STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. xx–xx, https://doi.org/10.59139/stattrans-2026-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

The bias of the estimation of income distribution parameters in non-probability surveys: an experiment based on EU-SILC microdata for Poland

Arkadiusz Kozłowski[1]

Abstract

The selection bias is a major issue when using non-probability samples for inference about finite populations. Minimizing it requires sufficient auxiliary power from external sources and proper modelling. This article examines the potential bias of estimates of income distribution parameters using real-life data from a household survey (European Union Statistics on Income and Living Conditions, EU-SILC) and computer simulations. It investigates the change in the performance of the estimates depending on data defectiveness, the sample size, the pool of auxiliary variables, sources of auxiliary information, and the estimation method. The non-probability samples had non-ignorable bias intentionally introduced by designing selection mechanisms directly tied to income or the access to a computer and the Internet. The results indicated that the bias of the estimates behaved predictably for simple parameters, such as the median and mean. However, for more complex parameters, the bias patterns were less stable. The use of sociodemographic variables for weighting did not consistently reduce the bias; on the contrary, in certain cases, it led to the aggravation of the bias. The number of auxiliary variables and the sample size were as important for the bias reduction as the actual estimation methods, of which the doubly robust estimation performed best.

Key words: selection bias, non-probability sampling, non-ignorable mechanism, income inequality, data weighting.

1. Introduction

Survey research is typically conducted on samples of elements with the aim of estimating the parameters of the whole finite population from which a sample was drawn. The method of sample selection is the key element in determining the possibility and validity of subsequent inference about the population parameters. Over the past

[1] Faculty of Management, Department of Statistics, University of Gdansk, Poland.
E-mail: arkadiusz.kozlowski@ug.edu.pl. ORCID: https://orcid.org/0000-0001-6282-1494.

70 years, the predominant paradigm in survey research methodology has been probability (random) sampling for sample selection and a design-based (randomisation) approach for inference about finite population parameters. The foundation of this approach was established by Neyman (1934) and subsequently developed by Hansen and Hurwitz (1943) and Horvitz and Thompson (1952), among others. The classic works on survey sampling fall within this framework; see, for example, Cochran (1963), Kish (1995), and Särndal et al. (1997).

However, random sampling is costly and not always feasible, primarily due to the need to have a list of all units belonging to the study population (sampling frame) and the need to reach randomly selected (inconvenient) units. In the practice of social research, a sample designed to be random is rarely truly random in the strict sense because of the incorrect mapping of units in the sampling frame (coverage error) and the presence of non-responses. Another common obstacle in list-based samples is the lack of anonymity of the respondents (the entities in a sampling frame must be identifiable so that a researcher can reach them if they are sampled), which makes it difficult to research sensitive topics.

In contrast, non-probability sampling (e.g. quota sampling, haphazard sampling, purposive sampling, see Baker et al., 2013) offers many practical advantages for the researcher. These include convenience, low costs, and the speed of the attainment of results. Recent trends pertaining to the conditions in which surveys are conducted, such as declining response rates in traditional probability surveys, the increased burden on respondents, the ever-growing demand for timely statistics, and the abundance of non-probability data sources (including *big data*), have prompted questions about the possible replacement of traditional research based on probability samples with the non-probability equivalents, even for official statistics (Beaumont, 2020; Beaumont and Rao, 2021; Kalton, 2023).

For non-probability samples to be a valid source for inference about finite populations, their two main deficiencies must be properly handled. The first is the lack of a widely acceptable framework for the assessment of estimation error. In probability sampling, the inference is based on the framework of possible outcomes from repeated draws of a sample with known inclusion probabilities. In non-probability sampling, the inclusion probabilities are not known (the sheer concept of inclusion probability in the non-probability sample and the possible repetition of the selection process are not clear, see, e.g. Lohr 2022, p. 526). The next section provides a short overview of the most important approaches to inference from non-probability samples.

The second weakness is the selection bias - systematic differences between the composition of a sample and the true population characteristics caused by the selection mechanism. Selection bias arises when the inclusion into the sample is somehow correlated with the studied variables. For example, to estimate how frequently people with

disabilities participate in public events, the sample may be selected as a convenience sample. People who go out more often, attend concerts, go to restaurants, etc., will be easier (more convenient) to recruit; therefore, their participation in the sample will be greater than their proportion in the general population. Such a sample will be biased and will generate systematically erroneous results (Szreder and Kozłowski, 2024, pp. 73–74).

Selection bias can be reduced if the researcher knows the selection mechanism well and can properly model it using auxiliary variables and additional reliable data sources, such as a probability sample or administrative records. These auxiliary variables and the right model are crucial for bias reduction. If the variables at hand have weak explanatory power for the selection mechanism, the bias could still be eliminated, provided they have strong explanatory power for the target variables. If a model captures all the forces that influence entry into a non-probability sample and/or the variability of the study variable, the units in and out of the non-probability sample are then exchangeable given the variables in the model (Mercer et al., 2017; i.e. conditionally exchangeable; see Li, 2024). Being so is the crucial and usually unmet assumption for the validity of inference from non-probability samples.

In most social surveys, it is reasonable to assume that the right model for the selection mechanism or the study variables could never be found due to the complexity of phenomena such as opinions, behaviors, attitudes, etc. In addition, the issue of finding the right model is complicated by the fact that the selection bias depends on the study variable; thus, the model properly capturing the selection forces may be different for different variables. Most surveys based on probability samples in practice use one set of demographic variables to calibrate weights assigned to sampled units to known population totals in order to, inter alia, reduce bias stemming from non-responses and other non-random errors. The question is how useful such variables could be in mitigating the effect of selection bias. Several researchers have noted that demographics and simple weighting are not sufficient to reduce the bias; see the review in Cornesse et al. (2020), as well as studies by Lavrakas et al. (2022) and Mercer and Lau (2023). They uncovered this by comparing probability and non-probability surveys with similar objectives with a benchmark. In this article, we wanted to check to what extent demographics and some weighting procedures can mitigate the selection bias, but instead of having a couple of real non-probability samples, we turned to a Monte Carlo simulation to have a broader scope of possibilities. Still, our experiment leveraged natural relationships between variables in a social survey, not artificially generated.

Using computer simulations on real-life data for generating non-probability samples is not obvious. When testing the performance of the proposed estimation methods for non-random samples, artificially generated data are often used (e.g. Chen, Li and Wu, 2020; Yang, Kim and Song, 2020; Kim and Morikawa, 2023). The use of real data

(from random and non-random samples) is often limited to a one-off application where, rather than assessing replicative properties, results are compared to known population values (e.g. Wang et al., 2015; Chen, Valliant and Elliott, 2019) or reliable probability survey estimates (e.g. Beaumont et al., 2024), or the estimates themselves and some aspects of estimation procedures are evaluated (e.g. Chen, Li and Wu, 2020; Yang, Kim and Song, 2020; Kim et al., 2021).

