Workshop of the Baltic-Nordic-Ukrainian Network on Survey Statistics 2018

Central Statistical Bureau of Latvia

LATVIJAS UNIVERSITATE ANNO 1919 UNIVERSITY OF LATVIA

Latvia University of Life Sciences and Technologies

isi

International Association of Survey Statisticians

Nordplus

Central Statistical Bureau of Latvia

WORKSHOP OF THE BALTIC-NORDIC-UKRAINIAN NETWORK ON SURVEY STATISTICS 2018

August 21–24, 2018
Jelgava, Latvia

Organised by

Central Statistical Bureau of Republic of Latvia,

Faculty of Business, Management and Economics
(*University of Latvia*)

Faculty of Economics and Social Development
(*Latvia University of Life Sciences and Technologies*)

Lecture Materials and Contributed Papers

Rīga, 2018

Workshop of the Baltic-Nordic-Ukrainian Network on Survey Statistics. August 21–24, 2018, Jelgava, Latvia. Central Statistical Bureau of Latvia, Rīga, 2018, 128 p.

Baltijas-Ziemeļvalstu-Ukrainas apsekojumu statistikas seminārs. 2018. gada 21.–24. augusts, Jelgava, Latvija. Latvijas Republikas Centrālā statistikas pārvalde, Rīga, 2018, 128 lpp.

Editors: Mārtiņš Liberts, Laura Jozefa
Cover: Maija Graudiņa

PDF: ISBN 978-9984-06-528-1
Print: ISBN 978-9984-06-527-4

Preface

Baltic-Nordic co-operation on survey statistics started in 1992, initiated by Professor Gunnar Kulldorff. This led to the establishment of a Baltic-Nordic network for co-operation on education and research in survey statistics in 1996. The network was expanded in 2008, since when it has been called the Baltic-Nordic-Ukrainian (BNU) Network on Survey Statistics.

The network has been organising annual events as summer schools, workshops or conferences since 1997. The August 2018 Workshop of the Baltic-Nordic-Ukrainian Network on Survey Statistics in Jelgava, Latvia is the 22nd event in the series. The main theme of the 2018 workshop is **Population census based on administrative data**. We are expecting 61 participants at the workshop, representing twelve countries.

Prof. Li-Chun Zhang (*University of Southampton, UK & Statistics Norway*) and Dr. Anders Holmberg (*Statistics Norway*) are the keynote speakers of the workshop. They will give a set of six lectures titled "*The past, present and future of population censuses: Methodology and quality aspects when data sources are being reused and combined to transform a census system*". Maciej Beręsewicz (*Poznań University of Economics and Business, Poland*), Natallia Bokun (*Belarus State Economic University*), Juris Breidaks (*Central Statistical Bureau of Latvia*), Danutė Krapavickaitė (*Vilnius Gediminas Technical University, Lithuania*), Manuela Lenk (*Statistics Austria*), Melike Oguz Alper (*Statistics Norway*), and Carl-Erik Särndal (*Statistics Sweden*) are invited speakers at the workshop.

Most of the workshop participants will present contributed papers. 27 contributed papers have been announced. A discussant is assigned to each contributed paper. For the first time at a BNU event, an award for the best student paper will be given. More information about the workshop is available on the workshop website (*http://home.lu.lv/~pm90015/workshop2018/*).

I would like to express my thanks to the Organising Committee members for their very active involvement in the organisation of the workshop. Many thanks to the Programme Committee led by Thomas Laitila for the very rich workshop programme. Special thanks are due to Ieva Ainâre, Dina Brīdaka, Janīna Dišereite, Maija Graudiņa, Ilga Puisāne, Gunta Purviņa, Signe Saliņa, Salvis Staģis, Kaspars Vasaraudzis, Vija Vizule, Aija Žīgure and other staff members of the Central Statistical Bureau of Latvia who have contributed to the organisation of the workshop. Many thanks to Andra Zvirbule for hosting the workshop at the Faculty of Economics and Social Development (Latvia University of Life Sciences and Technologies).

Finally, I wish to express my gratitude to the workshop sponsors – the International Statistical Institute (for supporting the participants from Ukraine and Belarus), the International Association of Survey Statisticians (for financial support), the Nordplus Higher Education program (for supporting students and teachers from the Baltic and Nordic countries), and the Central Statistical Bureau of Latvia (for organisational support).

I wish all participants a successful and inspiring workshop and an enjoyable stay in Jelgava.

Rīga, August 2018

Mārtiņš Liberts, chair of the Baltic-Nordic-Ukrainian Network on Survey Statistics

Contents

Preface 3

Contents 4

Lectors and Lecture Titles 6

Natallia Bokun. Tourism surveys in Belarus 7

Juris Breidaks. At-risk-of-poverty threshold variance estimations using Gaussian kernel and smoothing splines in R package vardpoor 9

Manuela Lenk. Quality assessment of register based statistics 18

Melike Oguz-Alper. Sampling and Estimation in Finite Networks 19

Carl-Erik Särndal. Comments on the development of Survey Statistics theory and practice in the last fifty years, some personal reflections 21

Natallia Bandarenka. Population Census in the Republic of Belarus: Experience and Perspectives 23

Andris Fisenko, Janis Lapins. Use of Register Data in Latvian Household Finance and Consumption Survey 28

Yuliia Halytsia. Calibration Weighting in Survey Sampling (Based on Sample Socio-Demographic Survey) 30

Julia Hellstrand. All-time low period fertility in Finland: tempo or quantum effect? 35

Miika Honkala. Estimation of response propensities using the previous survey 38

Tetiana Ianevych, Veronika Serhiienko. Effect of using Tobit and Heckit models in regression estimation for data with many zeros 42

Inguna Jurgelane-Kaldava. Correlation between Logistics and Higher Education in Latvia 46

Viktoria Kirpu. Handling missing data and errors in Estonian eHealth information system 50

Seppo Laaksonen. Multiple Imputation for Income 55

Mārtiņš Liberts. Population Estimation Beyond 2021 61

Vytautas Pankūnas, Julija Janeiko. Coherence studies in time series 66

Oona Pentala-Nikulainen. Case study: The effect of text message reminder on survey nonresponse 70

Iryna Rozora. Impact Factors Modeling of Households Deposit Dollarization in Ukraine 74

Natalia Rozora. Brexit: challenges to estimate UK population 79

Natallia Sakovich. Tourist Incomes and Expenditures Surveys in Belarus 80

Volodymyr Sarioglo. Sample Weights Calibration with Aim to Reduce the Estimation Bias Due to Under Coverage of the Well-Off Population 84

Elvijs Siliņš. Calibration of Register Based Census Data 88

Alina Sinisalo. Reporting tool for annual change studies by using survey data 89

Milda Slickute-Sestokiene. Preparation for the register-based census 92

Diana Sokurova. The Local Pivotal Method and its Application on StatVillage Data 93

Mykola Sydorov. Sample models in monitoring survey UniDOS 97

Kaja Sõstra. Using Mobile Positioning for Improving the Quality of Register Data 107

Maria Valaste. Handling Nonsampling Errors - Case Salo 111

Olga Vasylyk. Modelling of Survey Data 114

Peteris Vegis. Combining data from registers, surveys and 2011 Population and Housing Census to prepare database for 2021 register-based Population and Housing Census in Latvia 117

Markus Gintas Šova. Rethinking Sampling for UK Business Surveys 122

Workshop programme 126

List of participants 127

Lectors and Lecture Titles

Keynote Speakers

Prof. Li-Chun Zhang (University of Southampton, Statistics Norway)

Dr. Anders Holmberg (Statistics Norway)

The past, present and future of population censuses: Methodology and quality aspects when data sources are being reused and combined to transform a census system

Invited Speakers

Maciej Beręsewicz (Poznań University of Economics and Business, Poland)

An overview of methods for treating selectivity in big data sources

Natallia Bokun (Belarus State Economic University)

Tourism surveys in Belarus

Juris Breidaks (Central Statistical Bureau of Latvia)

At-risk-of-poverty threshold variance estimations using Gaussian kernel and smoothing splines in R package vardpoor

Danutė Krapavickaitė (Vilnius Gediminas Technical University, Lithuania)

Population size estimation

Manuela Lenk (Statistics Austria)

Quality assessment of register based statistics

Melike Oguz Alper (Statistics Norway)

Sampling and estimation in finite networks

Carl-Erik Särndal (Statistics Sweden)

Comments on the development of Survey Statistics theory and practice in the last fifty years; some personal reflections

Tourism Surveys in Belarus

Natalia Bokun
Belarusian State Economic University, e-mail: nataliabokun@rambler.ru

Abstract

Tourism is a growing and complex phenomenon which is becoming one of the world's largest economic activities. International tourist arrivals have increased from 25 million globally in 1950 to 674 million in 2000, and 1,235 million in 2016. International tourism receipts earned by destinations worldwide have surged from US $ 2 billion in 1950 to US 1,220 billion in 2016. Tourism represents 7 % of the world's export of goods and services, has grown faster than world trade for the last five years. In Belarus the number of foreign visitors has increased from 120 thousand in 2010 to 217 thousand in 2016; the value of tourism export has increased, too (more than twice).

