*mathematics*

MDPI

*Article*

# An Approach to Integrating a Non-Probability Sample in the Population Census

**Ieva Burakauskaitė and Andrius Čiginas ***

Institute of Data Science and Digital Technologies, Vilnius University, Akademijos Str. 4, LT-08412 Vilnius, Lithuania
* Correspondence: andrius.ciginas@mif.vu.lt

**Abstract:** Population censuses are increasingly using administrative information and sampling as alternatives to collecting detailed data from individuals. Non-probability samples can also be an additional, relatively inexpensive data source, although they require special treatment. In this paper, we consider methods for integrating a non-representative volunteer sample into a population census survey, where the complementary probability sample is drawn from the rest of the population. We investigate two approaches to correcting non-probability sample selection bias: adjustment using propensity scores, which models participation in the voluntary sample, and doubly robust estimation, which has the property of persisting possible misspecification of the latter model. We combine the estimators of population parameters that correct the selection bias with the estimators based on a representative union of both samples. Our analysis shows that the availability of detailed auxiliary information simplifies the applied estimation procedures, which are efficient in the Lithuanian census survey. Our findings also reveal the biased nature of the non-probability sample. For instance, when estimating the proportions of professed religions, smaller religious communities exhibit a higher participation rate than other groups. The combination of estimators corrects such selection bias. Our methodology for combining the voluntary and probability samples can be applied to other sample surveys.

**Keywords:** population census; auxiliary information; missing at random; propensity score adjustment; inverse probability weighting; semiparametric estimation; doubly robust estimation; variance estimation; composite estimation

**MSC:** 62D05



**Citation:** Burakauskaitė, I.; Čiginas, A. An Approach to Integrating a Non-Probability Sample in the Population Census. *Mathematics* **2023**, *11*, 1782. https://doi.org/10.3390/math11081782

Academic Editor: Stefano Bonnini

Received: 6 March 2023
Revised: 30 March 2023
Accepted: 4 April 2023
Published: 8 April 2023



## 1. Introduction

Population censuses are traditionally understood as large-scale surveys conducted once every ten years, where all individuals provide their data through census questionnaires. However, such data collections are expensive and require a lot of other resources. Therefore, new ways of conducting population censuses are being discussed more frequently, particularly as administrative data become more accessible [1–4]. Lithuania is an example, as Statistics Lithuania (the State Data Agency) has conducted the Population and Housing Census 2021 primarily based on administrative data from state registers and information systems.

Unfortunately, certain census variables, such as professed religion or mother tongue, cannot be derived from administrative sources. For these variables, a sample survey could be a compromise that supports the idea of optimizing the population census. Probability sampling methods, along with sample design-based inference, are an accepted approach to surveying finite populations in many areas of statistics [5], particularly in official statistics. Probability samples are also being used in censuses, as seen in [6]. On the other hand, the use of alternative data sources, such as big data or non-probability samples, has

been studied extensively in recent years since these data are cheaper and much easier to obtain [7–10]. However, unlike with probability samples, the inclusion mechanisms into non-probability samples are typically unknown; therefore, inclusion probabilities need to be estimated to correct the sample selection bias. Even a very large non-probability sample may lead to worse estimation results than a small probability sample if the sample selection bias is not taken into account [11]. To our knowledge, non-probability samples have not yet been directly used in censuses, including contemporary corrections of their biases. We present the results of our research on the integration of such a non-probability sample in a population census, which might be useful for other researchers and practitioners working with censuses and sample surveys.

We considered the sampling framework created in the survey of the Lithuanian census: firstly, the data were collected through the voluntary (non-probability) sample and then the probability sample was drawn from the rest of the census population. Our scenario is similar to the one considered in [12,13] but differs from another one often provided in the literature, where the study variables were assumed to be unobserved in a probability sample [14,15]. Moreover, we had access to complete auxiliary information from administrative sources and previous censuses, which may not have always been the case in other surveys. This enabled us to simplify the estimation procedures used in [14] and apply non-probability sample integration similar to [12,13].

Our goal is to efficiently combine both the non-probability and probability samples to estimate the parameters of interest. In Section 2.3.1, we consider a natural post-stratified generalized regression (calibrated) estimator of population means, as described in [12]. This estimator is based on the union of the probability and voluntary samples, with inclusion probabilities initially set to one for units in the latter sample. In Section 2.3.2, we explore an alternative to the post-stratified estimator, as presented in [14], which is the inverse probability weighting estimator based on estimated inclusion probabilities (propensity scores) for the non-probability sample. We adapt the methodology of [14] to our framework to derive the variance estimation formula for this estimator. Finally, we combine both estimators in Section 2.3.5 by taking into account their estimated variances. In Section 2.3.4, we investigate the doubly robust estimator for the non-probability sample, which provides protection against possible misspecification of the propensity score model [14]. This estimator incorporates model-based prediction estimators for the parameters of interest and exploits complete auxiliary information. We combine the doubly robust estimator with an analogous generalized difference estimator from [16], which we describe in Section 2.3.3. Our aim is to determine which combination works best, at least in our application to the population census.

The application of the considered estimators to the Lithuanian census survey is elaborated on in Section 3. A discussion of the results of the study and some future insights are reviewed in Section 4. The most relevant conclusions are outlined in Section 5. By summarizing our findings supported by the analysis of the real census data, 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.

## 2. Methods

### *2.1. Sampling from the Finite Population*

We consider any continuous or binary study variable $y$ with the fixed values $y_1, \ldots, y_N$ in a finite population $\mathcal{U} = \{1, \ldots, N\}$ of size $N$. We estimate the population mean or proportion

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

To estimate population parameters (1), assume that, at first, a non-probability sample $s_A$ of size $n_A$ is obtained from $\mathcal{U}$, and 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 $\mathcal{U} \backslash \{s_A\}$ afterward. We can then interpret that the union of both samples $s = s_A \cup s_B$ of size

$n = n_A + n_B$ is drawn according to the probability sampling design $p(\cdot)$ with 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, we use the notation $P_p$, $E_p$, and $V_p$ to denote the probability, expectation, and variance, calculated according to the randomness induced by $p(\cdot)$, respectively. We write $d_k = 1/\pi_k$ to denote the sampling weights.