To assess the replicative properties of estimation from non-probability samples using real data, probability sampling is used, but with different selection probabilities, in a way that mimics the possible non-probability mechanism. Valliant (2020) and Buelens et al. (2018) used cut-off sampling (planned undercoverage), which deliberately excluded certain elements from the frame. Also, Valliant (2020) and Lee & Valliant (2009) used stratified sampling with disproportional allocation. Marella (2023) used Poisson sampling. Liu et al. (2023) used unequal probability random systematic sampling without replacement.

Most authors in simulation studies (on artificial or real non-probability data) test the ignorable mechanism. Testing a non-ignorable mechanism, such as where the probability of getting into a non-random sample depends on the target variable, is done by Kim & Morikawa (2023) and Marella (2023).

Virtually all simulations on non-probability samples focus on estimating means (including proportions) and totals. A notable exception is the study by Wiśniowski et al. (2020), where regression parameters were estimated.

The simulation in this study aimed to test the extent to which the bias caused by non-ignorable selection mechanisms can be minimized using socio-demographic auxiliary variables when estimating parameters of household income distribution. Contrary to most studies, where only means are tested, the parameters estimated include non-linear functions of the target variable. The estimation methods cover various techniques, but only the variants where a single vector of weights can be derived have been used. This was to mimic the production of estimates in most probability surveys. The estimation methods also do not take into account the non-ignorability of the selection process. In reality, it is usually not known whether the selection process is ignorable or not. If the conditions are ignorable, one can expect unbiased estimates with manageable precision. The interesting part is what happens when the process is non-ignorable, and the performance of estimators depends on the natural relationship between income and socio-demographic characteristics.

The scenario tested in the simulation is that the target variable is only observed in a non-probability sample, and the estimation is strengthened by additional information on auxiliary variables from an independent probability sample and/or the entire popu-

lation. The data from a large-scale probability survey is used for the experiment. Although the experiment is based on one specific survey, it aims to introduce a methodological framework for such an assessment and identify certain regularities.

The rest of the article is structured as follows. The next section is devoted to the primary approaches to estimation in non-probability sample surveys. Our method of investigating estimation bias is then presented, starting with the real data description, followed by methods of generating non-probability samples, and ending with the estimation procedures. The results of the experiment are presented in Section 4, and some intriguing outcomes are discussed in Section 5. The article ends with a conclusion section.

2. Estimation from non-probability samples

We are interested in the finite population of elements indexed by $U = \{1, 2, \dots, N\}$. The elements may be observed on a study variable y and a vector of covariates $\boldsymbol{x}$. The population is considered constant, meaning that it is assumed that the surveying element i ($i = 1, 2, \dots, N$) would always yield the same values y_i and $\boldsymbol{x}_i$. A survey aims to estimate population parameters (functions of y) based on a sample of elements and possibly some auxiliary information. The unknown part of this setup is which elements are included in the sample. Following Meng (2018), we introduce the sample membership indicator R_i ($R_i = 1$ if element i is in the sample, $R_i = 0$ otherwise). The indicator is the result of all forces that determine whether an individual enters the sample. These include the selection mechanism (probabilistic or not), non-response, coverage errors, and others. Most of them are not controlled by the researcher. In probability sampling and pure conditions (without non-random errors), the distribution of R_i is known; notably, $E(R_i) = P(R_i = 1) = \pi_{P,i}$ (first-order inclusion probabilities) are known for each population element prior to sampling. These probabilities, along with second-order inclusion probabilities, $\pi_{P,ij} = P\big(R_i R_j = 1\big)$, play a central role in the generalization of the sample outcomes to the entire population (Särndal, Swensson and Wretman, 1997). In practice, the pure conditions rarely occur, and the selection mechanism is often a non-probability one.

Consider the non-probability selection mechanism leading to a sample s_{NP}. The inclusion probabilities are not known; thus, the usual randomization approach for inference about the population parameters cannot be applied. Nevertheless, the body of statistical theory offers some possibilities of inference from non-probability samples. Two primary approaches are quasi-randomization and superpopulation modelling (Zhang, 2019; Valliant, 2020). Other solutions can be regarded as variations or combinations of these two approaches. Every method requires some additional data sources

aside from a non-probability one, some auxiliary variables besides the ones studied, and explicit or implicit modelling.

To effectively use non-probability data for the production of estimates, a single vector of weights is desirable. The weights should be independent from any study variable and constructed in such a manner that a linear combination of the values of the study variable with the weights as coefficients yields an estimate of the total:

$$\hat{y} = \sum_{i \in s_{NP}} w_{NP,i} \cdot y_i \tag{1}$$

where $w_{NP,i}$ is the weight for unit i in the non-probability sample. Using a single vector of weights is the usual practice in probability surveys. It is convenient because the weights could be used for different y variables and parameters. In this study, only estimation methods that allow the construction of a single weight vector will be used.

The quasi-randomization approach (also: inverse probability weighting or propensity score adjustment, PS) is conceptually equivalent to the method introduced earlier to address nonresponse bias (see Valliant et al. 2013, pp. 321–338). In the nonresponse case, the mechanism leading to nonresponse is treated as the second phase of sampling; in this phase, unlike in the usual sampling design treated here as the first phase, the inclusion probabilities are not known and must be modelled after the survey is conducted based on information regarding both respondents and non-respondents. In the non-probability sample case, the aim is to estimate the (pseudo)probabilities of being included in s_{NP}. The estimation is performed after the sample is selected and is based on covariates that are believed to affect the inclusion in the non-probability sample. After the pseudo-inclusion probabilities are estimated, the inverses of the estimated inclusion probabilities can be used as weights in the production of estimates:

$$w_{NP,i}^{(PS)} = \frac{1}{\hat{\pi}_{NP,i}} \tag{2}$$

where $\hat{\pi}_{NP,i}$ is the estimated pseudo-inclusion probability for unit i in the s_{NP}. There are other propensity-score-based methods (see, e.g. Rivers, 2006), but they are beyond the scope of this article.

Superpopulation modelling (SM; model-based approach) is a prediction-based paradigm in which the values of variables of interest for unsampled units are predicted using specified models, and population estimates are generated as functions of these predictions (Valliant, Dorfman and Royall, 2000). This approach is predicated upon the idea that a finite population is merely a sample from a superpopulation. This superpopulation is defined by the model binding the variable of interest with specific covariates. In other words, it is assumed that finite population elements are generated by the model with certain fixed parameters and a random error. The model parameters are estimated from the sample elements, and the estimator of the total is then constructed

simply as a sum of y-values over the sample elements plus the sum of predictions over the non-sampled elements (Elliott and Valliant, 2017). There are several variants of the model-based approach, such as multilevel regression and poststratification method (Gelman and Little, 1997), mass imputation (Kim et al., 2021), using penalized estimation (Liu, Wang and Pan, 2025) or machine learning techniques (Ferri-García, Castro-Martín and Rueda, 2021).