However this belief frequently faces a careless, partial and discordant set of information. The multiplicity of stakeholders involved in the tourism system (international organizations, national, regional, local administrations) implies different needs in terms of typologies of information: from tourism demand to the economic role and impacts of tourism; from statistical data to quantitative analyses. The final result is an enormous and growing request for information which requires different methodologies. This is why increasing efforts to harmonize methodologies, develop tourism satellite account.

Nowadays the National Statistical Committee of the Republic of Belarus does preparatory work on development of tourism satellite account. In 2017 Methodological Recommendations for construction of Tourism satellite account were adopted. Since 2017-2018, the first tables of this account are calculated.

The main sources and instruments of development of Tourism satellite account include tourism industry enterprises censuses and the system of different surveys: establishment samples, households samples. The first results of their use indicated the appearance of significant problems: non-responses, enough high level of errors, sample and non-sample errors, discrepancies between data from these surveys, the need for localization of the sample.

This lecture has the next parts:

1) tourism in Belarus: main indicators and trends;

2) tourism satellite account and possible information sources;

3) tourism households surveys;

4) tourism establishment surveys;

5) accommodation surveys;

6) Border surveys.

The use of combination of univariate and multivariate samples, quasicausal samples, expert estimates, tertiary sources, increase of sample size of Border surveys, updating existing questionnaires will provide more reliable information over larger number of tourism demand and tourism supply indicators.

Keywords: satellite account, tourism, establishment survey, household survey, sample, questionnaire.

References

Bokun, N. (2013). Sample survey of households in Belarus: state and perspectives. *Statistics in transition*, Warsaw, 110-121.

Bokun, N. (2016). Micro-entities and small enterprises survey in Belarus. *In Proc. of the XI Intern. Conf. "Compute Data and Modeling"*, Minsk, Sept. 6-10, 2016. – P. 240-245.

International Recommendation for Tourism Statistics 2008. Complication Guide (2010).

European Implementation Manual on TSA (2014).

UN WTO. Tourism Highlights, 2017 Edition (2017).

Metodologicheskie polozheniya po postroeniyu vspomogatelnogo scheta turizma Respubliki Belarus (2017). Minsk, Belstat.

Metodika po raschetu vjezdnogo turisticheskogo potoka (2018). Minsk, Belstat.

Metodika po raschetu obschego objema vvoza i vyvoza fizicheskimi litsami, peresekauschimi Gosudarstvennuyu granitsu Respubliki Belarus tovarov, ne uchityvaemyh tamozhennoy statistikoy (2015). Minsk, Belstat.

At-risk-of-poverty threshold variance estimations using Gaussian kernel and smoothing splines in R package vardpoor

Juris Breidaks[1]

[1]Central Statistical Bureau of Latvia, e-mail: juris.breidaks@csb.gov.lv

Abstract

The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team (2018)) package vardpoor (Breidaks *et al.* (2018)) (a set of functions for statistical calculation in programme R). The package vardpoor was developed with the objective to modernise the sample error estimation in sample surveys. Before the package was developed, sampling errors were estimated using the chargeable programme SUDAAN (www.rti.org/sudaan). Use of SUDAAN had several shortcomings:

- Only obsolete SUDAAN version was available at CSB, which had to be updated;
- Updating of SUDAAN version would require financial resources;
- It is difficult to integrate SUDAAN into work with other data processing programmes (IBM SPSS Statistics or R);
- With the help of SUDAAN it was possible to linearize only non-linear statistics, as the ratio of two totals, but in the EU-SILC survey there were several other non-linear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Given the above shortcomings, it was decided to develop the vardpoor package, which would be designed as R extension. First of all, R is an open-source free statistical calculation environment; secondly, R is currently the most popular computing environment among statisticians; and thirdly R environment is very convenient and suitable for development of such solutions. It should also be mentioned that, upon developing vardpoor package as R extension, it was easily integrated in the statistical production processes.

The theoretical basis of vardpoor was borrowed from G. Osier article The Linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards? (Osier & Di Meglio (2012)), which was presented at the workshop devoted to the evaluation of the standard errors and other issues related to the EU-SILC survey in March 2012.

Keywords: BNU2018, vardpoor, risk of poverty threshold, smoothing splines

1 Introduction

The Central Statistical Bureau of Latvia (CSB) in 2012 developed R (R Core Team (2018)) package vardpoor (Breidaks *et al.* (2018)) (a set of functions for statistical calculation in

programme R). The package vardpoor was developed with the objective to modernise the sample error estimation in sample surveys. Before the package was developed, sampling errors were estimated using the chargeable programme SUDAAN (www.rti.org/sudaan). Use of SUDAAN had several shortcomings:

- Only obsolete SUDAAN version was available at CSB, which had to be updated;
- Updating of SUDAAN version would require financial resources;
- It is difficult to integrate SUDAAN into work with other data processing programmes (IBM SPSS Statistics or R);
- With the help of SUDAAN it was possible to linearize only non-linear statistics, as the ratio of two totals, but in the EU-SILC survey there were several other non-linear statistics, which had to be linearized separately;
- SUDAAN sampling error estimation did not include the effect of weight calibration.

Given the above shortcomings, it was decided to develop the vardpoor package, which would be designed as R extension. First of all, R is an open-source free statistical calculation environment; secondly, R is currently the most popular computing environment among statisticians; and thirdly R environment is very convenient and suitable for development of such solutions. It should also be mentioned that, upon developing vardpoor package as R extension, it was easily integrated in the statistical production processes.

The theoretical basis of vardpoor was borrowed from G. Osier article "The Linearisation approach implemented by Eurostat for the first wave of EU-SILC: what could be done from the second wave onwards?" (Osier & Di Meglio (2012)), which was presented at the workshop devoted to the evaluation of the standard errors and other issues related to the EU-SILC survey in March 2012.

2 Sampling error estimation mechanism

Sampling error estimation mechanism consists of a sequence of procedures:

1. Calculation of the domain-specific study variables, if the sampling error is to be estimated for population domains;
2. At-risk-of-poverty threshold linearization using Gaussian kernel (Osier (2009) and smoothing splines (Asmuss *et al.* (2016));
3. Calculation of regression residual if the weights are calibrated;
4. Variance estimation with the ultimate cluster method (Hansen *et al.* (1953));
5. Variance estimation for the simple random sampling design.