### *2.2. Auxiliary Data and Outcome Regression Model*

We associate with the unit $k \in \mathcal{U}$ the values $\mathbf{x}_k$ of the auxiliary variables $\mathbf{x}$, and assume that these values are known for all population units. Hence, the complete auxiliary information is available.

Suppose that the relationship between the variables $y$ and $\mathbf{x}$ can be described by a semiparametric outcome regression model $\xi$:

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

where $\boldsymbol{\beta}$ and $\sigma^2$ 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, for example, $m(\mathbf{x}_k, \boldsymbol{\beta}) = \mathbf{x}_k'\boldsymbol{\beta}$. Here, $E_\xi$ and $V_\xi$ denote the expectation and variance with respect to the model $\xi$. We assume (without any loss of generality) that 1 is the first component of the vector $\mathbf{x}_k$ for all $k \in \mathcal{U}$.

### *2.3. Estimation of Population Parameters*

#### 2.3.1. Post-Stratified Generalized Regression Estimator

Let us consider the combined sample $s$ with the accompanying probability sampling design $p(\cdot)$. Taking $m(\mathbf{x}_k, \boldsymbol{\beta}) = \mathbf{x}_k'\boldsymbol{\beta}$ in (2), we have the linear regression model, which is used to build the generalized regression estimator [17]

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

of (1), where

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

with positive constants $c_k$, for instance, $c_k = v_k^2$. The quantity $\widehat{\mathbf{B}}$ estimates the population characteristic

$$\mathbf{B} = \left(\sum_{k\in\mathcal{U}} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \sum_{k\in\mathcal{U}} \frac{\mathbf{x}_k y_k}{c_k},$$

which is, in turn, the generalized least squares estimator of $\boldsymbol{\beta}$ of the linear regression model, which is also called the assisting model. Estimator (3) is equivalent to the calibrated estimator [18]

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

often used in practice, where the calibration 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, i.e., $E_p(\hat{\mu}^{\mathrm{GR}}) \approx \mu$. The generalized regression estimator (3) is also referred to as the post-stratified estimator in [12], with two

post-strata, i.e., $s_A$ and $\mathcal{U}\backslash\{s_A\}$. The authors of [12] argue that such estimation is efficient if the non-probability sample $s_A$ is very large.

According to [17], 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 inclusions into the sample probabilities.

2.3.2. Inverse Probability Weighting Estimator Based on the Propensity Score Model

Let us consider only the non-probability sample $s_A$. The non-probability sample itself does not represent the target population, and naive estimators based on it are typically biased [11]. The main obstacle is the unknown selection mechanism for a unit to be included in the sample.

Let $R_k = \mathbb{I}(k\in s_A)$ be the indicator variable for a unit $k\in\mathcal{U}$ selected to the sample $s_A$. The probabilities

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

are called the propensity scores, where the subscript $q$ refers to the propensity score model. Probabilities (6) are analogous to the inclusion into the sample probabilities $\pi_k$ (for probability samples) since they describe the inclusion into the sample $s_A$. The propensity scores $\pi_k^A$, $k\in s_A$, need to be estimated before using them to weigh the units of the non-probability sample.

The following assumptions are considered to simplify the propensity score model [14]:

A1 The indicator $R_k$ and the study variable $y_k$ are independent given the covariates $\mathbf{x}_k$.
A2 All units have 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 $\mathbf{x}_k$ and $\mathbf{x}_l$ for $k\neq l$.

Due to assumption A1, 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 $k\in\mathcal{U}$, and the selection mechanism is called ignorable. It is similar to the notion of missing at random (MAR) used in missing data analyses [19].

We model the propensity scores $\pi_k^A = P_q(R_k=1\,|\,\mathbf{x}_k)$ parametrically using the inverse logit function

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

where $\boldsymbol{\theta}$ is the model parameter with the unknown true value $\boldsymbol{\theta}_0$. The propensity score estimates $\hat{\pi}_k^A$ under the logistic regression model (7) are obtained from the maximum likelihood estimator $\hat{\pi}_k^A = \pi(\mathbf{x}_k,\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function

$$l(\boldsymbol{\theta}) = \sum_{k\in s_A}\log\left\{\frac{\pi(\mathbf{x}_k,\boldsymbol{\theta})}{1-\pi(\mathbf{x}_k,\boldsymbol{\theta})}\right\} + \sum_{k\in\mathcal{U}}\log\{1-\pi(\mathbf{x}_k,\boldsymbol{\theta})\} = \sum_{k\in s_A}\mathbf{x}_k'\boldsymbol{\theta} - \sum_{k\in\mathcal{U}}\log\{1+\exp(\mathbf{x}_k'\boldsymbol{\theta})\}.$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is found by solving the score equations

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta}) = \sum_{k\in\mathcal{U}}\{R_k-\pi(\mathbf{x}_k,\boldsymbol{\theta})\}\mathbf{x}_k = 0.$$

A conventional way to do it is to apply the Newton–Raphson iterative procedure.

The estimated propensity scores $\hat{\pi}_k^A = \pi(\mathbf{x}_k,\hat{\boldsymbol{\theta}})$, $k\in s_A$, are then used to compute the inverse probability weighting (IPW) estimator [14]

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

of the population mean $\mu$. This is the adaptation of the Hájek estimator used for the probability samples. Estimator (8) can correct the sample selection bias efficiently if the propensity score model is well-specified.