The model-based estimators are not typically presented as a linear combination of sample elements, but they could be written in this form for simple models such as the linear model below:

$$E_M(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad V_M(y_i|\boldsymbol{x}_i) = v \tag{3}$$

where E_M and V_M are expectation and variance with respect to the superpopulation model. Estimating parameter vector $\boldsymbol{\beta}$ from the non-probability sample using the ordinary least squares method allows the weights to be presented in the following form:

$$w_{NP,\,i}^{(SM)} = \widehat{\boldsymbol{X}}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{4}$$

where $\widehat{\boldsymbol{X}}$ is the vector of population totals for auxiliary variables (these totals could be known or estimated from the probability reference sample), and $\boldsymbol{X}$ is a matrix of values for auxiliary variables from the non-probability sample.

A combination of quasi-randomization and superpopulation approaches yields a doubly robust estimation (DR). In this method, the PS weights are involved in making predictions based on the model. The doubly robust estimation is robust in the sense that it protects against the misspecification of either the inclusion or the superpopulation model (Chen, Li and Wu, 2020; Yang, Kim and Song, 2020).

In the doubly robust estimation, adopting the same linear model (3) translates into the following formula for the weights (see Valliant 2020, p. 243):

$$w_{NP,\,i}^{(DR)} = w_{NP,\,i}^{(PS)}\left[1 + \left(\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{x}}_{PS}\right)^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}_i\right] \tag{5}$$

where $\widehat{\boldsymbol{x}}_{PS} = \sum_{i\in s_{NP}} w_{NP,\,i}^{(PS)} \cdot \boldsymbol{x}_i$ is the vector of estimated population totals for x-variables using the PS method and $\boldsymbol{W} = diag\left\{w_{NP,\,i}^{(PS)}\right\}$. The weights in (5) are the same as the linear calibration weights, where the initial weights, rather than design weights, are the weights (2) from the PS method (see Deville and Särndal, 1992).

3. Methodology of the experiment

3.1. Data

The data for the experiment come from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. The aim of the survey is to obtain estimates

of income, poverty, social exclusion, and other aspects of living conditions. The survey is harmonized across 32 European countries, and its implementation is overseen by Eurostat. In each country, the survey "must be based on a nationally representative probability sample of the population residing in private households within the country" (European Commission 2022, p. 59). The data used in the experiment are microdata obtained from Statistics Poland pertaining to the 2022 edition conducted in Poland (Eurostat, 2024). The sample in this survey was selected using a two-stage sampling scheme, with enumeration census areas as primary and dwellings as secondary sampling units (all the households in the selected dwellings are surveyed). Stratified sampling with disproportional allocation was used at the first stage and simple random sampling without replacement in the second stage. Stratifying variables were NUTS 2 regions, city size classes and share of richest dwellings in the primary sampling unit. The last feature was calculated using information obtained from tax records (GUS, 2023).

The variables considered in the simulation concerned the household as a whole or the person completing the questionnaire. As the main objectives of the EU-SILC survey relate to incomes and income is also one of the variables in social surveys that might be most heavily affected by non-probabilistic sample selection (in the same way as it is affected by non-random errors, concerning measurement, coverage and non-response, especially for the top income units; see e.g. Brzeziński, Sałach and Wroński, 2020; Carranza, Morgan and Nolan, 2023), the main parameters of interest in our experiment are various parameters of income distribution. The main study variable is the equivalized annual disposable income (y_i), which is not directly observed but is calculated as follows:

$$y_i = \frac{TDHI_i}{EHS_i} \tag{6}$$

where: $TDHI_i$ is the total disposable household income for a household i, and EHS_i is the equivalized household size, calculated as a sum of weights: 1 for the first person aged 14 and over, 0.5 for each additional person at this age, and 0.3 for each person aged 13 or less. For brevity, the equivalized annual disposable income will hereinafter be referred to as "equivalized income".

The value of y_i is assigned to each person in the household i. The following parameters describe the population of persons, but the statistical unit in the dataset is a household. Therefore, the household data is considered as grouped data of persons, which is reflected in the way the parameters are calculated (i.e. they are weighted by the number of persons in a household). Additionally, for all the formulas that follow, except the mean, it is assumed that the households are ordered in a non-decreasing way with

respect to y. The parameters to be estimated and their population formulas are as follows:

- Mean (Mean of equivalized income for a person):

$$Y_{Mean} = \frac{\sum_{i=1}^{N} y_i \cdot z_i}{\sum_{i=1}^{N} z_i} \tag{7}$$

where: z_i is the household size (total number of household members).

- Median (Half of the persons in the population have equivalized income not higher than the median):

$$Y_{Median} = min\left\{ y_i \middle| \sum_{i'=1}^{i} p_{i'} \geq 0.5 \right\} \tag{8}$$

where: $p_{i'} = \frac{z_{i'}}{\sum_{i=1}^{N} z_i}$.

- Poverty (At-risk-of-poverty rate - the percentage of persons with an equivalized income below the at-risk-of-poverty threshold set at 60% of the median):

$$Y_{Poverty} = \frac{\sum_{i=1}^{N} I(y_i < 0.6 \cdot Y_{Median}) \cdot z_i}{\sum_{i=1}^{N} z_i} \tag{9}$$

where: $I(u)$ is the indicator function, $I(u) = 1$ if u is true, and $I(u) = 0$ otherwise.

- Depth (Relative median at-risk-of-poverty gap - the difference between the median equivalized income of persons below the at-risk-of-poverty threshold and this threshold, expressed as a percentage of the at-risk-of-poverty threshold):

$$Y_{Depth} = \frac{0.6 \cdot Y_{Median} - \tilde{Y}_{Median}}{0.6 \cdot Y_{Median}} \tag{10}$$

where $\tilde{Y}_{Median} = min\{y_i \mid \sum_{i'=1}^{i} p'_{i'} \geq 0.5 \ \wedge \ y_i < 0.6 \cdot Y_{Median}\}$ is the median equivalized income of persons below the at-risk-of-poverty threshold, and $p'_{i'} = \frac{z_{i'}}{\sum_{i:\, y_i < 0.6 \cdot Y_{Median}} z_i}$.

- Inequality (Income quintile share ratio - the ratio of total income received by the 20% of the population with the highest income to that received by the 20% of the population with the lowest income):

$$Y_{Inequality} = \frac{\sum_{i \in top20\%} TDHI_i}{\sum_{i \in bottom20\%} TDHI_i} \tag{11}$$

where $i \in top20\%$ and $i \in bottom20\%$ mean such households (i) for which $\sum_{i'=1}^{i} \frac{z_{i'}}{\sum_{i=1}^{N} z_i} > 0.8$ and $\sum_{i'=1}^{i} \frac{z_{i'}}{\sum_{i=1}^{N} z_i} < 0.2$, respectively (assuming units are sorted in non-decreasing order with respect to y, not the $TDHI$).

