Non-probability sample integration in the survey of Lithuanian census

Ieva Burakauskaitė\textsuperscript{1} and Andrius Čiginas\textsuperscript{1,2}

\textsuperscript{1}Statistics Lithuania
\textsuperscript{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 \( U = \{1, \ldots, k, \ldots, N\} \) of \( N = 2810761 \) individuals.

- We aim to estimate the population proportion

\[
\mu = \frac{1}{N} \sum_{k \in U} y_k
\]

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

A 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.

B 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.

C 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**.
The sampling frame was divided into $H = 113$ strata: municipality × area of residence (i.e., urban or rural).

The sample $s \subset U$ of size $n < N$, $n = 436404$, was drawn according to the sampling design $p(\cdot)$ with inclusion into the sample probabilities $\pi_k = P_p\{k \in s\} > 0$, $k \in U$:

- 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 $h$th 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 $d_k = 1/\pi_k$. 

6 / 28
Calibration (generalized regression) estimator

The generalized regression estimator with calibrated weights $w_k$ is used to evaluate the proportion $\mu$:

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

where $\hat{N} = \sum_{k \in s} w_k$.

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

$$\sum_{k \in s} \left( w_k - d_k \right)^2 \rightarrow \min \quad \text{subject to the calibration equations}$$

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

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 $U$. 
Estimation of $\hat{\mu}^{\text{GR}}$ variance

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

$$
\hat{\psi}^{\text{GR}} = \frac{1}{\tilde{N}^2} \sum_{k \in s} \sum_{l \in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{(y_k - x_k' \hat{\mathbf{B}})(y_l - x_l' \hat{\mathbf{B}})}{\pi_k \pi_l},
$$

where

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

with $x_k = \left(x_k^{(1)}, \ldots, x_k^{(P)}\right)'$ and $\pi_{kl} = \mathbb{P}_p\{k, l \in s\} > 0$. 
Table 1: Comparison of proportions of some sociodemographic characteristics in the voluntary sample and the whole population.

<table>
<thead>
<tr>
<th></th>
<th>Voluntary sample</th>
<th>Population</th>
<th>Difference in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td>Pole</td>
<td>0.35</td>
<td>0.07</td>
</tr>
<tr>
<td>Education higher</td>
<td>0.48</td>
<td>0.20</td>
<td>134</td>
</tr>
<tr>
<td>County Vilnius</td>
<td>0.64</td>
<td>0.29</td>
<td>121</td>
</tr>
<tr>
<td>Employment employed</td>
<td>0.63</td>
<td>0.45</td>
<td>41</td>
</tr>
<tr>
<td>Age group ⩾ 30, &lt; 50</td>
<td>0.37</td>
<td>0.27</td>
<td>37</td>
</tr>
<tr>
<td>Marital status married</td>
<td>0.52</td>
<td>0.42</td>
<td>25</td>
</tr>
<tr>
<td>Gender male</td>
<td>0.41</td>
<td>0.46</td>
<td>-11</td>
</tr>
<tr>
<td>Ethnicity Lithuanian</td>
<td>0.56</td>
<td>0.85</td>
<td>-34</td>
</tr>
<tr>
<td>Education (lower) secondary</td>
<td>0.24</td>
<td>0.37</td>
<td>-35</td>
</tr>
<tr>
<td>Education primary</td>
<td>0.09</td>
<td>0.20</td>
<td>-55</td>
</tr>
</tbody>
</table>
Table 2: Comparison of religion proportions in the voluntary sample and the whole population.

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

Consider the non-probability sample $s_A$ consisting of $n_A$ units from the census population $\mathcal{U}$. Let $R_k = I(k \in s_A)$ 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

$$
\pi^A_k = \mathbb{E}_q(R_k \mid x_k, y_k) = \mathbb{P}_q(R_k = 1 \mid x_k, y_k), \quad k \in \mathcal{U},
$$

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 $x_k$.
2. All units have a nonzero propensity score: $\pi_k^A > 0$ for all $k \in U$.
3. The selection indicators $R_k$ and $R_l$ are independent given the respective covariates $x_k$ and $x_l$ for $k \neq l$.