2.1 Calculation of the domain-specific study variables

Often separate estimates for subpopulations are needed. Subpopulations are called domains. The domains concerned are denoted as $(U_1, ..., U_d, ..., U_D)$ It is assumed that y total value in each domain must be estimated. The aim is to estimate $(Y_1, ..., Y_d, ..., Y_D)$, where

$$Y_d = \sum_{k \in U_d} y_k, d = 1, ..., D \tag{1}$$

The domain total can be expressed with a new variable y_{dk}, constructed from y specifically for domain U_d (Lundström & Särndal (2001)). The new variable is denoted with y_{dk} and its values for each element k are defined as

$$y_{dk} = \begin{cases} y_k, & \text{if } k \in U_d, \\ 0, & \text{if } k \notin U_d. \end{cases} \tag{2}$$

Then Y_d can be expressed as a total from the new variable y_{dk} for the whole population:

$$Y_d = \sum_{k \in U} y_{dk} \tag{3}$$

2.2 Linearization approach

The linearisation method (Särndal *et al.* (1992), Deville (1999), Osier (2009)) uses Taylor-like series approximation to reduce non-linear statistics to a linear form, justified by asymptotic properties of the estimator (Verma & Betti (2005)). The method based on influence functions (Deville (1999)) is general enough to handle all the complex non-linear indicators of poverty and inequality based on EU-SILC such as the at-risk-of-poverty threshold. The estimated variance of the estimator $\widehat{\theta}$ can be approximated by a linear function of the sample observations:

$$\widehat{Var}(\widehat{Y}) \cong \widehat{Var}\Big(\sum_{k \in s} w_k \cdot \widehat{u}_k\Big), \tag{4}$$

where the value of the estimated linearized variable $\widehat{u}_k$ is determined by calculating the following functional derivative:

$$\widehat{u}_k = \lim_{t \to 0} \frac{T\big(\widehat{M} + t\delta_k\big) - T\big(\widehat{M}\big)}{t}, \tag{5}$$

where the estimated population parameter $\widehat{\theta}$ is expressed T as a functional of the measure $\widehat{M}$, i.e.,

$$\widehat{\theta} = T\big(\widehat{M}\big), \tag{6}$$

and the measure $\widehat{M}$ allocates the sample weight w_k to each unit k in the sample s:

$$\widehat{M}\big(k\big) = \widehat{M}_k = w_k, k \in s, \tag{7}$$

δ_k is the Dirac measure at k: for each unit k in the sample, $\delta_k(i) = 1$ if and only if $k = i$. The functional derivative (18) is called the influence function.

2.3 Weighted quantile estimation in the domain

Quantiles are defined as $Q_{D,p}^{-1} = F_D^{-1}(p)$, where F_D is the income distribution function on the population in the domain D, i.e.,

$$F_{D,y}(x) = \frac{1}{N_D} \sum_{k \in U_D} 1_{[y_k \leq x]} \tag{8}$$

and $0 \leq x \leq 1$. The median is given by $p = 0.5$. For the following definitions, let n_D be the number of observations in the domain D of the sample, let $x_D := (x_1, ..., x_{n_D}^{'})$, denote the equalized disposable income with $x_1 \leq ... \leq x_{n_D}$, and let $w_D := (w_1, ..., w_{n_D}^{'})$

be the corresponding personal sample weights. Weighted quantiles for the estimation of the population values in the domain D according are then given (M. (2013)) by

$$\widehat{Q}_{D;p} = \widehat{Q}_{D;p}(x_D, w_D) := \begin{cases} \frac{1}{2}(x_j + x_{j+1}), & \text{if } \sum_{i=1}^{j} w_i = p \sum_{i=1}^{n_D} w_i, \\ x_{j+1}, & \text{if } \sum_{i=1}^{j} w_i < p \sum_{i=1}^{n_D} w_i < \sum_{i=1}^{j+1} w_i. \end{cases} \tag{9}$$

2.4 Calculation of the at-risk-of-poverty threshold in domain and its linearization

The at-risk-of-poverty threshold (ARPT) in the domain D is defined as 60% of the median income in the domain D:

$$ARPT_D = 0.6 \cdot F_D^{-1}(0.5) \tag{10}$$

$$ARPT_D = 0.6 \cdot \widehat{Q}_{D;p}^{-1}(0.5) \tag{11}$$

The linearized variable of the ARPT in the domain D is defined by Osier (Osier (2009)):

$$\widehat{u}_{D;k}^{ARPT} = I(ARPT_D)_k = 0.6 \cdot I(\widehat{Q}_{D;0.5})_k = \frac{-0.6}{f(\widehat{Q}_{D;0.5})} \cdot \frac{1_{[k \in D]}}{\widehat{N}_D} \big[1_{[y_k \leq \widehat{Q}_{D;0.5}]} - 0.5\big], \tag{12}$$

where y_i is i-th equalized disposable income, $\widehat{N}_D$ is estimated size of the population in the domain D.

$f(.)$ is estimator of the density function which in the next subsections will be descripted using smoothing splines estimation and Gausian kernel estimation.

2.4.1 Calculation of the density function using Gaussian kernel estimation

Deville (1999) and Osier (2009) suggest using Gaussian kernel estimation for the calculation of the density function. The density functions can be estimated on the basis of the Gaussian kernel function as follows (Preston (1995))

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i K\Big(\frac{x - y_i}{h_D}\Big) \tag{13}$$

where

$$K(o) = \frac{1}{h_D \sqrt{2\pi}} e^{-\frac{o^2}{2}} \tag{14}$$

is the Gaussian kernel. $\widehat{N}_D = \sum_{i \in D} w_i$ is the Horvitz and Thompson (Horvitz & Thompson (1952)) estimator of the population size in domain D; h_D is the bandwidth parameter in the domain D. For normally distributed population densities, the following bandwidth parameter was recommended by Silverman (Silverman (1986))

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{15}$$

$\widehat{\sigma}_D$ is the estimated standard deviation of the empirical income distribution:

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \Big(\sum_{i \in s_D} w_k y_k\Big)^2}. \tag{16}$$

2.4.2 Calculation of the density function using smoothing splines function estimation

The density functions can be estimated on the basis of the smoothing splines function as follows

$$f_D(x) = \frac{1}{\widehat{N}_D \widehat{h}_D} \sum_{i \in D} w_i s\Big(\frac{x - y_i}{h_{Di}}\Big) \tag{17}$$

where $s(x)$ is the smoothing spline, $\widehat{N}_D = \sum_{i \in D} w_i$ is the Horvitz and Thompson (Horvitz & Thompson (1952)) estimator of the population size in domain D; h_D is the bandwidth parameter in the domain D. For smoothing population densities, the following bandwidth parameter was recommended by Silverman (Silverman (1986))

$$\widehat{h}_D = \widehat{\sigma}_D \widehat{N}_D^{-0.2} \tag{18}$$

$\widehat{\sigma}_D$ is the estimated standard deviation of the empirical income distribution:

$$\widehat{\sigma}_D = \frac{1}{\widehat{N}_D} \sqrt{\widehat{N}_D \sum_{i \in s_D} w_k y_k^2 - \Big(\sum_{i \in s_D} w_k y_k \Big)^2}. \tag{19}$$

Smoothing spline s is solution for the following problem of histopolation in the Sobolev space $W_2^q[a, b]$.

$$\int_a^b (g^{(q)}(t))^2 \, dt \longrightarrow \min_{\substack{g \in W_2^q[a,b], \\ \int_{t_{i-1}}^{t_i} g(t)dt = f_i h_i, \;\; i = 1, \ldots, n.}}$$

A solution of the spline s is in the form

$$s(t) = \sum_{j=0}^{r-1} \varrho_j t^j + \frac{(-1)^{r+1}}{(2r)!} \sum_{i=1}^{n} \alpha_i ((t - t_i)_+^{2r} - (t - t_{i-1})_+^{2r}) \tag{20}$$

with the following conditions on the coefficients:

$$\sum_{i=1}^{n} \frac{\alpha_i}{j+1} (t_i^{p+1} - t_{i-1}^{p+1}) = 0, \quad p = 0, 1, \ldots, r-1. \tag{21}$$

2.5 Regression residual calculation

If the weights are calibrated, then calibration residual estimates $\widehat{e}_k$ are calculated (Lundström & Särndal (2001)) by formula

$$\widehat{e}_k = y_k - x_k^{'} \widehat{B}, \tag{22}$$

where

$$\widehat{B} = \Big(\sum_{k \in s} d_k q_k x_k x_k^{'} \Big)^{-1} \Big(\sum_{k \in s} d_k q_k x_k y_k \Big) \tag{23}$$

2.6 Variance estimation with the ultimate cluster method

If we assume that $n_h \geq 2$ for all h, that is, two or several primary sampling units (PSUs) are sampled from each stratum, then variance of $\widehat{\theta}$ can be estimated from the variation among the estimated PSU totals of y (Hansen *et al.* (1953); Osier & Di Meglio (2012); Di Meglio *et al.* (2013)):

$$\widehat{V}(\widehat{\theta}) = \sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{k=1}^{n_h}\left(y_{hk*}-\bar{y}_{h**}\right)^2 \tag{24}$$

where

- $y_{hk*} = \sum_{j=1}^{m_{hk}} w_{hkj}y_{hkj}$
- $y_{h**} = \dfrac{\sum_{k=1}^{n_h} y_{hk*}}{n_h}$
- f_h is a sampling fraction of PSUs for stratum h,
- h is the stratum number, with a total of H strata,
- k is the number of PSU within the sample of stratum h, with a total of n_h PSUs,
- j is the household number within PSU k of stratum h, with a total of m_{hi} households,
- w_{hkj} is the sampling weight for household j in PSU k of stratum h,
- y_{hkj} denotes the observed value of study variable y for household j in PSU k of stratum h.