- Gini (Gini coefficient; For the household data, the Gini coefficient is calculated using the trapezoid rule, because the household data are treated here as grouped data of individuals):

$$Y_{Gini} = 1 - \sum_{i=1}^{N} \left(\sum_{i'=0}^{i} \frac{z_{i'} y_{i'}}{Y_{total}} + \sum_{i'=0}^{i-1} \frac{z_{i'} y_{i'}}{Y_{total}} \right) \cdot p_i \tag{12}$$

where $Y_{total} = \sum_{i=1}^{N} z_i y_i$, $p_i = \frac{z_i}{\sum_{i=1}^{N} z_i}$, $z_0 = 0, y_0 = 0$.

The sociodemographic variables utilized as auxiliaries concerned the household as a whole or the person completing the questionnaire. These variables are as follows (numbers in parentheses indicate the number of categories): region (7), degree of urbanization (3), tenure status (2), educational attainment level (5), self-defined current economic status (3), household type (10), number of rooms available to household (6), sex (2), marital status (4), age at the end of the income reference period (7). All auxiliary variables are categorical (or categorized), which reflects the usual way these variables are used for weight adjustment in probability sampling. To use these variables in the models, dummy variables were created representing all but one variant from each original variable. Only additive effects were considered in the models. This yielded a total of 39 explanatory variables. The experiment examined two variants: using all variables (39) or only selected ones, most correlated with income (first 5 variables; a total of 15 binary variables).

In the original sample from EU-SILC 2022, 19,757 households were surveyed. 19,655 household left after removing the units with missing values on auxiliary variables. Since it was a probability sample, each household in the dataset came with a weight that reflected different inclusion probabilities and subsequent adjustments. These weights were used to construct the probable population of 12,633,327 households by copying each entry as many times as its weight, which was rounded to the nearest integer. This dataset was too big to run the simulations described below, so a simple random sample with replacement was drawn of size $N = 100{,}000$. This sample served as the target population in further analysis (the parameters of this population were practically the same as the weighted statistics from the original EU-SILC sample). It was assumed that this sample was representative of the general population of households in Poland in terms of relationships between variables that were exploited in the experiment.

3.2. Sampling

When selecting a non-probability sample, there is no defined randomness involved, as in probability sampling, but one could imagine the accidental nature of various activities related to the selection of respondents (e.g. the respondent visiting a website where an invitation to participate in the survey will be displayed). If it were possible to

select a non-probability sample again, with the same relevant research conditions, a different sample composition would be obtained. Therefore, when writing an algorithmic selection of a non-probability sample for a computer simulation, a probabilistic mechanism can be used. However, in such a mechanism, some kind of selection bias should be incorporated. The experiment tested two non-ignorable mechanisms: The first one was dependent directly on income (hereinafter "income-dependent"), and the second one was dependent on having a computer and Internet access (hereinafter "Internet-dependent").

In the first mechanism, first-order inclusion probabilities for all members of the target population were generated using the logistic function as follows:

$$\pi_{NP,i}^{(1)} = \frac{1}{1 + e^{-(b_0 + b_1 \cdot y_i)}} \tag{13}$$

where b_0 is an intercept chosen in such a manner that $\sum_{i=1}^{N} \pi_{NP,i}^{(1)} = n_{NP}$ for a chosen non-probability sample size n_{NP}, and b_1 is the parameter affecting the strength of the selection bias. Two values of b_1 have been tested, -1 and -0.2 (see Table 1). Both values are negative, which means that households with higher income were assigned smaller inclusion probabilities. Similar to Liu et al. (2023), the sampling scheme was based on random systematic sampling (Tillé 2006, pp. 127–128), which allowed for a fixed sample size in a sampling without replacement with unequal probabilities. The population was randomly sorted before each selection.

The second mechanism depends on owning a computer and having access to the Internet. The sampling was done as planned undercoverage - the samples were drawn only from the portion of the households that met both criteria. 79.6% of households in the target population own a computer and have Internet access. From this part of the population, the samples were drawn using simple random sampling without replacement.

Following Meng (2018), we can measure the selection bias (or data defect) for the estimation of the mean as $E_R\left[\rho_{R,y}^2\right]$, where E_R is the expected value over an R-mechanism, and $\rho_{R,y}$ is the correlation coefficient between the sample membership indicator R and the study variable y, calculated as follows:

$$\rho_{R,y} = \frac{Cov_K(R,y)}{D_K(R) \cdot D_K(y)} = \frac{\frac{1}{N}\sum_{i=1}^{N} R_i y_i - f_{NP} \cdot \bar{Y}}{\sqrt{f_{NP}(1 - f_{NP})} \cdot \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2\right)}} \tag{14}$$

where Cov_K and D_K denote covariance and standard deviation with respect to the empirical distribution of a variable (i.e. population parameters), and $f_{NP} = n_{NP}/N$, $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$. Since the sampling unit in our experiment was a household and the pa-

rameters of interest pertaining to income are always weighted by the number of household members, the coefficient (14) should also be weighted. The formula is then as follows:

$$\rho_{R,y_w} = \frac{\sum_{i=1}^{N} p_i R_i y_i - \bar{R}_w \cdot \bar{Y}_w}{\sqrt{(\bar{R}_w - \bar{R}_w^2)} \cdot \sqrt{(\sum_{i=1}^{N} p_i Y_i^2 - \bar{Y}_w^2)}} \tag{15}$$

where $p_i = z_i / \sum_{i=1}^{N} z_i$ is the weight of an element i, z_i denotes the number of household members, $\bar{R}_w = \sum_{i=1}^{N} p_i R_i$, and $\bar{Y}_w = \sum_{i=1}^{N} p_i Y_i = Y_{Mean}$.

Three selection mechanisms and two non-probability sample sizes ($n_{NP} = 1000$ or $n_{NP} = 2000$) were tested, yielding six designs with different defectiveness. For each design $M = 1000$ samples were generated, and for each sample the correlation coefficient (15) was calculated. Table 1 presents the means of (square) correlation coefficients for each sampling design. Negative mean correlation coefficients in the first two R-mechanisms were induced by negative coefficients b_1 in (13). The positive correlation for the third mechanism arose naturally as the effect of households with a computer and Internet access being more affluent. Each subsequent mechanism is by one order of magnitude less defective in terms of mean ρ^2_{R,y_w}. What is important, the defectiveness of a sampling mechanism increases with increasing sample size. For comparison, if the R-mechanism were simple random sampling (SRS), the value of $E_R\left[\rho^2_{R,y_w}\right]$ would be $1/(N-1) = 0.00001$ and $E_R\left[\rho_{R,y_w}\right] = 0$. The expected squared correlation in SRS is thus about seven times smaller than in the least defective non-probability mechanism introduced in the experiment.