▶ Propensity scores $\pi_k^A = P(\pi_k = 1 | x_k)$ can be modelled parametrically as

$$
\pi_k^A = \pi(x_k, \theta_0) = \frac{\exp(x_k'\theta_0)}{1 + \exp(x_k'\theta_0)},
$$

where $\theta_0$ 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 $\hat{\pi}_k^A = \pi(x_k, \hat{\theta})$, where $\hat{\theta}$ maximizes the log-likelihood function

\[
l(\theta) = \sum_{k \in s_A} \log \left\{ \frac{\pi(x_k, \theta)}{1 - \pi(x_k, \theta)} \right\} + \sum_{k \in U} \log \{1 - \pi(x_k, \theta)\}
= \sum_{k \in s_A} x_k' \theta - \sum_{k \in U} \log \{1 + \exp(x_k' \theta)\}.
\]

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

\[
U(\theta) = \frac{\partial}{\partial \theta} l(\theta) = \sum_{k \in U} \{R_k - \pi(x_k, \theta)\} x_k = 0
\]

using the Newton-Raphson iterative procedure.
Inverse probability weighted estimator

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

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

where $\hat{N}_A = \sum_{k \in s_A} 1/\hat{\pi}_k^A$. 
Estimation of asymptotic variance $\hat{V}_{IPW}$

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

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

where

$$\hat{b}' = \left\{ \sum_{k \in s_A} \left( \frac{1}{\hat{\pi}_k^A} - 1 \right) (y_k - \hat{\mu}_{IPW}) x_k' \right\} \left\{ \sum_{k \in U} \hat{\pi}_k^A (1 - \hat{\pi}_k^A) x_k x_k' \right\}^{-1}.$$
Generalized difference estimator

Assume the relationship between $y$ and $x$ can be described by a superpopulation (outcome regression) model $\xi$:

$$E_{\xi}(y_k | x_k) = m(x_k, \beta), \quad V_{\xi}(y_k | x_k) = v_k^2 \sigma^2, \quad k \in U,$$

where $\beta = (\beta_0, \ldots, \beta_{P'})'$ and $\sigma^2$ are unknown parameters. In our case, $m(x_k, \beta)$ takes the parametric form $m(x_k, \beta) = x_k' \hat{\beta}$ where $x_k$ are the same as in the propensity score model and model $\xi$ 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 $\mu$:

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

where $\hat{m}_k = m(x_k, \hat{\beta}) = x_k' \hat{\beta}$. 
Design-variance of $\hat{\mu}^{GD}$ (Wu and Sitter, 2001) is estimated as

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

where $\pi_{kl} = P_p\{k, l \in s\} > 0$ and $u_k = y_k - m(x_k, \hat{\beta})$ with $x_k = (x_k^{(1)}, \ldots, x_k^{(P')})'$. 

Estimation of $\hat{\mu}^{GD}$ variance
We consider the same parametric model $\mathbb{E}_{\xi}(y_k|x_k) = m(x_k, \beta)$ where subscript $\xi$ 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 $\mu$ is computed using the estimated propensity scores $\hat{\pi}_k^A = \pi(x_k, \hat{\theta})$, $k \in s_A$ (Chen et al., 2020):