We construct the estimator of variance $V_q(\hat{\mu}^{\text{IPW}})$ using asymptotic properties of estimator (8). Let $\mathcal{U}_\nu$ be a sequence of finite populations of size $N_\nu$, indexed by $\nu$. For each $\mathcal{U}_\nu$, there is an associated non-probability sample $s_{A,\nu}$ of size $n_{A,\nu}$. The population size $N_\nu \to \infty$ and the sample size $n_{A,\nu} \to \infty$ as $\nu \to \infty$. Further, the index $\nu$ is suppressed to simplify the notation. Consider the following regularity conditions [14]:

C1 The population size $N$ and the sample size $n_A$ satisfy $\lim_{N\to\infty} n_A/N = f_A \in (0,1)$.

C2 There exist $c_1$ and $c_2$ such that $0 < c_1 < N\pi_k^A/n_A \le c_2$ for all units $k \in \mathcal{U}$.

C3 The finite population and the propensity scores satisfy $N^{-1}\sum_{k\in\mathcal{U}} y_k^2 = O(1)$, as well as $N^{-1}\sum_{k\in\mathcal{U}} \|\mathbf{x}_k\|^3 = O(1)$, and $N^{-1}\sum_{k\in\mathcal{U}} \pi_k^A(1-\pi_k^A)\mathbf{x}_k\mathbf{x}_k'$ is a positive definite matrix.

**Proposition 1.** *Under assumptions A1–A3 and regularity conditions C1–C3, and assuming the logistic regression model (7) for the propensity scores, estimator (8) is asymptotically unbiased, i.e.,* $\hat{\mu}^{\text{IPW}} - \mu = O_p(n_A^{-1/2})$*, and an asymptotic variance of (8) can be derived as*

$$V_q(\hat{\mu}^{\text{IPW}}) = V^{\text{IPW}} + o(n_A^{-1}),$$

*where*

$$V^{\text{IPW}} = \frac{1}{N^2}\sum_{k\in\mathcal{U}}(1-\pi_k^A)\pi_k^A\left(\frac{y_k-\mu}{\pi_k^A} - \mathbf{b}'\mathbf{x}_k\right)^2$$

*with* $\pi_k^A = \pi(\mathbf{x}_k,\boldsymbol{\theta}_0) = \exp(\mathbf{x}_k'\boldsymbol{\theta}_0)/\left(1+\exp(\mathbf{x}_k'\boldsymbol{\theta}_0)\right)$ *and*

$$\mathbf{b}' = \left\{\sum_{k\in\mathcal{U}}(1-\pi_k^A)(y_k-\mu)\mathbf{x}_k'\right\}\left\{\sum_{k\in\mathcal{U}}\pi_k^A(1-\pi_k^A)\mathbf{x}_k\mathbf{x}_k'\right\}^{-1}.$$

**Proof.** The proposition is actually the corollary of Theorem 1 in [14] for complete auxiliary data. An inspection of the proof of the latter theorem leads to simpler assumptions required for the proposition statement. □

Using the asymptotic variance from Proposition 1, the variance of estimator (8) can be estimated by

$$\widehat{V}^{\text{IPW}} = \frac{1}{(\widehat{N}^A)^2}\sum_{k\in s_A}(1-\hat{\pi}_k^A)\left(\frac{y_k-\hat{\mu}^{\text{IPW}}}{\hat{\pi}_k^A} - \widehat{\mathbf{b}}'\mathbf{x}_k\right)^2, \tag{9}$$

where

$$\widehat{\mathbf{b}}' = \left\{\sum_{k\in s_A}\left(\frac{1}{\hat{\pi}_k^A}-1\right)(y_k-\hat{\mu}^{\text{IPW}})\mathbf{x}_k'\right\}\left\{\sum_{k\in\mathcal{U}}\hat{\pi}_k^A(1-\hat{\pi}_k^A)\mathbf{x}_k\mathbf{x}_k'\right\}^{-1},$$

given the non-probability sample $s_A$.

#### 2.3.3. Generalized Difference Estimator

Consider the combined sample $s$ together with the sampling design $p(\cdot)$. If the outcome regression model (2) is not assumed to be linear, one can apply the generalized difference estimator [16]

$$\hat{\mu}^{\text{GD}} = \frac{1}{N}\left(\sum_{k\in s} d_k(y_k - m(\mathbf{x}_k,\hat{\boldsymbol{\beta}})) + \sum_{k\in\mathcal{U}} m(\mathbf{x}_k,\hat{\boldsymbol{\beta}})\right) \tag{10}$$

to estimate the population mean $\mu$. The estimator exploits the available complete auxiliary data. Here, $\hat{\boldsymbol{\beta}}$ can be the quasi-maximum likelihood estimator of the regression parameter $\boldsymbol{\beta}$ based on the dataset $\{(y_k,\mathbf{x}_k),\, k\in s\}$ by [20].

A design-consistent estimator of the variance $V_p(\hat{\mu}^{\mathrm{GD}})$ is provided in [16] as

$$\hat{\psi}^{\mathrm{GD}} = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \left( 1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right) \frac{(y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}))(y_l - m(\mathbf{x}_l, \hat{\boldsymbol{\beta}}))}{\pi_k \pi_l}. \tag{11}$$

2.3.4. Doubly Robust Estimator

Let us use only the non-probability sample $s_A$. A drawback of the IPW estimator (8) may be its sensitivity to a misspecified model for the propensity scores, especially if some units have very small values in $\hat{\pi}_k^A$ [10]. The efficiency and robustness of the IPW estimator can be improved by incorporating outcome regression model (2), also called the prediction model. The doubly robust (DR) estimator for the mean $\mu$ is [14]

$$\hat{\mu}^{\mathrm{DR}} = \frac{1}{\widehat{N}^A} \sum_{k \in s_A} \frac{y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}})}{\hat{\pi}_k^A} + \frac{1}{N} \sum_{k \in \mathcal{U}} m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}). \tag{12}$$

Due to the so-called model transportability implied by the assumption A1 [21], model (2), fitted using standard methods based on the dataset $\{(y_k, \mathbf{x}_k),\ k \in s_A\}$, can be applied to compute the predicted values $m(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$ for all $k \in \mathcal{U}$ used in (12). A doubly robust estimator (12) is an analog of the generalized difference estimator (10).