Table 1. Estimated sampling design defects measured by the mean (square) correlation coefficients between sample membership indicator R and equivalized income, by R-mechanism and sample size

Selection mechanism dependent on	n_{NP}	Mean over M samples of	
		ρ_{R,y_w}	ρ^2_{R,y_w}
Income (strongly), with $b_1 = -1$	1000	-0.05269	0.00278
	2000	-0.07438	0.00554
Income (weakly), with $b_1 = -0.2$	1000	-0.01549	0.00025
	2000	-0.02185	0.00049
Having a computer and internet access	1000	0.00741	0.00007
	2000	0.01046	0.00012

Source: own calculations.

As Table 1 shows, both non-probability sampling mechanisms introduced in the experiment led to data defects. In both cases, the mechanisms depended on variables

that were not used at the estimation stage. Therefore, they were non-ignorable – the assumption for propensity score modelling was not met. In such a situation, the effectiveness of the estimation methods in reducing bias depends on how well the auxiliary variables can explain the variation in the variable responsible for selection bias.

Probability reference samples, needed for auxiliary information at the estimation stage, were drawn independently for each non-probability sample using stratified sampling with proportional allocation. The strata were formed by cross-classifying *Region* and *Degree of urbanization*, which resulted in 21 strata. The size of these samples was $n_P = 1000$.

3.3. Estimation

It was assumed that the estimation methods tested in the experiment should have similar practical properties to the estimation methods used in most probability sample surveys. There were, therefore, two criteria for selecting the methods: no need to create a separate model for various study variables and the possibility of creating a single weight vector, thus allowing for the estimation of multiple parameters for distinct target variables. An additional assumption was to use different estimation approaches discussed in Section 2. Therefore, three solutions were used in the experiment: the propensity-score adjustment method, the model-based method, and doubly robust estimation. In each method, the essential part was the formula for weights.

Several propensity-score adjustment methods have been proposed in the literature (see, e.g. Valliant and Dever, 2011; Wang, Valliant and Li, 2021; Kim and Kwon, 2024), including methods based on machine learning techniques (Rueda et al., 2023; Ferri-García et al., 2024). In this experiment, the pseudo-inclusion probabilities were estimated using the logistic regression model:

$$\hat{\pi}_{NP,\,i} = \frac{e^{\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}}} \tag{16}$$

where $\boldsymbol{x}_i$ is the vector of values for auxiliary variables for unit i, and $\widehat{\boldsymbol{\beta}}$ is the vector of estimated model parameters. The model parameters were estimated using the maximum pseudo-likelihood estimation method as proposed by Chen et al. (2020). In this method, the estimates $\widehat{\boldsymbol{\beta}}$ maximise the pseudo-log-likelihood function:

$$l^*(\boldsymbol{\beta}) = \sum_{i \in s_{NP}} \boldsymbol{x}_i{}^T\boldsymbol{\beta} - \sum_{i \in s_P} \frac{1}{\pi_{P,i}} \log\left(1 + e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}}\right) \tag{17}$$

where $\pi_{P,i}$ is the inclusion probability for element i in the reference sample. The maximum is obtained by solving the score equation $\frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\beta}) = \sum_{i \in s_{NP}} \boldsymbol{x}_i - \sum_{i \in s_P} \frac{1}{\pi_{P,i}} \cdot \pi_{NP,i} \cdot \boldsymbol{x}_i = 0$ by means of the Newton-Raphson iterative procedure. The weights were then calculated using formula (2).

The weights for the model-based and the doubly robust methods were calculated as described in Section 2, formulas (4) and (5), respectively. The linear model assumed in the derivation of weights in those methods may not be the best choice when all auxiliary variables are of a categorical nature, as was the case in the experiment, but it allows the single weight vector to be derived. This practical advantage makes these estimation procedures resemble the production of estimates in probability surveys.

In the practice of sample surveys, the distribution of weights is also examined, and possibly some trimming is performed. The use of categorical auxiliary variables and additive models in our experiment should keep the weights in a reasonable range. Nevertheless, the effect of trimming extreme weights was also examined. The trimming procedure employed consists in setting negative or large weights to zero or to an upper bound, respectively, and spreading the total weight trimmed away among the other sample units. The upper bound was arbitrarily chosen as three times the median weight.

The estimators of parameters described in Section 3.1. were then constructed straightforwardly by plugging the weights into the parameter formulas. The differences in the formulas for estimators compared to formulas for parameters are as follows:

- Summation is done over households in a non-probability sample.
- The expanded values of household size, $\check{z}_i = z_i \cdot w_{NP,i}$, and total disposable household income, $\widetilde{TDHI}_i = TDHI_i \cdot w_{NP,i}$, replaced the observed ones.

The conditions that were controlled in the experiment yielded 144 experimental setups. They comprised 6 sampling procedures simulating non-probability sample selection (3 selection mechanisms and 2 sample sizes) and 24 estimation procedures (2 sets of auxiliary variables, 2 sources of auxiliary information, 3 methods of weights construction, and 2 options for weight trimming). Despite using methods from different approaches to estimation, all solutions were tested using the randomization paradigm (i.e. assuming constant population values and averaging over possible samples). The performance of each setup was assessed by the relative bias of estimates of a finite population parameter θ, using estimator $\hat{\theta}$, which was calculated as follows (for $M = 1000$ iterations):

$$RB(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m - \theta}{\theta} 100\% \tag{18}$$

4. Results

To assess the sampling bias induced by non-ignorable selection mechanisms, we examined naïve estimates of income distribution parameters (i.e. without applying any weighting). Table 2 presents the relative biases of naïve estimators. The bias was highest

in the case of a selection mechanism strongly dependent on income. The weaker dependence on income in most cases resulted in considerably lower bias. The lowest bias was typically observed for the mechanism dependent on possession of a computer and Internet access.

Table 2. Relative bias (%) of naïve estimates of parameters of income distribution by non-probability sampling mechanisms

Selection mechanism dependent on	n_{NP}	Relative bias (%) for estimation of parameters:					
		mean	median	poverty	depth	inequality	Gini
Income (strongly)	1000	-27.8	-22.9	22.1	29.0	-1.0	-4.7
	2000	-27.6	-22.6	21.8	27.7	-1.7	-4.8
Income (weakly)	1000	-8.0	-5.4	2.6	10.7	-4.8	-4.1
	2000	-8.0	-5.4	3.1	10.9	-4.9	-4.1
Having a computer and internet access	1000	3.7	4.3	-3.9	3.9	6.7	-1.6
	2000	3.6	4.2	-4.4	3.4	6.2	-1.7

Source: own calculations.