2.7 The design effect estimation and effective sample size

The design effect of sampling is estimated by

$$\widehat{Deff}_{sam}(\widehat{\theta}) = \frac{\widehat{Var}_{CUR,HT}(\widehat{\theta})}{\widehat{Var}_{SRS,HT}(\widehat{\theta})} \tag{25}$$

where $\widehat{Var}_{SRS,HT}(\widehat{\theta})$ is the variance of HT estimator under SRS, $\widehat{Var}_{SRS,HT}(\widehat{\theta})$ is the variance of HT estimator under current sampling design.

The design effect of estimator is estimated by

$$\widehat{eff}_{est}(\widehat{\theta}) = \frac{\widehat{Var}_{CUR,CAL}(\widehat{\theta})}{\widehat{Var}_{CUR,HT}(\widehat{\theta})} \tag{26}$$

where $\widehat{Var}_{CUR,CAL}(\widehat{\theta})$ is the variance of calibrated estimator under current sampling design.

The overall design effect of sampling and estimator is estimated by

$$\widehat{Deff}(\widehat{\theta}) = \widehat{Deff}_{sam}(\widehat{\theta}) \cdot \widehat{eff}_{est}(\widehat{\theta}) \tag{27}$$

The effective sample size is estimated by

$$\widehat{n}_{eff}(\widehat{\theta}) = \frac{n}{\widehat{Deff}_{sam}(\widehat{\theta})}, \tag{28}$$

where n is the sample size or the number of respondents (in case of non-response).

3 R package vardpoor

3.1 Function varpoord description

Function varpoord is used to estimate sampling errors for indicators on social exclusion and poverty. Data is given at the person level, but information for the calibration is given at the household level. At the beginning of the function execution a range of tests is performed in order to test if there are any mistakes in data. Function varpoord consist argument type, if it is chosen $linarpt$, then calculate the at-risk-of-poverty threshold (ARPT) in the domain and linearized values in the domain D using Gaussian kernel (Osier (2009) and smoothing splines (Asmuss *et al.* (2016))

If calibration matrix X and g weights are used at household level, function calculates the residuals at the household level. Function varpoord outputs several results:

- point estimates for statistics,
- variance estimates,
- relative standard error,
- absolute margin of error,
- relative margin of error,
- lower and upper bound of the confidence interval,
- variance of HT estimator under current design,
- variance of calibrated estimator under SRS,
- the sample design effect, the estimated design effect of estimator,
- the overall design effect of sample design and estimator,
- the effective sample size.

3.2 varpoord function testing results

Function was tested on simulated Austria data of EU-SILC. In this function will test ARPT quality indicator using smoothing splines (Asmuss *et al.* (2016)), the function varpoord() is used:

```
smooth_cal <- varpoord(inc = "INC_ekv20",
                       w_final = "db090",
                       income_thres = "INC_ekv20",
                       wght_thres = "db090",
                       ID_household = "db030n",
                       H = "db050",
                       PSU = "db060",
                       sort = NULL,
                       dataset = dataset2,
                       type = c("linarpt"),
                       method = "smooth_splines",
                       r = 2,
                       ro = 0.01)
```

Table 1: ARPT quality in 2012

method	estim	se	cv
Gaussian kernal	1876.67	50.59	2.69
Smoothing spline r=2 $\rho = 0.01$	1876.67	70.18	3.74

In this function will test ARPT quality indicator using Gaussian kernel (Osier (2009), the function varpoord() is used:

```
gausian_cal <- varpoord(inc = "INC_ekv20",
                        w_final = "db090",
                        income_thres = "INC_ekv20",
                        wght_thres = "db090",
                        ID_household = "db030n",
                        H = "db050",
                        PSU = "db060",
                        sort = NULL,
                        dataset = dataset2,
                        type = c("linarpt"),
                        method = "Gaussian")
```

In table was shown has calculated standard errors, coefficient of the variance.

References

Asmuss, S., Breidaks, J. & Budkina, N. (2016). On approximation of density function by shape preserving smoothing histospline. *Proceedings of the 15th Conference on Applied Mathematics (APLIMAT 2016)* , 30 – 43.

Breidaks, J., Liberts, M. & Ivanova, S. (2018). vardpoor: Variance estimation for sample surveys by the ultimate cluster , 1 – 27.

Deville, J. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25(2)**, 193 – 203.

Di Meglio, E., Osier, G., Goedemé, T., Berger, Y. G. & Di Falco, E. (2013). Standard error estimation in EU-SILC – first results of the Net-SILC2 project .

Hansen, M., H., W. N., Hurwitz & Madow, W. G. (1953). *Sample survey methods and theory*, vol. Volume I Methods and applications. Wiley.

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite verse , 663 – 685.

Lundstöm, S. & Särndal, C. E. (2001). *Estimation in the presence of nonresponse and frame imperfections*.

M., A. A. . T. (2013). Estimation of social exclusion indicators from complex surveys: The r package laeken. **54(15)**.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods* **3 (3)**, 167 – 195.

Osier, G. & Di Meglio, E. (2012). The linearisation approach implemented by eurostat for the first wave of EU-SILC: what could be done from second wave onwards? Tech. rep., Institut National de la Statistique et des Etudes Economiques (STATEC Luxembourg), Luxembourg.

Preston, I. (1995). Sampling distributions of relative poverty statistics. *Applied Statistics* **44**, 91 – 99.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London:Chapman and Hall.

Verma, V. & Betti, G. (2005). *Sampling errors and design effects for poverty measures and other complex statistics*. Working Paper 53, Siena: Dipartimento di Metodi Quantitativi, Universita degli Studi.

Quality assessment of register based statistics

Manuela Lenk[1]

[1]Statistics Austria, e-mail: manuela.lenk@statistik.gv.at

Abstract

In the scope of the transformation to a register-based Census 2011 in Austria, a quality framework for statistical data based on administrative sources has been developed. Now, this quality framework is used annually to evaluate the quality of the register-based labour market statistics (RBLMS). These quality indicators offer a wide range of possibilities to analyse the attributes on their own but also in combinations. Moreover, the calculated quality indicators i.e. on register level can also be used for other projects and deliver essential information of the quality. The presentation reaches from the fundaments of the Austrian Census to the basic concept of the quality framework and provides detailed results from the application of the framework to the Austrian register-based labour market statistics. Starting with a look on key topics like the Principle of Redundancy and Analysis of Residence, the talk gives an overview of the actually quality framework and its three stages (raw data level, census data base, final data pool) as well as the different types of attributes (simple, multiple, derived). The quality of each attribute is evaluated on these three different stages. Therefore, the whole data editing process can be monitored. This comes along with key figures of the RBLMS. Finally, the prospects are certain findings and further developments.

Keywords: administrative data, register-based census, quality assessment, Austrian Census 2011.

Sampling and Estimation in Finite Networks

Melike Oguz–Alper[1] and Li–Chun Zhang[2,3,4]

[1]Statistics Norway, e–mail: Melike.Oguz.Alper@ssb.no
[2]University of Southampton, e–mail: L.Zhang@soton.ac.uk
[3]Statistics Norway
[4]University of Oslo

Abstract

The conventional design–based inference or the randomisation approach conceives population data as a list of units where a set of measurements, which are assumed to be constant, are attributed to the individual units. The randomness is solely specified by a well–defined probability sampling design that assign probabilities to the samples selected repeatedly from a finite population. Population data may have a hierarchical structure where the lower–level units are nested within the higher–level units. Sampling and estimation techniques are well established for such finite list populations, regardless of whether the units are in a hierarchical structure or not.