We turn to the estimation of variance $V_q(\hat{\mu}^{\mathrm{DR}})$. Let $\boldsymbol{\beta}_0$ be the unknown true value of parameter $\boldsymbol{\beta}$ in the prediction model. Consider additional regularity conditions imposed on the mean function $m(\mathbf{x}, \boldsymbol{\beta})$ as given in [14]:

C4 For each $\mathbf{x}$, $\partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ is continuous in $\boldsymbol{\beta}$ and $|\partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}| \leq h(\mathbf{x}, \boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}_0$, and $N^{-1} \sum_{k \in \mathcal{U}} h(\mathbf{x}_k, \boldsymbol{\beta}_0) = O(1)$.

C5 For each $\mathbf{x}$, $\partial^2 m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'$ is continuous in $\boldsymbol{\beta}$ and $\max_{i,j} |\partial^2 m(\mathbf{x}, \boldsymbol{\beta})/\partial \beta_i \partial \beta_j| \leq k(\mathbf{x}, \boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}_0$, and $N^{-1} \sum_{k \in \mathcal{U}} k(\mathbf{x}_k, \boldsymbol{\beta}_0) = O(1)$.

**Proposition 2.** *Estimator* (12) *is doubly robust in the sense that it is a consistent estimator of the mean $\mu$ if either the propensity score model or the prediction model is correctly specified. Under assumptions A1–A3 and regularity conditions C1–C5, and assuming correctly specified logistic regression model* (7) *for the propensity scores, an asymptotic variance of* (12) *can be derived as*

$$V_q(\hat{\mu}^{\mathrm{DR}}) = V^{\mathrm{DR}} + o(n_A^{-1}),$$

*where*

$$V^{\mathrm{DR}} = \frac{1}{N^2} \sum_{k \in \mathcal{U}} (1 - \pi_k^A) \pi_k^A \left( \frac{y_k - m(\mathbf{x}_k, \boldsymbol{\beta}_0) - h_N}{\pi_k^A} - \mathbf{b}_{\mathrm{DR}}' \mathbf{x}_k \right)^2 \tag{13}$$

*with* $\pi_k^A = \pi(\mathbf{x}_k, \boldsymbol{\theta}_0) = \exp(\mathbf{x}_k' \boldsymbol{\theta}_0)/\left(1 + \exp(\mathbf{x}_k' \boldsymbol{\theta}_0)\right)$ *and*

$$\mathbf{b}_{\mathrm{DR}}' = \left\{ \sum_{k \in \mathcal{U}} (1 - \pi_k^A)(y_k - m(\mathbf{x}_k, \boldsymbol{\beta}_0) - h_N) \mathbf{x}_k' \right\} \left\{ \sum_{k \in \mathcal{U}} \pi_k^A (1 - \pi_k^A) \mathbf{x}_k \mathbf{x}_k' \right\}^{-1},$$

$$h_N = \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - m(\mathbf{x}_k, \boldsymbol{\beta}_0)).$$

**Proof.** The proposition follows from Theorem 2 of [14] for the complete auxiliary data. □

**Remark 1.** *Regularity conditions C4–C5 are redundant for the linear outcome regression model.*

Using asymptotic variance (13), a simple plug-in variance estimator for doubly robust estimator (12) is

$$\widehat{V}^{\mathrm{DR}} = \frac{1}{(\widehat{N}^A)^2} \sum_{k \in s_A} (1 - \hat{\pi}_k^A) \left( \frac{y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}) - \hat{h}_N}{\hat{\pi}_k^A} - \widehat{\mathbf{b}}'_{\mathrm{DR}} \mathbf{x}_k \right)^2, \tag{14}$$

where

$$\widehat{\mathbf{b}}'_{\mathrm{DR}} = \left\{ \sum_{k \in s_A} \left( \frac{1}{\hat{\pi}_k^A} - 1 \right) (y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}) - \hat{h}_N) \mathbf{x}'_k \right\} \left\{ \sum_{k \in \mathcal{U}} \hat{\pi}_k^A (1 - \hat{\pi}_k^A) \mathbf{x}_k \mathbf{x}'_k \right\}^{-1},$$

$$\hat{h}_N = \frac{1}{\widehat{N}^A} \sum_{k \in s_A} \frac{y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}})}{\hat{\pi}_k^A},$$

given the non-probability sample $s_A$.

2.3.5. Composite Estimators

We linearly combine design-based post-stratified estimators (3) and (10) based on the combined sample $s$ with, correspondingly, model-based IPW and DR estimators supported only by the non-probability sample $s_A$. We consider two composite estimators

$$\hat{\mu}^{\mathrm{C1}} = \hat{\lambda}_1 \hat{\mu}^{\mathrm{GR}} + (1 - \hat{\lambda}_1) \hat{\mu}^{\mathrm{IPW}} \quad \text{with} \quad \hat{\lambda}_1 = \frac{\widehat{V}^{\mathrm{IPW}}}{\hat{\psi}^{\mathrm{GR}} + \widehat{V}^{\mathrm{IPW}}} \tag{15}$$

and

$$\hat{\mu}^{\mathrm{C2}} = \hat{\lambda}_2 \hat{\mu}^{\mathrm{GD}} + (1 - \hat{\lambda}_2) \hat{\mu}^{\mathrm{DR}} \quad \text{with} \quad \hat{\lambda}_2 = \frac{\widehat{V}^{\mathrm{DR}}}{\hat{\psi}^{\mathrm{GD}} + \widehat{V}^{\mathrm{DR}}} \tag{16}$$

of population mean (1). Similar combinations are investigated in [13]. Here, the quantities $\hat{\lambda}_1$ and $\hat{\lambda}_2$ estimate optimal coefficients of the combinations, ignoring covariance terms since the estimators we combine, in principle, do not have common sources of randomness. Compositions (15) and (16) give more weight to the estimators with smaller variances. The respective variance estimators are

$$\widehat{V}^{\mathrm{C1}} = \hat{\lambda}_1 \hat{\psi}^{\mathrm{GR}} \tag{17}$$

and

$$\widehat{V}^{\mathrm{C2}} = \hat{\lambda}_2 \hat{\psi}^{\mathrm{GD}}. \tag{18}$$

The interpretation of variance estimators (17) and (18) is that the variances of the design-based estimators may be reduced by the factors $\hat{\lambda}_1$ and $\hat{\lambda}_2$, respectively.