Notably, estimations of different parameters uncovered different patterns of bias. The sample mean and median behaved similarly. They had a strong negative bias for mechanisms dependent on income because inclusion probabilities were negatively correlated with income, and there was a positive bias for the mechanism dependent on having a computer and Internet access because this mechanism favored more affluent households. The poverty estimator (i.e. sample at-risk-of-poverty rate) showed the opposite pattern to median and mean in terms of signs. In income-dependent mechanisms, where poorer households appeared in the sample more frequently than in the population, the poverty rate was overestimated, and in the other mechanism, due to the opposite effect, the rate was underestimated. The depth estimator (i.e. sample relative median at-risk-of-poverty gap) differed from the previous ones in that its bias was always positive and the greatest of all estimators for the mechanism related to income. The inequality and Gini estimators appeared to be most resilient to the strong income-dependent selection mechanism. The Gini parameter was always underestimated.

As for the sample size, one cannot observe any distinctive differences in estimation bias between smaller and larger samples. This stands in contrast to the data defect indexes presented in Table 1, which were always higher for a larger sample size. This results from the specific nature of $\rho_{R,y}$, which is the correlation between binary variable R and study variable y. The variance of a binary variable depends on its mean, and the mean of R is the sampling fraction $f_{NP} = n_{NP}/N$. Changing the number of 1s in R changes the correlation, even though the underlying mechanism stays the same. Thus, the quantity $E_R[\rho_{R,y}^2]$ should not be interpreted in separation from the sample size

(additional information from a larger sample size is not reflected in $E_R[\rho^2_{R,y}]$ but in other components of the mean squared error; see Meng, 2018).

For each of the three selection mechanisms simulating non-probability sampling, various estimation strategies were examined. The relative bias of these strategies for each estimated parameter is presented in Figure 1. For each parameter and selection mechanism, there are 48 circles in the same color representing complex estimation methods (24 different estimation strategies for each of the two sample sizes) and, for comparison, two triangles representing naïve estimation, one for each tested sample size. The points are mapped exactly on the Y-axis but are slightly jittered on the X-axis to better visualize their position. The graphic aims to illustrate the general effect of using auxiliary information and different weighting procedures on bias. If the circles are closer to the horizontal line, which represents unbiasedness (bias = 0%), than the triangles, then using sociodemographic variables and various weighting procedures generally reduces the bias of naïve estimation.

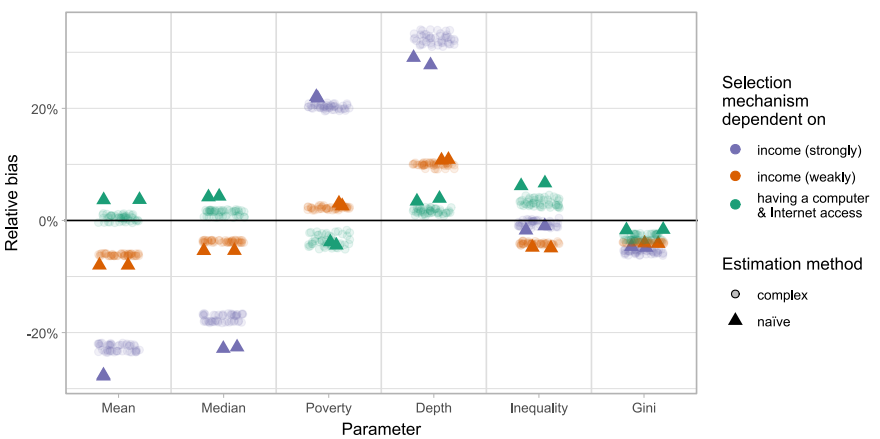

Figure 1. Relative bias (%) of naïve and complex estimates of parameters of income distribution by non-probability sampling mechanisms

Source: own calculations.

The pattern of the reduction (or increase) of the relative bias again hinged on the type of parameter estimated. For median and mean, there was always a visible improvement, irrespective of the selection mechanism. The effect was satisfactory for the Internet-dependent mechanism, sometimes reducing the bias almost to zero, but insufficient for income-dependent mechanisms. For poverty and inequality estimators, the bias was also reduced on average, but to a lesser extent. The strangest pattern was exhibited by the depth estimates. The least biased naïve estimation in the Internet-related mechanism was slightly corrected by weighting; the same applied to the estimation in the

weaker form of the income-dependent mechanism. However, in the stronger form of the income-dependent mechanism, weighting always increased the already large bias of naïve estimates for the depth parameter. The adverse effect of weighting was also present in the estimation of the Gini coefficient.

The 144 estimation strategies examined were the result of the combination of all variants of six controlled factors. To assess the impact of a particular factor, the differences in the absolute value of relative bias between complex and corresponding naïve estimations were calculated and averaged separately over specific factors. The results of these operations are presented in Table 3. Negative values mean improvement, and positive values mean deterioration in terms of estimation bias after the introduction of weights.

Table 3. The average difference in the absolute value of relative bias (in percentage points) between the complex and corresponding naïve estimations of parameters of income distribution by experimental conditions

Experimental condition		**The average difference in relative bias between the complex and naïve estimation of parameters:**					
name	**value**	**mean**	**median**	**poverty**	**depth**	**inequality**	**Gini**
Non-probability sample size	1000	-3.26	-3.23	-1.01	0.58	-1.64	0.67
	2000	-3.35	-3.32	-1.08	-0.16	-1.42	0.69
Number of auxiliary variables	small	-3.04	-2.91	-0.63	0.27	-1.65	0.75
	large	-3.57	-3.63	-1.46	0.15	-1.41	0.61
Source of auxiliary information	population	-3.30	-3.27	-1.06	0.10	-1.56	0.68
	reference sample	-3.31	-3.27	-1.03	0.33	-1.50	0.68
Estimation method	PS	-3.16	-3.10	-0.89	0.04	-1.51	0.67
	SM	-3.32	-3.36	-1.23	0.44	-1.48	0.74
	DR	-3.43	-3.36	-1.01	0.16	-1.59	0.63
Weight trimming	no	-3.32	-3.29	-1.03	0.21	-1.57	0.69
	yes	-3.29	-3.25	-1.06	0.22	-1.48	0.67

Source: own calculations.

The type of estimated parameter differentiated the effect of changing a factor on the reduction (or increase) of estimation bias compared to naïve estimation. Increasing the non-probability sample size typically resulted in a greater reduction in bias or a smaller increase in it. The exceptions were for the estimation of inequality and Gini, but the differences are small. The larger set of auxiliary variables usually had a positive effect. The only exception was for the inequality estimation. The source of auxiliary

information (population totals known exactly or estimated from a probability reference sample) exhibited no significant effect on bias, except for the depth estimator.

The effect of the estimation method was noticeable but could be outweighed by a larger set of auxiliary variables. Nevertheless, the largest bias reduction (or smallest increase) was typically observed when the doubly robust (DR) estimation method was used. It was either the best (in four cases) or the second-best (in two cases) solution for estimation. In the case of depth estimation, the best was PS weighting, which increased the bias the least, and in the case of poverty estimation, the best was SM weighting, which decreased the bias the most.