The variety of data available today, however, raises other possibilities of representation. Data may contain non–nested relationships with many–to–many linkage between the units, unlike the conventional envision of data structure as a tree of units with only one–to–one or one–to–many relationships. There may be multiple types of relationship between the units. Such complex structures can often be represented by a graph consisting of a set of nodes and a set of edges. A valued graph where the measurements are attributed to the graph objects is called *network*. We can think of social networks, transportation networks, labour–flow network, communication networks, computer networks, etc. The parameter of interest is not necessarily defined only on the nodes, but the relational structure itself may be of interest. There is a large literature on the model-based inference in networks. However, the modelling approach may not always be viable especially when the underlying dynamics are too complicated or transient or subject to shocks. The randomisation approach to finite networks may be more useful in that case.

The theory of sampling and inference in finite networks is relatively underdeveloped, and the techniques are rarely applied in Official Statistics, despite some notable exceptions in the past such as multiplicity sampling including indirect sampling and adaptive cluster sampling. There is a recent article by Zhang and Patone (2017) which is a synthesis and extension of the graph sampling theory by covering all the existing network sampling techniques as special cases. They develop a general Horvitz & Thompson's (1952) estimator under arbitrary T–stage snowball sampling.

In this talk, at first, we will move quickly from the conventional design-based approach to finite list populations, to providing a formal definition of sampling in finite population networks. The existing multiplicity sampling techniques will be discussed in order to provide a better picture of how they could be, in fact, envisaged as network sampling techniques. When it comes to the inference in finite networks, target parameters are not limited to the

totals of measures attributed to the nodes. Thus, we consider a design–based inference for higher–order target parameters which are defined based on the measures associated with the relationships or edges. Examples of such network parameters are network density, reciprocity, number of dyads or triads, transitivity, etc. We establish generally the relative efficiency of two types of Horvitz–Thompson estimators for network sampling. Results from a limited simulation study with an application to a labour–flow network will be presented, where the industrial sectors are the nodes and the labour flows between the sectors form the edges. The data used is retrieved from the Norwegian Income and Employment data in the administrative sources.

Keywords: Graph sampling, finite networks, network parameter, multiplicity–sampling

Comments on the development of Survey Statistics theory and practice in the last fifty years; some personal reflections.

Carl-Erik Särndal

Statistics Sweden

Abstract

The continued development of the field of Survey Statistics will be interesting. One reason is that the data collection phase for producing official statistics is likely to change, possibly to alternatives other than the probability sampling data collection that has been a standard, or seen as an ideal. This presentation cannot predict the development; it looks instead at some important ideas in the progression of Survey Statistics over the last five decades.

In the more than one hundred years history of survey sampling, a more than sixty year old result has had a particular significance, namely, that unbiased estimation of a finite population total is obtained by weighting the observed survey variable values by the inverse of the inclusion probabilities. This works, because in probability sampling, these probabilities are known for all population units. The unbiased estimator that expresses this fundamental and mathematically simple result bears the name of the two auteurs, Horvitz and Thompson, of a classical 1952 JASA article. But behind the result lay a long development, from the early attempts of statisticians to convince users of statistics that "observing just a small sample from the large finite population can be enough". Although interesting in a historical perspective, this long period is not considered in this presentation.

Inverse inclusion probability weighting, and its modifications and extensions, have had a strong impact on survey statistics over the last fifty years, which is the period examined here. Such weighting is the basis for what we now call *design-based inference*.

By contrast, an alternative approach known as *model-based inference* will, at least in its most pure forms, deny that any important role be given to probability sampling and to inverse inclusion probability weighting. Modeling, and trust in the assumed models, is the justification for the inference. Although not design unbiased, the resulting estimates may be advantageous in other ways.

A feature of the last fifty years of development is the importance of auxiliary variables in

the estimation process for official statistics. This has been particularly evident in northern European countries, with their access to a vast supply of auxiliary variables, from administrative registers, or in the form of paradata.

Several areas of research and practice have extended the design-based inference paradigm. The two areas mentioned below were, in their original form, presented for a survey background that national statistical institutes cannot count on now, several decades later: a full, or almost full, 100% response from the selected probability sample.

1) The generalized regression (GREG) approach originated in the realization that while inverse inclusion probability weighting is needed for design unbiased estimation, such unbiasedness is not the only important factor. The estimation also needs to be variance efficient. The GREG estimation approach realizes a low variance from a strong regression existing between survey variable y and auxiliary vector $\mathbf{x}$. One can explain it by saying that accurate prediction of the unobserved y-values is derived from the information on $\mathbf{x}$ known for the population.

2) The calibration (CAL) approach had its origin in a search for a weighting of the observed sample y-values that is not far from the basic inverse probability weighting, but better than it, because required to respect a condition called a calibration equation, where the known population total of the x-vector, or a design unbiased estimate of it, appears on one of the two sides of the equation. But a secondary purpose is to explain the survey variable y through the auxiliary vector x. The calibration approach is thus double-natured: The weighting aspect is combined with an implicit relationship between y and $\mathbf{x}$. Although the outlook is different, the CAL approach is in special cases identical to the GREG approach.

Both 1) and 2) can be called *design-based model assisted inference*. However, the last few decades have witnessed a strong adverse trend for the conditions for probability sampling surveys: high rates of nonresponse in the drawn probability sample. It has become necessary to adapt the inference – which can perhaps no longer be called design-based - to these new conditions.

High nonresponse causes a more or less pronounced bias in the survey estimates. This can happen even under conditions of quite strong relationship between y and $\mathbf{x}$. The objective is then to hold this bias as low as possible. The CAL approach that has been particularly important and useful for nonresponse weighting adjustment. Auxiliary variables are also important for managing the data collection so as to get a well-balanced set of respondents from the drawn probability sample.

The presentation reviews briefly the approaches 1) and 2), then focuses on approaches to inference under (high) survey nonresponse. A question arising is: How important will the probability sampling paradigm and the inverse inclusion probability weighting be in the future?

Population Census in the Republic of Belarus: Experience and Perspectives

Natalia Bandarenka [1]

[1]State Institute of Management and Social Technologies of the Belarusian State University,
e-mail: bondnata@mail.ru

Abstract

The paper considers the main questions of the program, methodology, design and distinctive features of carrying out the population census in the Republic of Belarus
Keywords: population census, census program.

1 Introduction

The population census provides unique information about the size and structure of the population by sex, age, nationality, education, marital status, occupation and other characteristics and is the major information resource about the population received at the state level by interviewing citizens.

The main advantages of the population census are that it provides demographic indicators at the level of the smallest administrative-territorial units and these indicators are comparable between the territories since they relate to a single moment of time.

2 Population Census in Belarus

2.1 History and conducting principles

Within the borders of modern Belarus the population census took place 9 times: in 1897, 1926, 1937, 1959, 1970, 1979 and 1989. The census was not only in 1949 and in 1969: in 1949 because the country's leadership did not want to show the true losses of the Second World War; in 1969 because of the difficult economic situation (the census was held in 1970 coinciding with the centenary of Lenin's birth).

In the history of independent Belarus there were two censuses: in 1999 and in 2009.

There are the next basic principles of carrying out a population census in Belarus:

- The generality (a census covers all territory and all population).

- Simultaneity (the choice of the critical moment (date and time of a census)). All data collected during the census belong to one certain and in advance determined moment.
- Uniform program of a census: collecting data on the same signs, characteristics specified in the census form, for all rewrites.
- Collection of personal and easily identifiable information about each individual person.
- Self-determination: all information is collected only from the words of the respondents; it is prohibited to require their documentary confirmation. The only exception is when the respondent claims to be 100 or more years old.
- Confidentiality (the prohibition of informing someone about personal information received about the census during the census).
- Strict centralization of census management: the state assumes responsibility for conducting, monitoring and financing the census.
- Regularity of the census. In Belarus the requirement to conduct a population census at least once every 10 years is legally fixed.

The census program traditionally consists of three sections:

1. An address part (name and the address of the rewritable, its relation to the head of a family/household);

2. The actually a census program (personal demographic characteristics (sex, age, marriage status), social and economic characteristics (education level, profession, occupations, income, social status), ethnic (ethnic origin, native language, language skills, religion / confessional accessory) characteristics; the questions connected with studying of population reproduction; migration);

3. The questions connected with other survey (determined by the objectives of this particular census).

2.2 Population Census in 1999

From February 16 to February 23, 1999 in Belarus the first population census among the CIS (Commonwealth of Independent States) countries has been conducted. Difficulties in the organization of carrying out a census have been caused by the fact that before Belarus was a part of the USSR and all leadership in process of collecting and the analysis information was carried out through Moscow. 33 thousand specialists were attracted for the census; each of specialists was supposed to interview people from 25 to 40 apartments on the day. The method of the census is «face-to-face» interviewing.