## 3. Application to the Survey of the Lithuanian Census

### *3.1. Motivation*

In 2020, the COVID-19 pandemic highlighted the need to promptly produce statistical information from national statistical institutions. This led Statistics Lithuania (the State Data Agency) to take on a new role as the governing organization for state data, forming a unified database of the main state registers and information systems with a vast amount of data that are ready to be used for statistical purposes. Therefore, Statistics Lithuania was able to carry out the following 2021 census 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, a statistical survey for such data collection had to be launched. Hence, a statistical survey on population by ethnicity, native language, and religion was conducted in 2021. It aimed to evaluate population proportions for the following variables: religion professed

(16 categories), mother tongue (more than 12 categories), knowledge of other languages (16 languages), and ethnicity. For the latter variable, mass imputation was used since relevant information was known from the Ethnicity Register for approximately 87% of the census population. The research was conducted to achieve the objective of efficiently estimating these proportions by exploiting complete data from the previous censuses and other auxiliary information.

### *3.2. Sample Selection*

The survey sample $s \subset \mathcal{U}$ was drawn and consisted of three parts, $s = s_A \cup s_O \cup s_B$:

(i) At first, a voluntary online survey was carried out from 15 January to 28 February, 2021, which allowed for the collection of statistical data from approximately 2% of the census population (about 54,000 respondents), resulting in the non-probability sample $s_A$.

(ii) After the end of the online survey, a sampling frame for probability sampling was constructed. It excluded certain addresses, e.g., if at least one individual from the address participated in the online survey, if it was an institution, if more than 15 individuals were permanent residents, among other rules. These units, which were not included in the sampling frame, comprised the part $s_O$ of the sample $s$.

(iii) Lastly, the probability sample $s_B$ was drawn from the sampling frame $\mathcal{U} \setminus \{s_A \cup s_O\}$, which was divided into $H = 113$ strata according to the municipality intersected with the area of residence, i.e., urban or rural. The number of addresses sampled from a particular stratum was proportional to the size of the stratum, resulting in around 40,000 addresses sampled from the Population Register in total; approximately 6% of the census population was interviewed through the telephone survey (about 171,000 respondents).

The working sampling design $p(\cdot)$ is characterized by the inclusion probabilities: $\pi_k = 1$ if $k \in s_A \cup s_O$, and

$$\pi_k \approx \frac{m_k n'_h}{N'_h} \quad \text{if} \quad k \in s_B,$$

where $N'_h$ denotes the size of the $h$th 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.

### *3.3. Imputation of Missing Values*

The response rate in the probability sample $s_B$ reached approximately 88%. Missing values in the whole sample $s$ were filled in using three imputation methods: historical, deductive, and k-nearest neighbor.

Missing values were first filled in using historical information from the 2011 and 2001 censuses consecutively, as variables of interest were fully known for the populations of those censuses. The remaining missing values accounted for 2.3% of the sample.

Additional sociodemographic characteristics of previous and current censuses, such as age, gender, marital status, household structure, country of birth, citizenship, education, and employment status, were used for further deductive imputation. For instance, if the same religion was observed for each household member except one, the corresponding religion was imputed where missing. After the deductive imputation, only 0.3% of the sample remained with missing values.

Eventually, the remaining missing values in the sample were filled in by applying the $k$-nearest neighbor method [22].

### *3.4. Application to Religion Proportions*

We focus on the non-probability sample integration for the estimation of religion proportions as the results are similar for every proportion of interest.

When we obtained the non-probability sample from the online survey, the question arose if the collected data could be used for estimation. We first checked the representativeness of the voluntary sample using sociodemographic characteristics known for the entire 2021 census population. A comparison of some proportions of sociodemographic characteristics between the voluntary sample and the whole population showed the biased nature of the non-probability sample. The results provided in Table 1 suggest that people with higher education, as well as those who are employed and married, tend to participate in such online surveys. Another interesting observation made was the willingness of some ethnic communities to participate in the online survey and represent their community. For instance, Polish people in Lithuania accounted for 35% of the voluntary sample but only about 7% of the whole population.

**Table 1.** Comparison of proportions of some sociodemographic characteristics in the voluntary sample and the whole population.

| | | 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 | $\geq$30, <50 | 0.37 | 0.27 | 37 |
| Marital status | married | 0.52 | 0.42 | 25 |
| Gender | male | 0.41 | 0.46 | −11 |
| Ethnicity | Lithuanian | 0.56 | 0.85 | −34 |
| Education | (lower) secondary | 0.24 | 0.37 | −35 |
| Education | primary | 0.09 | 0.20 | −55 |

Additionally, we compared the religion proportions of 2011 religion in the 2021 census population for the online survey respondents and the entire population; see Table 2. It was observed that the representatives of smaller religious communities were more likely to participate in the survey. For instance, the proportion of the Karaites religious community in the voluntary sample was 1307% larger than the corresponding proportion in the population.