Weight trimming had a very small and uneven effect on the bias. This is because only a few weights fell outside the specified bounds. As expected, the distributions of weights were moderately skewed. There were no negative weights for the PS method, almost none for the DR method, and only a few for the SM method. The incidence of extremely large weights was 0.5% on average.

Generally, the effects of various changes in complex estimation methods were in line with the expectations (better performance for larger sample size, larger set of auxiliary variables, and DR estimator) for the estimation of simple parameters, namely, median and mean. For other, more complex parameters, the effects were not as consistent.

5. Discussion

The use of socio-demographic variables for weighting sampled units in a non-ignorable selection mechanism mostly reduced the bias, at least to some extent, in our experiment. However, in some cases, the bias increased after weighting, regardless of the method used to construct the weights. The deterioration in estimator performance following the introduction of weights was particularly pronounced for the depth estimator under the strong income-dependent mechanism, and for the Gini estimator under the Internet-dependent and strong income-dependent mechanisms.

In the depth case, the cause of this behavior may lie in the fact that the measure is a ratio of two related quantities. Changes in both the numerator and the denominator influence its value, and even if both components move in the correct direction, the difference in relative change may cause the whole measure to shift in the other direction. In the strong income-dependent mechanism, the naïve median was heavily underestimated because the selection probabilities were negatively correlated with income. Weighting corrected this bias only partially, so the weighted median remained far below the true population median. The median equivalized income of persons below 60% of the overall median was therefore estimated using an unduly small fraction of

poor households. As a result, the median income of poor households was adjusted upward disproportionally less than the overall median. This increased the gap between medians (the numerator) more than the overall median (the denominator). In the selection mechanism weakly dependent on income, these rates of adjustment were favorable because the naïve median was much less underestimated, and the weighting increased it relatively less as well.

In the Gini case, the unfavorable effects of weighting may stem from the grouped nature of the household data and from certain relationships between variables. Regarding the former, in the EU-SILC survey, we had information on household income rather than individual income, which implies that household composition affected the value of the Gini coefficient. The same individuals grouped into larger households would always yield a smaller Gini coefficient (the trapezoid rule used for grouped data inflates the area under the Lorenz curve, and this inflation increases as the number of groups/trapezes decreases). As for the latter, there was a nonlinear relationship in the population between the number of household members and the equivalized income: starting from one-person households, income increased, peaked for three-person households, and then declined. The overall linear correlation coefficient was very close to zero. However, this relationship differed between sampled and non-sampled units. In the selection mechanism strongly dependent on income, the sample correlation between household size and equivalized income was always positive and not negligible (0.11 on average). Under the Internet-dependent mechanism, the sample correlation was always negative, because in the frame population (households having a computer and Internet access), this correlation was negative (-0.13), while in the remainder of the population it was positive (0.21).

The interplay of these two factors was likely responsible for the downward bias of the weighted Gini estimator. In samples drawn under the income-dependent mechanisms, the weights were positively correlated with income (hence the upward correction of mean and median estimators), but also wealthier households tended to be slightly larger in these samples. Larger weights assigned to wealthier households increased the estimated income dispersion, but they also increased the contribution of larger households, thereby reducing the granularity of the weighted data. This household-composition effect slightly outweighed the benefits of increased dispersion, deepening the downward bias of the weighted Gini estimator. In samples from the Internet-dependent mechanism, on the other hand, the weights were negatively correlated with income (leading to reductions in the weighted mean, variance, and, more often than not, the coefficient of variation), and wealthier households tended to be smaller. Although smaller households in these samples ultimately received larger weights on average, the reduction in variability appeared to dominate, causing the estimator to decrease further on average.

6. Conclusions

Inference about finite population parameters from non-probability samples is a challenging task. The researcher must account for the possible dependencies between the selection mechanism and the variables studied, which cause selection bias. Not having the comfort of controlling the probabilities of population elements being included in the sample compels us to rely on modelling. The models should explain the selection mechanism or the studied variables, or both. The keys to the successful usage of these models, aside from knowledge about the studied phenomenon, are the reliable sources of auxiliary information about the population and a set of explanatory variables, measured in both the non-probability sample and the additional source that captures the variability of at least one modelled element. Demographics and social variables are commonly used as auxiliary variables because they are relatively easy to obtain and reliable.

As the conducted experiment showed, the usage of these variables for data weighting reduces the bias for income distribution parameters, but the effect is only certain for simple parameters such as median and mean. For more complex ones, the weighting may well induce additional bias. One of the reasons this happened in the simulation was that the weighting tended to change sample composition in terms of household size, which affected the Gini estimator. Even for simple parameters, the bias was reduced to near zero only for the Internet-related selection mechanism. For mechanisms depending directly on income, the reduction was far from satisfactory. Generally, the bias of estimates of the mean and median predictably reacted to changes in estimation strategy, showing better performance with a larger sample size, a larger set of auxiliary variables, and the use of the doubly robust estimator. However, for other, more complex parameters, the effects were less consistent.

The results of the experiment provide valuable insight into how the bias of different non-probability selection mechanisms and for different parameters is affected by various estimation practices. The experiment was carried out on a large representative sample of a real population. The computer simulation tested realistic scenarios in which sampled observations were not exchangeable, and the estimation of many parameters was done using a single vector of weights. Nevertheless, the experiment has limitations that should be mentioned. The study examined the effect of the sample selection and the estimation methods, but did not account for the change in the mode of the survey – non-probability sample surveys are often carried out online, while data from the EU-SILC 2022 survey was collected face to face or by phone. Furthermore, the EU-SILC data were treated as real observations, but in practice, a significant proportion of income values were imputed.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. and Tourangeau, R., (2013). *Report of the AAPOR Task Force on Non-probability Sampling. Technical report*. Deerfield: The American Association for Public Opinion Research.

Beaumont, J.-F., (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46(1), pp. 1–29.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K., (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50(1), pp. 77–106.

Beaumont, J.-F., Rao, J. N. K., (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies. *The Survey Statistician*, 83, pp. 11–22.

Brzeziński, M., Sałach, K. and Wroński, M., (2020). Wealth inequality in Central and Eastern Europe: Evidence from household survey and rich lists' data combined. *Economics of Transition and Institutional Change*, 28(4), pp. 637–660. Available at: https://doi.org/10.1111/ecot.12257.

Buelens, B., Burger, J. and Brakel, J. A. van den, (2018). Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, 86(2), pp. 322–343. Available at: https://doi.org/10.1111/insr.12253.

Carranza, R., Morgan, M. and Nolan, B., (2023). Top Income Adjustments and Inequality: An Investigation of the EU-SILC. *Review of Income and Wealth*, 69(3), pp. 725–754. Available at: https://doi.org/10.1111/roiw.12591.

