Non-probability sample integration in the survey of Lithuanian census

Ieva Burakauskaitė[1] and Andrius Čiginas[1,2]

[1]Statistics Lithuania
[2]Vilnius University

IASS Webinar | 31 August, 2022

Objects of interest

- **The Statistical survey on population by ethnicity, native language and religion 2021** aimed to evaluate population proportions of:
  - *religion professed* (16 categories),
  - *mother tongue* (12 categories),
  - *knowledge of other languages* (16 languages),
  - *ethnicity* (mass imputation was used).
- Let us further consider **binary variables** where y denotes one of the above mentioned categories of a corresponding variable with the fixed values $y_1, \ldots, y_N$ in the finite census population $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ of $N = 2810761$ individuals.
- We aim to estimate the **population proportion**

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

for each binary variable y.

The use of administrative data in the Census 2021

- Variables of interest were completely observed in previous **population and housing censuses**.

  Based on the data of the last census carried out in 2011:
  - Population of Lithuania comprised people of *154 ethnicities*;
  - *One in three* residents indicated that they spoke *two foreign languages*;
  - The residents belonged to *59 different religious communities*.

- The main part of **Census 2021** was based on **administrative data**.

- Additional variables were collected through **the Statistical survey on population by ethnicity, native language and religion 2021**.

Combination of voluntary and probability samples

Ⓐ voluntary sample	An **online survey** was carried out from 15 January to 28 February, 2021. Approximately **2%** of the census population filled in the given questionnaire.
Ⓑ not in the sampling frame	After the end of the online survey, a sampling frame for probability sampling was constructed. It **excluded addresses if**: at least one individual from the address participated in the online survey, it was an institution, etc.
Ⓒ probability sample	Around **40 thousand addresses** were sampled from the Population Register. Approximately **6%** of the census population was interviewed.

Imputation of missing values: *historical*, *deductive* and *k-nearest neighbor* methods

- Missing values in the survey sample were filled in using **historical information from censuses 2011 and 2001** consecutively, as variables of interest are fully known for the populations of previous censuses.
- **Additional sociodemographic characteristics** of previous and current censuses (such as age, gender, marital status, household structure, country of birth, citizenship, education, employment status, etc.) were used for **the deductive imputation**.
- The remaining missing values in the survey sample were filled in using **the *k*-nearest neighbor method**.

Sampling design

- The sampling frame was divided into $H = 113$ strata:

  municipality $\times$ area of residence (i.e., urban or rural).

- The sample $s \subset \mathcal{U}$ of size $n < N$, $n = 436404$, was drawn according to the sampling design $p(\cdot)$ with **inclusion into the sample probabilities** $\pi_k = \mathrm{P}_p\{k \in s\} > 0$, $k \in \mathcal{U}$:
  - Inclusion into the sample probability for unit k in stratum h equals to

    $$\pi_k \approx \frac{m_k n'_h}{N'_h},$$

    where N'_h denotes the size of the hth stratum, n'_h is the number of addresses selected, and m_k is the number of individuals in the corresponding address;
  - $\pi_k = 1$ for voluntary sample respondents and addresses not in the sampling frame.

- **The primary sampling weights** then equal to $d_k = 1/\pi_k$.

Calibration (generalized regression) estimator

- **The generalized regression estimator** with calibrated weights w_k is used to evaluate the proportion μ:

$$\hat{\mu}^{\mathrm{GR}} = \frac{1}{\widehat{N}} \sum_{k \in s} w_k y_k, \qquad \text{where} \quad \widehat{N} = \sum_{k \in s} w_k.$$

- Weights are calibrated according to Deville and Särndal (1992):

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

subject to the calibration equations

$$\sum_{k \in s} w_k x_k^{(1)} = \sum_{k \in \mathcal{U}} x_k^{(1)}, \ \ldots \ , \sum_{k \in s} w_k x_k^{(P)} = \sum_{k \in \mathcal{U}} x_k^{(P)}$$

for P auxiliary variables $x^{(1)}, \ldots, x^{(P)}$ (binary variables on **age groups**, **gender** and **religions professed in 2011 intersected with counties**) with known values in the census population $\mathcal{U}$.