Unlike previous censuses only the permanent population was rewritten and not present population as before. This was done to save money (as the budget for one person was 0.5-0.7 USD), and also taking into account the experience of other countries.

In developing the census program the 1989 program was used as a basis, but with significant changes ((instead of 13 main questions the program has included 17). The

questionnaire was not only in Russian, but also in Belarusian.

The main differences in forming the questions of the census program were:

- For the first time not only registered but also unregistered marriage was taken into account;
- The question was included about how many children people not only have but also plan to have;
- Since the peculiarity of Belarus is the spread of the language, the questionnaire was asked not simply "What is your native language?" as it is done in other countries and in what language the person speaks at home and what other language is fluent;
- The question of how many sources of livelihood were available has been for the first time included.

2.3 Population Census in 2009

The population census in 2009 was conducted from October 14 to October 24. In comparison with a census of 1999 duration of the period of a census has increased from 8 to 11 days that has allowed reduced the burden on the specialist to 300 people and, thereby, to increase quality of the survey.

More than 48 thousand temporary specialist of whom 63% employees of the organizations, 33% – pupils and students, 4% – pensioners have participated in a population census. More than a half of participants of a census had the highest or average special educations.

In addition the pilot census showed the need to organize the work of stationary census plots (about 5% of the population living in the pilot census area visited such sites). Thus, for the first time in the population census of Belarus there were more than 3000 stationary census plots where respondents could indicate information about themselves.

Poll of respondents and filling of questionnaires were carried out in Belarusian or Russian at the request of the respondent.

The census program included 9 main thematic clusters:

- The number and location of the population;
- Demographic characteristics;
- Level of education;
- Socio-economic characteristics;
- National composition of the population, citizenship;
- Population migration;
- Characteristics of households;
- Housing conditions;
- Population temporarily residing (residing) in the territory of the Republic of Belarus.

The distinctive features of the 2009 census program were following:

Unlike the 1999 the census program in 2009 contained an additional question about the type of educational institution in which the respondent studied.

The program for the first time includes information characterizing the labor migration. The data indicate migration flows within the administrative region, country region, the republic, as well as the number of external labor migrants. At the same time, information on the location of the main work of labor migrants is given in combination with their age, level of education, status in employment, occupation, type of economic activity, place of residence.

2.4 Population Census in 2019

The next population census in the Republic of Belarus will be held in 2019 and will be conducted to the period from October 4 to October 30. Duration of a census will be 27 days.

On October 4,5,6 2019 the census will be conducted at stationary sites (according to the principle of elections), then from October 7 to October 11, 2019, the lists will be updated and from October 12 to October 30, 2019 a round of interviewing by specially trained people in households will be carried out (at the same time stationary sites continue to work).

Unlike the two previous censuses the population census-2019 will be conducted in three ways:

- «face-to-face» interviewing by a specialist;
- on stationary sits;
- by the Internet (the principle of self-registration).

The innovation of the forthcoming census is the replacement of traditional paper questionnaire with tablet computers in which census forms will be downloaded electronically, as well as maps of sites with addresses and house outlines. Automation of data entry at the survey stage of respondents will allow to provide high quality of filling out questionnaires due to the connection of the control system. It is provided to use the information system "Register of the Population" as a basis for filling of an address part of the questionnaire. It will allow to obtain automatically about 20% of the information needed to fill out the questionnaire (a full name, the identification number, date and place of birth, gender, citizenship, place of residence and place of stay). And only the remaining missing information will be obtained through an of interviewing.

Using tablets will also increase the load per specialist to 750 people (in 2009 were 300 people). As a result in 2019 it is planned to attract about 2.5 thousand people for the census (in 2009 there were 7,5 thousand people). Thus, the number of temporary census staff will be cut 3 times in comparison with the previous census.

On the first time within the population census of 2019 in the Republic of Belarus there

will take place the agricultural census (which according to the recommendations of FAO has to be conducted once in 10 years). In the Republic of Belarus the National Statistical Committee provides current statistical accounting of the main agricultural organizations activity and also makes selective monitoring of agricultural activity of the citizens having personal subsidiary farms and constantly living in rural areas (a survey of private subsidiary plots in rural areas (from 2011)). At the same time the population carrying out agricultural activities in urban areas, in garden associations, seasonal houses and summer cottages is not examined. It has caused need of inclusion of questions about their agricultural activity in the census program.

The questionnaire on agricultural activities includes a minimum set of indicators allowing to specify existence in the property (possession, use, rent) of the household, their location (urban or rural area, garden associations), determine the structure of arable land, the number of perennial plantations , the number of livestock, poultry and bee colonies.

References

In Belarus the first population census is conducted (1999) [Electronic resource] http://90s.by/years/1999/perepis.html

National Statistical Committee of the Republic of Belarus (2010). *Population census of 2009. Basic organizational and methodological provisions.*

National Statistical Committee of the Republic of Belarus (2017). *About preparation for carrying out in 2019 a population census of Republic of Belarus.*

Use of Register Data in Latvian Household Finance and Consumption Survey

Andris Fisenko[1] and Jānis Lapiņš[2]

[1] Bank of Latvia, e-mail: andris.fisenko@bank.lv
[2]Bank of Latvia, e-mail: janis.lapins@bank.lv

Abstract

The Household Finance and Consumption Survey (HFCS) is a statistical survey conducted in the euro area countries by collecting and compiling data on the real assets, financial assets, debt, income and consumption of households. The HFCS is carried out by the European Central Bank and the national central banks of the European Union Member States. The HFCS is conducted at the national level. To obtain comparable data, the participating countries follow common methodological guidelines (Household Finance and Consumption Network 2016), but do not necessarily use identical questionnaires.

The Latvian HFCS for the second time was conducted by the Bank of Latvia in 2017, again in a close cooperation with the Central Statistical Bureau of Latvia (CSB). CSB ensured the collection of the HFCS data and the adding of respondents' data from several administrative data sources to the survey data. These administrative data, as well as the comments and the paradata provided by interviewers at the conclusion of each interview, are used at the Bank of Latvia during the data editing phase to detect and correct possible mistakes in the survey data. Such quality checks aim to correct various kinds of inconsistencies, such as mistyped or erroneous answers.

The quality of the survey data on the participation of household members in the first and second level pension scheme collected in the previous HFCS wave, in 2014, was very poor. Therefore, in the current survey wave, it was decided to exclude the questions related to the first and second level pension scheme from the questionnaire, and to obtain the necessary data from the State Social Insurance Agency. This decision allowed to obtain high quality data on persons' participation in public pension schemes, as well as to reduce respondents burden, too.

Among administrative data sources that we use for editing of the HFCS data are:

- the State Revenue Service (SRS) data on all type of persons' income in 2016,

- SRS data on persons' participation in the voluntary (third level) pension schemes,
- the Land Cadastre's data on real estate properties that belong to the household members,
- the Credit Register data on persons' mortgages, loans and/or leasing contracts.

Data editing is one of the most important, intensive and time-consuming task of HFCS. For the current survey wave it is still ongoing. In our presentation we plan to report some first results showing usefulness of administrative data for editing of the HFCS data.

References

Household Finance and Consumption Network (2016) The Household Finance and Consumption Survey: methodological report for the second wave. ECB Statistical Paper Series, 17.

Calibration Weighting in Survey Sampling (Based on Sample Socio-Demographic Survey)

Yuliia Halytsia

Taras Shevchenko National University of Kiev, e-mail: halytsya2013@gmail.com

Abstract

Today, one of the most pressing informational and statistical problems is the problem of ensuring the reliability of the results of sample population surveys. The solution to these problems is closely linked to the creation and use (including further adjustment) of the system of statistical weights of the sample survey. This contributed paper contains a short overview of calibration method and its applying to sample socio-demographic survey.
Keywords: calibration, household survey, statistical weights

1 Introduction

Large-scale population surveys are unique on the basis of an array of received primary data, a system of indicators, evaluated on the results of the survey, the principles of organization and conduct of the survey, data processing, etc. First of all, it concerns sample surveys of the population, because, firstly, the design of the sample is developed on the basis of existing actual data sources and is, in a sense, unique; and secondly, at many stages of the survey it is necessary to take into account the fact that not all the general population is examined, but a certain, specially selected part of it.