**Table 2.** Comparison of religion proportions in the voluntary sample and the whole population.

| | Voluntary Sample | Population | Difference in % |
|---|---|---|---|
| Karaites | 0.00130 | 0.00009 | 1307 |
| New Apostolic Church | 0.00161 | 0.00014 | 1049 |
| Evangelical Reformed Believers | 0.00833 | 0.00207 | 302 |
| Other | 0.01596 | 0.00514 | 211 |
| Pentecostalists | 0.00198 | 0.00067 | 194 |
| Greek Catholics (Uniats) | 0.00048 | 0.00021 | 131 |
| Evangelical Lutherans | 0.01311 | 0.00585 | 124 |
| Judaists | 0.00074 | 0.00035 | 112 |
| Baptists and Free Churches | 0.00083 | 0.00048 | 74 |
| Sunni Muslims | 0.00130 | 0.00085 | 52 |
| Not indicated | 0.07621 | 0.10090 | −24 |
| Seventh Day Adventist Church | 0.00026 | 0.00032 | −20 |
| None | 0.07580 | 0.06424 | 18 |
| Old Believers | 0.00615 | 0.00683 | −10 |
| Orthodox | 0.04047 | 0.03787 | 7 |
| Roman Catholics | 0.75548 | 0.77398 | −2 |

The sociodemographic variables of Table 1 and the religion variables of Table 2 contain information that can explain the chance of being selected in the voluntary sample. Hence, these variables were used as covariates in the propensity score model.

To estimate the religion proportions in the 2021 census population, we first considered the post-stratified generalized regression and generalized difference estimators given by (3) and (10), respectively. The calibrated weights in (4) were calculated by taking such auxiliary information as binary variables on age groups, gender, and religions professed in 2011 intersected with counties in the calibration equations, while the same auxiliary variables as in the propensity score model were used for estimator (10). Comparing the results of the post-stratified calibrated estimator with the proportions of the previous censuses in Table 3 (and based on external evaluations), the estimator $\hat{\mu}^{\mathrm{GR}}$ tends to underestimate smaller religious communities due to the lack of data. On the other hand, the generalized difference estimator $\hat{\mu}^{\mathrm{GD}}$ seems to produce slightly higher estimates for the majority of these smaller religions.

**Table 3.** Religion proportions in 2001, 2011, and 2021 census populations.

| | $\mu^{(2001)}$ | $\mu^{(2011)}$ | $\hat{\mu}^{\mathrm{GR}}$ | $\hat{\mu}^{\mathrm{GD}}$ |
|---|---|---|---|---|
| Roman Catholics | 0.78391 | 0.77233 | 0.73664 | 0.74101 |
| Not indicated | 0.05671 | 0.10112 | 0.15701 | 0.15025 |
| None | 0.09696 | 0.06146 | 0.05408 | 0.05477 |
| Orthodox | 0.04150 | 0.04113 | 0.03433 | 0.03482 |
| Old Believers | 0.00806 | 0.00767 | 0.00434 | 0.00419 |
| Evangelical Lutherans | 0.00565 | 0.00604 | 0.00389 | 0.00398 |
| Other | 0.00282 | 0.00493 | 0.00566 | 0.00625 |
| Evangelical Reformed Believers | 0.00208 | 0.00221 | 0.00122 | 0.00126 |
| Pentecostalists | 0.00037 | 0.00061 | 0.00117 | 0.00158 |
| Sunni Muslims | 0.00075 | 0.00089 | 0.00058 | 0.00064 |
| Baptists and Free Churches | 0.00034 | 0.00044 | 0.00017 | 0.00016 |
| Judaists | 0.00039 | 0.00040 | 0.00025 | 0.00029 |
| Greek Catholics (Uniats) | 0.00010 | 0.00023 | 0.00030 | 0.00038 |
| Seventh Day Adventist Church | 0.00016 | 0.00030 | 0.00014 | 0.00013 |
| New Apostolic Church | 0.00012 | 0.00014 | 0.00015 | 0.00020 |
| Karaites | 0.00008 | 0.00010 | 0.00008 | 0.00010 |

As we observed relatively more representatives of minor religions in the voluntary sample (see Table 2), we expected smaller variances of the estimators based only on this non-probability sample with a condition that a selection bias is properly corrected. The IPW estimator is designed to correct such bias by incorporating the propensity scores evaluated using the auxiliary variables of Tables 1 and 2.

We integrated the non-probability sample through the combination $\hat{\mu}^{\mathrm{C1}}$ of the post-stratified generalized regression (calibrated) and IPW estimators. According to Table 4, it seems that the first composite estimator corrected the underestimation. Alternatively, we considered the combination $\hat{\mu}^{\mathrm{C2}}$ of the generalized difference estimator with its analog DR estimator based on the auxiliary variables of Tables 1 and 2. The second composite estimator seems to have produced even higher estimates for smaller religious communities; however, it also came with larger variances (see Table 5).

Nevertheless, as the calibrated and generalized difference estimators seemed to evaluate larger proportions of interest accurately, they were used in compositions (15) and (16) with weights equal to 1. That is, for religions None, Not indicated, and Roman Catholics, the proportions were evaluated using only the design-based calibrated or generalized difference estimators.