Estimation of $\hat{\mu}^{\mathrm{GR}}$ variance

Variance of $\hat{\mu}^{\mathrm{GR}}$ is estimated according to Deville and Särndal (1992):

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

where

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

with $\mathbf{x}_k = \left(x_k^{(1)}, \ldots, x_k^{(P)} \right)'$ and $\pi_{kl} = \mathrm{P}_p\{k, l \in s\} > 0$.

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	$\geqslant 30, < 50$	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

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

Propensity scores

- Consider the non-probability sample s_A consisting of n_A units from the census population $\mathcal{U}$. Let $R_k = \mathbb{I}(k \in s_A)$ be the indicator variable for a unit $k \in \mathcal{U}$ being selected to the sample s_A.
- **Propensity scores** (Rosenbaum and Rubin, 1983) are given by

$$\pi_k^A = \mathrm{E}_q(R_k \,|\, \mathbf{x}_k, y_k) = \mathrm{P}_q(R_k = 1 \,|\, \mathbf{x}_k, y_k), \quad k \in \mathcal{U},$$

  where the subscript q refers to the model for the selection mechanism for the sample s_A – **the propensity score model**.

Propensity score model

- **Model assumptions**:
  1. The selection indicator R_k and the response variable y_k are independent given the covariates $\mathbf{x}_k$.
  2. All units have a nonzero propensity score: $\pi_k^A > 0$ for all $k \in \mathcal{U}$.
  3. The selection indicators R_k and R_l are independent given the respective covariates $\mathbf{x}_k$ and $\mathbf{x}_l$ for $k \neq l$.

- Propensity scores $\pi_k^A = \mathrm{P}_q(R_k = 1 \,|\, \mathbf{x}_k)$ can be modelled **parametrically** as

$$\pi_k^A = \pi(\mathbf{x}_k, \boldsymbol{\theta}_0) = \frac{\exp(\mathbf{x}_k' \boldsymbol{\theta}_0)}{1 + \exp(\mathbf{x}_k' \boldsymbol{\theta}_0)},$$

  where $\boldsymbol{\theta}_0$ is the true value of the unknown model parameters.

Estimation of propensity scores

- **The maximum likelihood estimator** for $\hat{\pi}_k^A$ is computed as $\hat{\pi}_k^A = \pi(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function

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

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

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

using the Newton-Raphson iterative procedure.

Inverse probability weighted estimator

The estimated propensity scores $\hat{\pi}_k^A = \pi(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$, $k \in s_A$, are used to compute **the inverse probability weighted estimator** for the proportion μ (Chen et al., 2020):

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

where $\widehat{N}^A = \sum_{k \in s_A} 1/\hat{\pi}_k^A$.

Estimation of asymptotic variance $\widehat{V}^{\mathrm{IPW}}$

Under certain regularity conditions discussed in Chen et al. (2020) and assuming the logistic regression model for the propensity scores, we have $\hat{\mu}^{\mathrm{IPW}} - \mu = O_p(n_A^{-1/2})$, and an asymptotic **variance of** $\hat{\mu}^{\mathrm{IPW}}$ can be estimated using

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

where

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

Generalized difference estimator

- Assume the relationship between y and $\mathbf{x}$ can be described by a superpopulation (outcome regression) model ξ:

$$\mathrm{E}_\xi(y_k|\mathbf{x}_k) = m(\mathbf{x}_k, \boldsymbol{\beta}), \quad \mathrm{V}_\xi(y_k|\mathbf{x}_k) = v_k^2\sigma^2, \quad k \in \mathcal{U},$$

  where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{P'})'$ and σ^2 are unknown parameters. In our case, $m(\mathbf{x}_k, \boldsymbol{\beta})$ takes the parametric form $m(\mathbf{x}_k, \boldsymbol{\beta}) = \mathbf{x}_k'\boldsymbol{\beta}$ where $\mathbf{x}_k$ are the same as in the propensity score model and model ξ is fitted using the units from the sample s, and $v_k = 1$.