It is necessary to pay much attention to the problem of coordinating the results of sample surveys of the population with available high-quality external information to increase their representativeness and usefulness. Such coordination is most often appropriate at micro level - for individual units: individuals, households, etc. The expediency of coordinating the results of surveys with external information is conditioned, at least, by the fact that: the evaluation of the indicators based on the results of sample surveys is characterized by a certain error due to lack of observation, as well as non-sample mistakes; a sample population survey cannot provide estimates of certain characteristics of the general population (although they are obtained on the basis of the survey, but mainly reflect the parameters laid down in the formation of the sample); the main characteristics of the general population can change rapidly in time, so at the time of completion of the processing of the survey data - somewhat different from those at the time of the organization of the survey.

Ensuring the maximum quality and usefulness of the results of sample surveys in estimation of the target characteristics of the general population is the main purpose of the system of statistical weights. The most reasonable way of solving the problem of coordinating the results of a sample survey with external data is the corresponding adjustment of the system of statistical weights.

2 Calibration method

A method of calibration is the most theoretically developed and effective from the currently known generalized methods for adjusting the system of statistical weights in order to coordinate the results of the survey with several external distributions for different types of units. The method consists in solving a special task of minimizing the change in the value of weights in the process of coordinating the results of the survey with the external data. Usually, at the same time as the sample is formed, the basic statistical weights are calculated. On the following stages, which are implemented after the survey, the basic weights are adjusted precisely in order to take into account structural features of the general population.

The statement of the calibration problem as an optimization problem of minimizing the distance between design weights and calibrated (adjusted) weights, provided that in the household / population survey reliable external information on the total number of households and gender-age structure of the population is available, can be represented as follows:

$$\begin{cases} \sum\limits_{i=1}^{n} q \dfrac{\left(w_i^{(c)} - w_i^{(d)}\right)^2}{w_i^{(d)}} \to \min; \\ \sum\limits_{i=1}^{n} k_{ji}^{(f)} w_i^{(c)} = F(j); \;\; j = 1,2,...,J; \\ \sum\limits_{i=1}^{n} k_{li}^{(m)} w_i^{(c)} = M(l); \;\; l = 1,2,...,L; \\ \sum\limits_{i=1}^{n} w_i^{(c)} = H, \end{cases}$$

where:

n – the sample size of the population / households who participated in the survey;
q - parameter;

$w_i^{(d)}$ - statistical weight of i - th respondent / household, which needs to be adjusted;

$w_i^{(c)}$ - weight of i - th respondent / household after calibration;

$k_{ji}^{(f)}$ - the number of women in i - th household, which, according to the survey, belong to the same j - th gender-age group;

$F(j)$- the total number of women in j - th gender-age group according to external data;

J - number of gender-age groups for women;

$k_{li}^{(m)}$ - the number of men in i -th household, which, according to the survey, belong to the same l - th gender-age group;

$M(l)$- the total number of men in l - th gender-age group according to external data;

L - number of gender-age groups for men;

H - total population by external data.

The solution of the formulated problem can be accomplished using method of Lagrange multipliers.

3 Application of calibration method

For the practical part of the work we used the results of a sample survey "Social Inequalities: Perceptions by Ukrainian Society", conducted by the Center "Social Monitoring" in 2017.

As a subject of research and evaluation for the general population, we chose nominal variables that reflect the perceptions of the population about the minimum necessary and sufficient level of income, as well as the level at which a household can be considered poor.

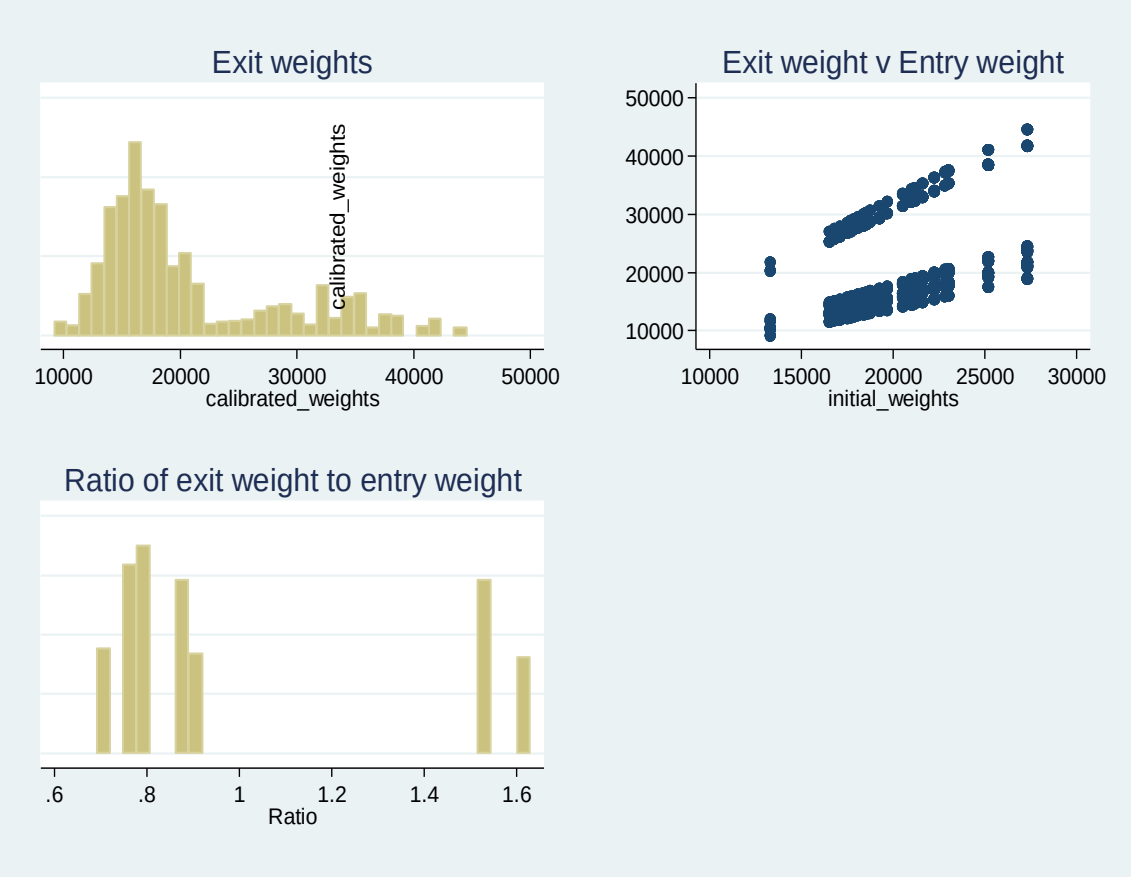
Base on the survey data, were calculated average sample values of: *monthly average per capita income per family member, which the respondent considers to be sufficient for normal life* and *monthly average per capita income, which, according to the respondent's opinion, provides the living wage*, as well as *monthly average per capita income for which family can be considered poor*. These indicators are to be estimated for the general population after calibration of statistical weights.

Given the peculiarities of the sample design and the used system of primary weights of the socio-demographic survey "Social Inequalities: Perceptions by Ukrainian Society", we consider it expedient to perform the calibration procedure using a combination of such variables as gender and age (gender-age structure of the population). Within this work, calibration was be carried out only at the national level, without considering the type of settlement or regions.