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

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

Doubly robust estimator
Estimation of asymptotic variance $\hat{V}_{\text{DR}}$

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 $\hat{\mu}_{\text{DR}}$ can be derived as

$$
\hat{V}_{\text{DR}} = \frac{1}{(\hat{N}A)^2} \sum_{k \in s_A} \left( 1 - \hat{\pi}_k^A \right) \left( \frac{y_k - \hat{m}_k - \hat{h}}{\hat{\pi}_k^A} - \hat{\beta}' x_k \right)^2,
$$

where

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

$$
\hat{\beta}' = \left\{ \sum_{k \in s_A} \left( \frac{1}{\hat{\pi}_k^A} - 1 \right) (y_k - \hat{m}_k - \hat{h}) x_k' \right\} \left\{ \sum_{k \in U} \hat{\pi}_k^A \left( 1 - \hat{\pi}_k^A \right) x_k x_k' \right\}^{-1}.
$$
Composite estimator (I)

- **Estimator of the population proportion** $\mu$ is constructed as
  \[ \hat{\mu}^{C1} = \hat{\lambda}_1 \hat{\mu}^{GR} + (1 - \hat{\lambda}_1) \hat{\mu}^{IPW}, \]
  where \[ \hat{\lambda}_1 = \frac{\hat{V}^{IPW}}{\hat{\psi}^{GRs} + \hat{V}^{IPW}}, \]
  and $\hat{\psi}^{GRs}$ is a smoothed version of the variance estimator $\hat{\psi}^{GR}$.

For the smoothing of variance, similarly as in Dick (1995), we assume $\text{var}_p(\hat{\mu}^{GR}) \approx K \tilde{N} \gamma$, with $\tilde{N}$ 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** $\hat{\mu}^{C1}$ is taken as
  \[ \hat{V}^{C1} = \hat{\lambda}_1 \hat{\psi}^{GRs}. \]

- Estimates $\hat{\mu}^{C1}$ are **benchmarked** using $\hat{V}^{C1}$ as weights.
Composite estimator (II)

- Estimator of the population proportion $\mu$ is constructed as

$$\hat{\mu}^{C2} = \hat{\lambda}_2 \hat{\mu}^{GD} + (1 - \hat{\lambda}_2) \hat{\mu}^{DR},$$

where

$$\hat{\lambda}_2 = \frac{\hat{V}^{DR}}{\hat{\psi}^{GDs} + \hat{V}^{DR}},$$

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

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

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

- Estimates $\hat{\mu}^{C2}$ are benchmarked using $\hat{V}^{C2}$ as weights.
Table 3: Religion proportions in 2001, 2011 and 2021 census populations.

<table>
<thead>
<tr>
<th>Religion</th>
<th>$\mu^{(2001)}$</th>
<th>$\mu^{(2011)}$</th>
<th>$\hat{\mu}_{GR}$</th>
<th>$\hat{\mu}_{GD}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roman Catholics</td>
<td>0.78391</td>
<td>0.77233</td>
<td>0.73664</td>
<td>0.74101</td>
</tr>
<tr>
<td>Not indicated</td>
<td>0.05671</td>
<td>0.10112</td>
<td>0.15701</td>
<td>0.15025</td>
</tr>
<tr>
<td>None</td>
<td>0.09696</td>
<td>0.06146</td>
<td>0.05408</td>
<td>0.05477</td>
</tr>
<tr>
<td>Orthodox</td>
<td>0.04150</td>
<td>0.04113</td>
<td>0.03433</td>
<td>0.03482</td>
</tr>
<tr>
<td>Old Believers</td>
<td>0.00806</td>
<td>0.00767</td>
<td>0.00434</td>
<td>0.00419</td>
</tr>
<tr>
<td>Evangelical Lutherans</td>
<td>0.00565</td>
<td>0.00604</td>
<td>0.00389</td>
<td>0.00398</td>
</tr>
<tr>
<td>Other</td>
<td>0.00282</td>
<td>0.00493</td>
<td>0.00566</td>
<td>0.00625</td>
</tr>
<tr>
<td>Evangelical Reformed Believers</td>
<td>0.00208</td>
<td>0.00221</td>
<td>0.00122</td>
<td>0.00126</td>
</tr>
<tr>
<td>Pentecostalists</td>
<td>0.00037</td>
<td>0.00061</td>
<td>0.00117</td>
<td>0.00158</td>
</tr>
<tr>
<td>Sunni Muslims</td>
<td>0.00075</td>
<td>0.00089</td>
<td>0.00058</td>
<td>0.00064</td>
</tr>
<tr>
<td>Baptists and Free Churches</td>
<td>0.00034</td>
<td>0.00044</td>
<td>0.00017</td>
<td>0.00016</td>
</tr>
<tr>
<td>Judaists</td>
<td>0.00039</td>
<td>0.00040</td>
<td>0.00025</td>
<td>0.00029</td>
</tr>
<tr>
<td>Greek Catholics (Uniats)</td>
<td>0.00010</td>
<td>0.00023</td>
<td>0.00030</td>
<td>0.00038</td>
</tr>
<tr>
<td>Seventh Day Adventist Church</td>
<td>0.00016</td>
<td>0.00030</td>
<td>0.00014</td>
<td>0.00013</td>
</tr>
<tr>
<td>New Apostolic Church</td>
<td>0.00012</td>
<td>0.00014</td>
<td>0.00015</td>
<td>0.00020</td>
</tr>
<tr>
<td>Karaites</td>
<td>0.00008</td>
<td>0.00010</td>
<td>0.00008</td>
<td>0.00010</td>
</tr>
</tbody>
</table>
Table 4: *Religion* proportion estimates in 2021 census population.