**Table 4.** Religion proportion estimates in 2021 census population.

| | $\hat{\mu}^{\mathbf{GR}}$ | $\hat{\mu}^{\mathbf{C1}}$ | $\hat{\mu}^{\mathbf{GD}}$ | $\hat{\mu}^{\mathbf{C2}}$ |
|---|---|---|---|---|
| Roman Catholics | 0.73664 | 0.73349 | 0.74101 | 0.73811 |
| Not indicated | 0.15701 | 0.15452 | 0.15025 | 0.14832 |
| None | 0.05408 | 0.05319 | 0.05477 | 0.05401 |
| Orthodox | 0.03433 | 0.03804 | 0.03482 | 0.03592 |
| Old Believers | 0.00434 | 0.00503 | 0.00419 | 0.00486 |
| Evangelical Lutherans | 0.00389 | 0.00460 | 0.00398 | 0.00475 |
| Other | 0.00566 | 0.00636 | 0.00625 | 0.00709 |
| Evangelical Reformed Believers | 0.00122 | 0.00151 | 0.00126 | 0.00175 |
| Pentecostalists | 0.00117 | 0.00126 | 0.00158 | 0.00191 |
| Sunni Muslims | 0.00058 | 0.00069 | 0.00064 | 0.00094 |
| Baptists and Free Churches | 0.00017 | 0.00024 | 0.00016 | 0.00037 |
| Judaists | 0.00025 | 0.00031 | 0.00029 | 0.00051 |
| Greek Catholics (Uniats) | 0.00030 | 0.00034 | 0.00038 | 0.00060 |
| Seventh Day Adventist Church | 0.00014 | 0.00017 | 0.00013 | 0.00031 |
| New Apostolic Church | 0.00015 | 0.00017 | 0.00020 | 0.00035 |
| Karaites | 0.00008 | 0.00009 | 0.00010 | 0.00022 |

Table 5 provides comparisons of the relative percent difference between (i) the smoothed version of variance estimator (5) and variance estimator (17), (ii) the smoothed version of variance estimator (11) and variance estimator (18), and (iii) variance estimators (17) and (18). We used smoothed variances $\hat{\psi}^{GRs}$ and $\hat{\psi}^{GDs}$ instead of $\hat{\psi}^{GR}$ and $\hat{\psi}^{GD}$ in compositions (15) and (16), respectively, due to the estimation of very small proportions. For the smoothing of variance (5), similarly as in [23], we assumed that $V_p(\hat{\mu}^{\mathrm{GR}}) \approx K\widetilde{N}^{\gamma}$, with $\widetilde{N}$ as the sum of values of 2011 variable of interest in the 2021 census population. Parameters $K > 0$ and $\gamma \in \mathbb{R}$ were evaluated through a log–log regression, using the data pairs $(\hat{\psi}^{\mathrm{GR}}, \widetilde{N})$ calculated from all categories of the variable of interest. The smoothing of variance (11) was performed analogously.

**Table 5.** Comparison of the relative difference (in %): (i) $(\hat{\psi}^{\mathrm{GRs}} - \widehat{V}^{\mathrm{C1}})/\widehat{V}^{\mathrm{C1}}$, (ii) $(\hat{\psi}^{\mathrm{GDs}} - \widehat{V}^{\mathrm{C2}})/\widehat{V}^{\mathrm{C2}}$, (iii) $(\widehat{V}^{\mathrm{C1}} - \widehat{V}^{\mathrm{C2}})/\widehat{V}^{\mathrm{C2}}$.

| | **(i)** | **(ii)** | **(iii)** |
|---|---|---|---|
| New Apostolic Church | 1 | 6 | −88 |
| Karaites | 4 | 43 | −88 |
| Greek Catholics (Uniats) | 2 | 16 | −84 |
| Seventh Day Adventist Church | 4 | 22 | −79 |
| Judaists | 6 | 35 | −76 |
| Pentecostalists | 1 | 3 | −73 |
| Baptists and Free Churches | 9 | 42 | −71 |
| Sunni Muslims | 6 | 20 | −65 |
| Evangelical Reformed Believers | 6 | 12 | −46 |
| Other | 2 | 2 | −14 |
| Evangelical Lutherans | 7 | 7 | −7 |
| Old Believers | 19 | 22 | 4 |
| Orthodox | 18 | 9 | 146 |
| None | 0 | 0 | 261 |
| Not indicated | 0 | 0 | 366 |
| Roman Catholics | 0 | 0 | 1389 |

The compositions given by (15) and (16) assign more weight to estimators with a smaller variance. The first two numerical columns in Table 5, which correspond to cases (i)–(ii), provide an indication of how much composite estimators can improve the estimation accuracy of the calibrated and generalized difference estimators, respectively. The last column of the table includes the relative percent differences between the variance estimates of composite estimators $\hat{\mu}^{\mathrm{C1}}$ and $\hat{\mu}^{\mathrm{C2}}$. The first composite estimator gives more

satisfactory results for the proportions of smaller religious communities. However, the second composite estimator seems to correct the estimation accuracy of proportions of larger religious groups.

## 4. Discussion

The generalized regression and generalized difference estimators considered here are traditional model-assisted estimators based on the union of the non-probability and probability samples, together with the probability sampling design $p(\cdot)$. Since the selection bias is corrected through the known inclusion probabilities, the application of these estimators may lead to valid design-based inferences. However, the latter estimators incorporate unweighted units of the non-probability sample, which does not add more efficiency unless the size of the non-probability sample is of the same magnitude as the population size [12]. In the application to the Lithuanian census, the voluntary sample covers about 2% of the population, and such a contribution is too small.

Meanwhile, IPW and DR estimators exploit the non-probability sample in a more advanced way, i.e., through the propensity score and prediction models, and we may benefit from combining them with traditional design-based estimators. There are various ways to combine different data sources and estimators [24]; however, we choose to optimize the linear combinations of estimators by taking into account their estimated variances. This approach appears to be efficient in the application when estimating smaller population proportions that require larger sample sizes. Indeed, the first composite estimator improves the estimation accuracy for proportions of smaller religious communities as the estimated variances of the composition are smaller than the estimated variances of the generalized regression estimator by up to 19%. Alternatively, we consider the combination of the generalized difference estimator with the DR estimator as it allows us to better leverage the available complete auxiliary data through the prediction model. Surprisingly, while this composite estimator improves the estimation accuracy for proportions of larger religious communities, it does not give satisfactory results for smaller religions.

In the application, the detailed data available from administrative registers and previous complete censuses allowed us to efficiently apply the model-based IPW and DR estimators integrating the non-probability sample. In regard to future censuses, it is worth considering the possibility of collecting much larger non-probability samples by promoting voluntary participation more.