Chen, J. K. T., Valliant, R. L. and Elliott, M. R., (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), pp. 657–681. Available at: https://doi.org/10.1111/rssc.12327.

Chen, Y., Li, P. and Wu, C., (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532), pp. 2011–2021. Available at: https://doi.org/10.1080/01621459.2019.1677241.

Cochran, W.G. (1963) *Sampling techniques*. New York: Wiley.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., Leeuw, E. D. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B. and Wenz, A., (2020). A review of conceptual approaches and empirical evidence on probability

and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1), pp. 4–36. Available at: https://doi.org/10.1093/jssam/smz041.

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

Elliott, M. R., Valliant, R., (2017) Inference for Nonprobability Samples. *Statistical Science*, 32(2), pp. 249–264.

European Commission, (2022). *Methodological guidelines and description of EU-SILC target variables*. Eurostat. Available at: https://ec.europa.eu/eurostat/documents/203647/16993001/Methodological+guidelines+2022+operation+v7.pdf/ec6bc779-6462-34aa-a8bd-256f8af34d31?t=1703153300474 (Accessed: 16 July 2024).

Eurostat, (2024). *EU-SILC Microdata*. Available at: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Ferri-García, R., Castro-Martín, L. and Rueda, M. del M., (2021). Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling. *Mathematics and Computers in Simulation*, 186, pp. 19–28. Available at: https://doi.org/10.1016/j.matcom.2020.03.005.

Ferri-García, R., Rueda-Sánchez, J. L., Rueda, M. del M. and Cobo, B., (2024). Estimating response propensities in nonprobability surveys using machine learning weighted models. *Mathematics and Computers in Simulation*, 225, pp. 779–793. Available at: https://doi.org/10.1016/j.matcom.2024.06.012.

Gelman, A., Little, T. C., (1997). Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 23(2), pp. 127–135.

GUS, (2023). *Incomes and living conditions of the population of Poland – report from the EU-SILC survey of 2021*. Warszawa: GUS.

Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14(4), pp. 333–362. Available at: https://doi.org/10.2307/2235923.

Horvitz, D. G., Thompson, D. J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), pp. 663–685.

Kalton, G., (2023). Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day. *Statistics in Transition new series*, 24(3). Available at: https://doi.org/10.59170/stattrans-2023-029.

Kim, J. K., Kwon, Y., (2024). Comments on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples". *Survey Methodology*, 50(1), pp. 57–63.

Kim, J. K., Morikawa, K., (2023). An empirical likelihood approach to reduce selection bias in voluntary samples, arXiv Available at: https://doi.org/10.48550/arXiv.2211.02998.

Kim, J. K., Park, S., Chen, Y. and Wu, C., (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 184(3), pp. 941–963. Available at: https://doi.org/10.1111/rssa.12696.

Kish, L., (1995). *Survey sampling*. New York: John Wiley & Sons.

Lavrakas, P. J., Pennay, D., Neiger, D. and Phillips, B., (2022). Comparing Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia: A Total Survey Error Perspective. *Survey Research Methods*, 16(2), pp. 241–266. Available at: https://doi.org/10.18148/srm/2022.v16i2.7907.

Lee, S., Valliant, R., (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, 37(3), pp. 319–343. Available at: https://doi.org/10.1177/0049124108329643.

Li, Y., (2024). Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples. *Survey Methodology*, 50(1), pp. 37–55.

Liu, A.-C., Scholtus, S. and De Waal, T., (2023). Correcting Selection Bias in Big Data by Pseudo-Weighting. *Journal of Survey Statistics and Methodology*, 11(5), pp. 1181–1203. Available at: https://doi.org/10.1093/jssam/smac029.

Liu, Z., Wang, D. and Pan, Y., (2025). Superpopulation model inference for non probability samples under informative sampling with high-dimensional data. *Communications in Statistics – Theory and Methods*, 54(5), pp. 1370–1390. Available at: https://doi.org/10.1080/03610926.2024.2335543.

Lohr, S. L., (2022). *Sampling. Design and analysis*. Third edition. London, New York: CRC Press.

Marella, D., (2023). Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach. *Journal of Official Statistics*, 39(2), pp. 151–172. Available at: https://doi.org/10.2478/jos-2023-0008.

Meng, X. L., (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, 12(2), pp. 685–726. Available at: https://doi.org/10.1214/18-AOAS1161SF.

Mercer, A. W., Kreuter, F., Keeter, S. and Stuart, E. A., (2017). Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference. *Public Opinion Quarterly*, 81(S1), pp. 250–271. Available at: https://doi.org/10.1093/poq/nfw060.

Mercer, A. W., Lau, A., (2023). Comparing Two Types of Online Survey Samples, *Pew Research Center Methods*, 7 September. Available at: https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/ (Accessed: 15 August 2024).

Neyman, J., (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), pp. 558–625. Available at: https://doi.org/10.2307/2342192.

Rivers, D., (2006). *Understanding People. Sample Matching*. YouGovPolimetrix. Available at: http://www.websm.org/db/12/16528/Web%20Survey%20Bibliography/Understanding_people_Sample_matching/ (Accessed: 26 April 2023).

Rueda, M. del M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L. and Ferri-García, R., (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, 65(2), p. 2200035. Available at: https://doi.org/10.1002/bimj.202200035.

Särndal, C.-E., Swensson, B. and Wretman, J. H., (1997). *Model assisted survey sampling*. New York: Springer.

Szreder, M., Kozłowski, A., (2024). *Wnioskowanie na podstawie prób losowych i nielosowych*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Tillé, Y., (2006). *Sampling algorithms*. New York: Springer.

Valliant, R., (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263. Available at: https://doi.org/10.1093/jssam/smz003.

Valliant, R., Dever, J. A., (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, 40(1), pp. 105–137. Available at: https://doi.org/10.1177/0049124110392533.

Valliant, R., Dever, J. A. and Kreuter, F., (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Valliant, R., Dorfman, A. H. and Royall, R. M., (2000). *Finite population sampling and inference: A prediction approach*. New York: John Wiley & Sons.

Wang, L., Valliant, R. and Li, Y., (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24), pp. 5237–5250. Available at: https://doi.org/10.1002/sim.9122.

Wang, W., Rothschild, D., Goel, S. and Gelman, A., (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), pp. 980–991. Available at: https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wiśniowski, A., Sakshaug, J. W., Ruiz, D. A. P. and Blom, A. G., (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1), pp. 120–147. Available at: https://doi.org/10.1093/jssam/smz051.

Yang, S., Kim, J. K. and Song, R., (2020). Doubly Robust Inference when Combining Probability and Non-Probability Samples with High Dimensional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2), pp. 445–465. Available at: https://doi.org/10.1111/rssb.12354.

Zhang, L. C., (2019). On valid descriptive inference from non-probability sample, *Statistical Theory and Related Fields*, 3(2), pp. 103–113.