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\mu}^{GR} )</th>
<th>( \hat{\mu}^{C1} )</th>
<th>( \hat{\mu}^{GD} )</th>
<th>( \hat{\mu}^{C2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roman Catholics</td>
<td>0.73664</td>
<td>0.73349</td>
<td>0.74101</td>
<td>0.73811</td>
</tr>
<tr>
<td>Not indicated</td>
<td>0.15701</td>
<td>0.15452</td>
<td>0.15025</td>
<td>0.14832</td>
</tr>
<tr>
<td>None</td>
<td>0.05408</td>
<td>0.05319</td>
<td>0.05477</td>
<td>0.05401</td>
</tr>
<tr>
<td>Orthodox</td>
<td>0.03433</td>
<td>0.03804</td>
<td>0.03482</td>
<td>0.03592</td>
</tr>
<tr>
<td>Old Believers</td>
<td>0.00434</td>
<td>0.00503</td>
<td>0.00419</td>
<td>0.00486</td>
</tr>
<tr>
<td>Evangelical Lutherans</td>
<td>0.00389</td>
<td>0.00460</td>
<td>0.00398</td>
<td>0.00475</td>
</tr>
<tr>
<td>Other</td>
<td>0.00566</td>
<td>0.00636</td>
<td>0.00625</td>
<td>0.00709</td>
</tr>
<tr>
<td>Evangelical Reformed Believers</td>
<td>0.00122</td>
<td>0.00151</td>
<td>0.00126</td>
<td>0.00175</td>
</tr>
<tr>
<td>Pentecostalists</td>
<td>0.00117</td>
<td>0.00126</td>
<td>0.00158</td>
<td>0.00191</td>
</tr>
<tr>
<td>Sunni Muslims</td>
<td>0.00058</td>
<td>0.00069</td>
<td>0.00064</td>
<td>0.00094</td>
</tr>
<tr>
<td>Baptists and Free Churches</td>
<td>0.00017</td>
<td>0.00024</td>
<td>0.00016</td>
<td>0.00037</td>
</tr>
<tr>
<td>Judaists</td>
<td>0.00025</td>
<td>0.00031</td>
<td>0.00029</td>
<td>0.00051</td>
</tr>
<tr>
<td>Greek Catholics (Uniats)</td>
<td>0.00030</td>
<td>0.00034</td>
<td>0.00038</td>
<td>0.00060</td>
</tr>
<tr>
<td>Seventh Day Adventist Church</td>
<td>0.00014</td>
<td>0.00017</td>
<td>0.00013</td>
<td>0.00031</td>
</tr>
<tr>
<td>New Apostolic Church</td>
<td>0.00015</td>
<td>0.00017</td>
<td>0.00020</td>
<td>0.00035</td>
</tr>
<tr>
<td>Karaites</td>
<td>0.00008</td>
<td>0.00009</td>
<td>0.00010</td>
<td>0.00022</td>
</tr>
</tbody>
</table>
Table 4: *Religion* proportion estimates in 2021 census population.