Our study is based on a strong MAR assumption for the variable of interest in the propensity score model. Although the analysis of the Lithuanian census data allowed us to identify variables that clearly explain voluntary survey participation (and, thus, the propensity score model may be specified quite well), this assumption might be relaxed in future research. In this way, future investigations could explore data integration under the assumption of the non-ignorable selection mechanism.

## 5. Conclusions

The non-probability and probability samples can be combined into a single sample and used in design-based post-stratified estimators of parameters. This approach is a safe but inefficient way to exploit the non-probability sample.

The post-stratified estimators are, therefore, linearly combined with the model-based estimators based only on the non-probability sample. By applying the estimated optimal combinations to the Lithuanian census survey, there is a significant improvement in estimating the population proportions.

The success of integrating a voluntary sample into a census survey depends on the availability of proper auxiliary information, such as complete data from previous censuses. In the future, such information could be obtained by collecting much larger non-probability samples. In addition, it would then be possible to forego the selection of a probability sample.

Applications such as ours may be more efficient if the MAR assumption for the propensity score model is abandoned. This would lead to more complex estimation procedures that can be explored in the future.

**Author Contributions:** Conceptualization, A.Č.; methodology, A.Č.; software, A.Č. and I.B.; validation, A.Č. and I.B.; formal analysis, A.Č. and I.B.; investigation, A.Č. and I.B.; resources, A.Č. and I.B.; data curation, A.Č. and I.B.; writing—original draft preparation, I.B.; writing—review and editing, A.Č. and I.B.; visualization, A.Č. and I.B.; supervision, A.Č.; project administration, A.Č.; funding acquisition, A.Č. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are not publicly available due to privacy concerns, as the data contain sensitive information on an individual level.

**Acknowledgments:** This research was conducted as part of the Lithuanian Population and Housing Census 2021 and was supported by Statistics Lithuania (the State Data Agency), where both authors of the paper are employed.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MAR | missing at random |
| IPW | inverse probability weighting |
| DR | doubly robust |

## References

1. Axelson, M.; Holmberg, A.; Jansson, I.; Westling, S. A register-based census: The Swedish experience. In *Administrative Records for Survey Methodology*; Chun, A.Y., Larsen, M.D., Durrant, G., Reiter, J.P., Eds.; Wiley: Hoboken, NJ, USA, 2021; pp. 179–204.
2. Bernardini, A.; Brown, J.; Chipperfield, J.; Bycroft, C.; Chieppa, A.; Cibella, N.; Dunnet, G.; Hawkes, M.; Hleihel, A.; Law, E.; Ward, D.; Zhang, L.-C. Evolution of the person census and the estimation of population counts in New Zealand, United Kingdom, Italy and Israel. *Stat. J. IAOS* **2022**, *38*, 1221–1237. [CrossRef]
3. Bycroft, C. Census transformation in New Zealand: Using administrative data without a population register. *Stat. J. IAOS* **2015**, *31*, 401–411. [CrossRef]
4. Mule, V.T., Jr.; Keller, A. Administrative records applications for the 2020 census. In *Administrative Records for Survey Methodology*; Chun, A.Y., Larsen, M.D., Durrant, G., Reiter, J.P., Eds.; Wiley: Hoboken, NJ, USA, 2021; pp. 205–229.
5. Tille, Y. *Sampling and Estimation from Finite Populations*; Wiley Series in Survey Methodology; Wiley: Hoboken, NJ, USA, 2020.
6. Argüeso, A.; Vega, J.L. A population census based on registers and a "10% survey" methodological challenges and conclusions. *Stat. J. IAOS* **2014**, *30*, 35–39.
7. Beaumont, J.F. Are probability surveys bound to disappear for the production of official statistics? *Surv. Methodol.* **2020**, *46*, 71–96 .
8. Kim, J.-K. A gentle introduction to data integration in survey sampling. *Surv. Stat.* **2022**, *85*, 19–29.
9. Rao, J.N.K. On making valid inferences by integrating data from surveys and other sources. *Sankhya B* **2021**, *83*, 242–272. [CrossRef]
10. Wu, C. Statistical inference with non-probability survey samples. *Surv. Methodol.* **2022**, *48*, 283–311.
11. Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **2018**, *12*, 685–726. [CrossRef]
12. Kim, J.-K.; Tam, S.-M. Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.* **2021**, *89*, 382–401. [CrossRef]
13. Tam, S.-M.; Kim, J.-K. Big data ethics and selection-bias: An official statistician's perspective. *Stat. J. IAOS* **2018**, *34*, 577–588. [CrossRef]
14. Chen, Y.; Li, P.; Wu, C. Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.* **2020**, *115*, 2011–2021. [CrossRef]
15. Castro-Martín, L.; Rueda, M.d.M.; Ferri-García, R. Estimating general parameters from non-probability surveys using propensity score adjustment. *Mathematics* **2020**, *8*, 2096–2109. [CrossRef]

16. Wu, C.; Sitter, R.R. A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.* **2001**, *96*, 185–193. [CrossRef]
17. Särndal, C.-E.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*; Springer Series in Statistics; Springer: New York, NY, USA, 1992.
18. Deville, J.C.; Särndal, C.-E. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382. [CrossRef]
19. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
20. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: New York, NY, USA, 1989.
21. Kim, J.K.; Park, S.; Chen, Y.; Wu, C. Combining non-probability and probability survey samples through mass imputation. *J. R. Stat. Soc. Ser. A* **2021**, *184*, 941–963. [CrossRef]
22. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Softw.* **2016**, *74*, 1–16. [CrossRef]
23. Dick, P. Modelling net undercoverage in the 1991 Canadian census. *Surv. Methodol.* **1995**, *21*, 45–54.
24. Yang, S.; Kim, J.-K. Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.* **2020**, *3*, 625–650. [CrossRef]