- **The generalized difference estimator** (Wu and Sitter, 2001) is alternatively used to evaluate the proportion μ:

$$\hat{\mu}^{\mathrm{GD}} = \frac{1}{N}\left(\sum_{k\in s} d_k(y_k - \widehat{m}_k) + \sum_{k\in\mathcal{U}} \widehat{m}_k\right),$$

  where $\widehat{m}_k = m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}) = \mathbf{x}_k'\hat{\boldsymbol{\beta}}$.

Estimation of $\hat{\mu}^{\mathrm{GD}}$ variance

Design-variance of $\hat{\mu}^{\mathrm{GD}}$ (Wu and Sitter, 2001) is estimated as

$$\hat{\psi}^{\mathrm{GD}} = \frac{1}{\widehat{N}^2} \sum_{l \in s} \sum_{\substack{k \in s \\ k<l}} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \left(\frac{u_k}{\pi_k} - \frac{u_l}{\pi_l} \right)^2,$$

where $\pi_{kl} = \mathrm{P}_p\{k, l \in s\} > 0$ and $u_k = y_k - m(\mathbf{x}_k, \hat{\boldsymbol{\beta}})$ with $\mathbf{x}_k = \left(x_k^{(1)}, \ldots, x_k^{(P')} \right)'$.

Doubly robust estimator

- We consider the same parametric model $\mathrm{E}_\xi(y_k|\mathbf{x}_k) = m(\mathbf{x}_k, \boldsymbol{\beta})$ where subscript ξ indicates the model for the outcome regression that is fitted with only n_A units from the non-probability sample s_A.
- **The doubly robust estimator** for the proportion μ is computed using the estimated propensity scores $\hat{\pi}_k^A = \pi(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$, $k \in s_A$ (Chen et al., 2020):

$$\hat{\mu}^{\mathrm{DR}} = \frac{1}{\widehat{N}^A} \sum_{k \in s_A} \frac{y_k - \widehat{m}_k}{\hat{\pi}_k^A} + \frac{1}{N} \sum_{k=1}^{N} \widehat{m}_k,$$

where $\widehat{N}^A = \sum_{k \in s_A} 1/\hat{\pi}_k^A$ and $\widehat{m}_k = m(\mathbf{x}_k, \hat{\boldsymbol{\beta}}) = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$.

Estimation of asymptotic variance $\widehat{V}^{\mathrm{DR}}$

Under certain regularity conditions discussed in Chen et al. (2020) and assuming the correctly specified logistic regression model for the propensity scores, the estimator of the asymptotic **variance of** $\hat{\mu}^{\mathrm{DR}}$ can be derived as

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

where

$$\hat{h} = \frac{1}{(\widehat{N}^A)^2} \sum_{k \in s_A} \frac{y_k - \widehat{m}_k}{\hat{\pi}_k^A},$$

$$\widehat{\widetilde{\mathbf{b}}}' = \left\{ \sum_{k \in s_A} \left(\frac{1}{\hat{\pi}_k^A} - 1 \right) \left(y_k - \widehat{m}_k - \hat{h} \right) \mathbf{x}_k' \right\} \left\{ \sum_{k \in \mathcal{U}} \hat{\pi}_k^A \left(1 - \hat{\pi}_k^A\right) \mathbf{x}_k \mathbf{x}_k' \right\}^{-1}.$$

Composite estimator (I)

- **Estimator of the population proportion** μ is constructed as

$$\hat{\mu}^{\mathrm{C1}} = \hat{\lambda}_1 \hat{\mu}^{\mathrm{GR}} + (1 - \hat{\lambda}_1)\hat{\mu}^{\mathrm{IPW}}, \qquad \text{where} \quad \hat{\lambda}_1 = \frac{\widehat{V}^{\mathrm{IPW}}}{\hat{\psi}^{\mathrm{GRs}} + \widehat{V}^{\mathrm{IPW}}},$$

  and $\hat{\psi}^{\mathrm{GRs}}$ is a smoothed version of the variance estimator $\hat{\psi}^{\mathrm{GR}}$.