<table>
<thead>
<tr>
<th>Religion</th>
<th>$\hat{\mu}^{GR}$</th>
<th>$\hat{\mu}^{C1}$</th>
<th>$\hat{\mu}^{GD}$</th>
<th>$\hat{\mu}^{C2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roman Catholics</td>
<td>0.73664</td>
<td>0.73349</td>
<td>0.74101</td>
<td>0.73811</td>
</tr>
<tr>
<td>Not indicated</td>
<td>0.155701</td>
<td>0.15452</td>
<td>0.15025</td>
<td>0.14832</td>
</tr>
<tr>
<td>None</td>
<td>0.05408</td>
<td>0.05319</td>
<td>0.05477</td>
<td>0.05401</td>
</tr>
<tr>
<td>Orthodox</td>
<td>0.03433</td>
<td>0.03804</td>
<td>0.03482</td>
<td>0.03592</td>
</tr>
<tr>
<td>Old Believers</td>
<td>0.00434</td>
<td>0.00503</td>
<td>0.00419</td>
<td>0.00486</td>
</tr>
<tr>
<td>Evangelical Lutherans</td>
<td>0.00389</td>
<td>0.00460</td>
<td>0.00398</td>
<td>0.00475</td>
</tr>
<tr>
<td>Other</td>
<td>0.00566</td>
<td>0.00636</td>
<td>0.00625</td>
<td>0.00709</td>
</tr>
<tr>
<td>Evangelical Reformed Believers</td>
<td>0.00122</td>
<td>0.00151</td>
<td>0.00126</td>
<td>0.00175</td>
</tr>
<tr>
<td>Pentecostalists</td>
<td>0.00117</td>
<td>0.00126</td>
<td>0.00158</td>
<td>0.00191</td>
</tr>
<tr>
<td>Sunni Muslims</td>
<td>0.00058</td>
<td>0.00069</td>
<td>0.00064</td>
<td>0.00094</td>
</tr>
<tr>
<td>Baptists and Free Churches</td>
<td>0.00017</td>
<td>0.00024</td>
<td>0.00016</td>
<td>0.00037</td>
</tr>
<tr>
<td>Judaists</td>
<td>0.00025</td>
<td>0.00031</td>
<td>0.00029</td>
<td>0.00051</td>
</tr>
<tr>
<td>Greek Catholics (Uniats)</td>
<td>0.00030</td>
<td>0.00034</td>
<td>0.00038</td>
<td>0.00060</td>
</tr>
<tr>
<td>Seventh Day Adventist Church</td>
<td>0.00014</td>
<td>0.00017</td>
<td>0.00013</td>
<td>0.00031</td>
</tr>
<tr>
<td>New Apostolic Church</td>
<td>0.00015</td>
<td>0.00017</td>
<td>0.00020</td>
<td>0.00035</td>
</tr>
<tr>
<td>Karaites</td>
<td>0.00008</td>
<td>0.00009</td>
<td>0.00010</td>
<td>0.00022</td>
</tr>
</tbody>
</table>
Table 5: Comparison of relative difference (in %): (i) \( \frac{\hat{\psi}^{GRs} - \hat{V}^{C1}}{\hat{V}^{C1}} \), (ii) \( \frac{\hat{\psi}^{GDs} - \hat{V}^{C2}}{\hat{V}^{C2}} \), (iii) \( \frac{\hat{V}^{C1} - \hat{V}^{C2}}{\hat{V}^{C2}} \).

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


