dalyvauti apklausoje nepriklauso nuo tyrimo kintamojo ir kad nuo jo priklauso. Abiem atvejais polinkiai įvertinami naudojant didžiausiojo tikėtinumo metodą. Sudarome tiesines kombinacijas populiacijos vidurkio įvertinių, pagrįstų polinkiais dalyvauti apklausoje, ir nepaslinkto įvertinio, pagrįsto tikimybine imtimi. Sudarytus įvertinius palyginame atlikdami imitacinį tyrimo modeliavimą, vertindami populiacijos proporcijas naudojant Gyventojų ir būstų surašymo statistinių tyrimų duomenis.

Raktiniai žodžiai: duomenų integravimas; NMAR; polinkio vertinimas; gyventojų surašymas