  For the smoothing of variance, similarly as in Dick (1995), we assume $\mathrm{var}_p(\hat{\mu}^{\mathrm{GR}}) \approx K\widetilde{N}^{\gamma}$, with $\widetilde{N}$ as the sum of values of 2011 variable of interest in the 2021 census population. Parameters $K > 0$ and $\gamma \in \mathbb{R}$ are evaluated through a log-log regression with all categories of the variable of interest as auxiliary information.

- **Variance estimator for the composition** $\hat{\mu}^{\mathrm{C1}}$ is taken as

$$\widehat{V}^{\mathrm{C1}} = \hat{\lambda}_1 \hat{\psi}^{\mathrm{GRs}}.$$

- Estimates $\hat{\mu}^{\mathrm{C1}}$ are **benchmarked** using $\widehat{V}^{\mathrm{C1}}$ as weights.

Composite estimator (II)

- **Estimator of the population proportion** μ is constructed as

$$\hat{\mu}^{\mathrm{C2}} = \hat{\lambda}_2 \hat{\mu}^{\mathrm{GD}} + (1 - \hat{\lambda}_2)\hat{\mu}^{\mathrm{DR}}, \qquad \text{where} \quad \hat{\lambda}_2 = \frac{\widehat{V}^{\mathrm{DR}}}{\hat{\psi}^{\mathrm{GDs}} + \widehat{V}^{\mathrm{DR}}},$$

  and $\hat{\psi}^{\mathrm{GDs}}$ is a smoothed version of the variance estimator $\hat{\psi}^{\mathrm{GD}}$.

- **Variance estimator for the composition** $\hat{\mu}^{\mathrm{C2}}$ is taken as

$$\widehat{V}^{\mathrm{C2}} = \hat{\lambda}_2 \hat{\psi}^{\mathrm{GDs}}.$$

- Estimates $\hat{\mu}^{\mathrm{C2}}$ are **benchmarked** using $\widehat{V}^{\mathrm{C2}}$ as weights.

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

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

	$\hat{\mu}^{\mathrm{GR}}$	$\hat{\mu}^{\mathrm{C1}}$	$\hat{\mu}^{\mathrm{GD}}$	$\hat{\mu}^{\mathrm{C2}}$
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 4: *Religion* proportion estimates in 2021 census population.

	$\hat{\mu}^{\mathrm{GR}}$	$\hat{\mu}^{\mathrm{C1}}$	$\hat{\mu}^{\mathrm{GD}}$	$\hat{\mu}^{\mathrm{C2}}$
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: Comparison of 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

Summary

We considered a few ways of **non-probability sample integration**:

- **Natural post-stratified estimation** with elements of the non-probability sample, slightly modified by calibration. The post-stratification would have a more significant effect if the voluntary sample comprised a larger part of the census population.
- **Combination of the post-stratified calibration estimator with the inverse probability weighted estimator** correcting the selection bias arising from the non-probability sample. It exploits the latter sample better as the estimation of small proportions of interest is improved according to the assessment of subject matter experts.
- Alternatively, **combination of the generalized difference estimator with the doubly robust estimator** was considered as it lets to better exploit the unit level information.

Final remarks

- **Combination of the post-stratified calibration estimator with the inverse probability weighted estimator** improves the estimation accuracy of proportions of smaller religious communities as the variances of the first composite estimator are estimated smaller than the variances of the generalized regression estimator in up to 19 percent.
- **Combination of the generalized difference estimator with the doubly robust estimator**, while correcting the estimation accuracy of proportions of bigger religious communities, does not give satisfactory results for smaller religions.

Literature

- Chen, Y., Li, P. and Wu, C. (2020), Doubly Robust Inference With Nonprobability Survey Samples, *Journal of the American Statistical Association*, 115(532), pp. 1–25.
- Deville, J.-C. and Särndal, C.-E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87(418), pp. 376–382.
- Dick, P. (1995), Modelling net undercoverage in the 1991 Canadian census, *Survey Methodology*, 21(1), pp. 45–54.
- Rosenbaum, P. R. and Rubin, D. B. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70(1), pp. 41–55.
- Wu, C. and Sitter, R. R. (2001), A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, 96(453), pp. 185–193.