Table 1: Primary and calibrated statistical weights

Weights	Number of observations	Average	Standard deviation	Minimum value	Maximum value
Primary	2 046	20 730.7	3 135.8	13 313.9	27 317.5
Calibrated	2 046	20 730.7	7 814.7	9 217.1	44 516.9

Figure 1: Results of calibration of statistical weights



The obtained results of calculations of the mean values of the indicators for the sample and estimates for the general population (that is, after the calibration procedure) are given below:

- Monthly income per family member, which is considered sufficient for normal life:

Primary weights

```
Survey: Mean estimation

Number of strata =       1      Number of obs   =       2,046
Number of PSUs   =   2,046      Population size =  42,414,905
                               Design df       =       2,045
```

	Linearized Mean	Std. Err.	[95% Conf.	Interval]
f02	9043.391	297.785	8459.398	9627.385

Calibrated weights

```
Survey: Mean estimation

Number of strata =       1      Number of obs   =       2,046
Number of PSUs   =   2,046      Population size =  42,414,905
                               Design df       =       2,045
```

	Linearized Mean	Std. Err.	[95% Conf.	Interval]
f02	8663.918	280.934	8112.972	9214.865

- Monthly income per person, which provides a living wage for today:

Primary weights

```
Survey: Mean estimation

Number of strata =      1       Number of obs   =      2,046
Number of PSUs   =  2,046       Population size =  42,414,905
                                Design df       =      2,045

                    Linearized
             Mean   Std. Err.     [95% Conf. Interval]
-------------------------------------------------------------
f03      4699.368   79.73994      4542.988    4855.748
```

Calibrated weights

```
Survey: Mean estimation

Number of strata =      1       Number of obs   =      2,046
Number of PSUs   =  2,046       Population size =  42,414,905
                                Design df       =      2,045

                    Linearized
             Mean   Std. Err.     [95% Conf. Interval]
-------------------------------------------------------------
f03      4570.809   74.74176      4424.231    4717.387
```

- ***Family with such monthly income per capita can be considered poor:***

Primary weights

```
Survey: Mean estimation

Number of strata =      1       Number of obs   =      2,046
Number of PSUs   =  2,046       Population size =  42,414,905
                                Design df       =      2,045

                    Linearized
             Mean   Std. Err.     [95% Conf. Interval]
-------------------------------------------------------------
f04      2466.568   48.82644      2370.814    2562.323
```

Calibrated weights

```
Survey: Mean estimation

Number of strata =      1       Number of obs   =
Number of PSUs   =  2,046       Population size =  42,4
                                Design df       =

                    Linearized
             Mean   Std. Err.     [95% Conf. Int
-------------------------------------------------------------
f04      2395.195   46.65419      2303.701    2
```

Conclusion

Average values for all investigated nominal variables after calibration were lower than before the calibration procedure for statistical weights was performed. We believe that this is due to consideration of a high level of demographic aging in Ukraine in estimating indicators for the general population, since elderly people are usually distinguished by comparatively lower financial needs and showing lower self-esteem levels of necessary and sufficient income.

References

Sarioglo V.G. (2012) Estimation of socio-economic indicators: applied aspects of the application of indirect methods. *Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine*, 136

D'Souza, J. (2010). Calibrate: a Stata Program for Calibration Weighting.- *London: Stata User Group.*

Kish L. Survey sampling / Kish L. – *New York : John Wiley & Sons, 1995*. – 643 p.

Rozora I., Lukovych O. Mean Estimation with Robust Calibrated Estimators// *Baltic-Nordic Summer School on Survey Statistics.- Kyiv, 2016.*

All-time low period fertility in Finland: tempo or quantum effect?

Julia Hellstrand[1]

[1]University of Helsinki, e-mail: julia.hellstrand@helsinki.fi
or, Statistics Finland, e-mail: julia.hellstrand@stat.fi

Abstract

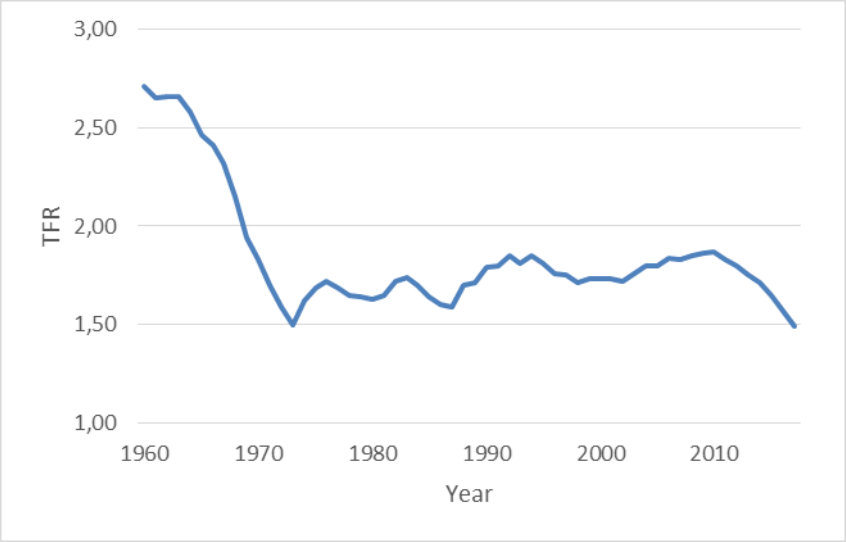
This is a short description of my Master's thesis that I am currently working on. It deals with the decreasing period fertility rates in Finland since 2010 and forecasts cohort fertility.
Keywords: Period fertility, postponement, tempo effect, quantum effect, cohort forecasting

1 Introduction

In Finland, the number of live births decreased from 60 980 in 2010 to 50 321 in 2017. The most commonly used fertility index, the total fertility rate (TFR), did also decrease rapidly in the 2010s and experienced an all-time low rate of 1.49 live births per woman in 2017 (see picture 1). The recent years' decline in the number of births and in the TFR is a subject of concern that has been frequently in the Finnish news lately, and the aim with my thesis is to understand the rapid decrease in the 2010s.

The total fertility rate (TFR) is the average number of children that would ever be born to a woman if she were to experience the exact current age-specific fertility rates through her lifetime and she were to live to the end of her child-bearing years. Since period-based measures are by nature synthetic, meaning that no real group of women necessarily will experience the fertility rates of one calendar year through their lifetime, the TFR comes with limitations. Shifts in the TFR depends both on fertility timing, tempo, and by changes in the total number of children women have, quantum (Myrskylä et al. 2013). It is known that postponement of first birth is an ongoing and persistent process in most developed countries (Andersson et al. 2009) and that fertility rates are depressed when women delay childbearing in a given period (Bongaarts and Feeney 1998). Thus, the recent decrease in the TFR could be due to a tempo effect, meaning that women are postponing their births but not necessarily having less children, or it could be due to a pure quantum effect which by time would be seen as a decrease in the completed cohort fertility rate as well.

Picture 1: TFR in Finland in 1960-2017



Source: Statistics Finland 2018

2 Goals and methods

My thesis has three main goals; (1) to describe period fertility trends in Finland among age, parity, regions and levels of education, (2) to calculate an alternative tempo adjusted fertility rate that adjust for fertility timing and (3) to forecast cohort fertility in Finland. Period fertility trends will be described by age-specific fertility rates and mother's mean age of childbearing at different points in time. The drop in the TFR will be examined by demographic decomposition (Andreev and Shkolnikov 2012) and the tempo adjusted fertility rate will be computed by the method developed by Bongaarts and Feeney (1998). Cohort fertility will be forecasted mainly by a Bayesian method developed by Schmertmann et al. (2014) but also by simpler methods like Freeze Rates (e.g. Frejka and Calot 2001a) and 5-year linear extrapolation (Myrskylä et al. 2013). By the time of the workshop in August, I will try to complete as many goals as possible. The results will be presented at the workshop.

References

Andersson, G., Rønsen, M., Knudsen, L. B., Lappegård, T., Neyer, G., Skrede, K., Teschner, K. & Vikat, A. (2009). *Cohort fertility patterns in the Nordic countries.* Demographic research, 20, 313-352.

Andreev, E. M., & Shkolnikov, V. M. (2012). *An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages.* MPIDR Technical Report TR-2012-002

Bongaarts, J. & Feeney, G. (1998). *On the quantum and tempo of fertility.* Population and Development Review 24(2): 271-291

Frejka, T., & Calot, G. (2001a). *Cohort reproductive patterns in low-fertility countries.* Population and Development Review 27(1): 103-132.

Myrskylä, M., Goldstein. J.R. & Chen, Y.A. (2013). *New Cohort Fertility Forecasts for the Developed World: Rises, Falls, and Reversals.* Population and Development Review, 39, 31-56.

Schmertmann, C., Zagheni E., Goldstein J. R., and Myrskylä M. (2014) *Bayesian Forecasting of Cohort Fertility*, Journal of the American Statistical Association, 109:506, 500-513

Estimation of response propensities using the previous survey

Miika Honkala

Statistics Finland, e-mail: miika.honkala@stat.fi

Abstract

This paper studies how response propensities, estimated using the dataset of the previous survey, predict actual response rates. In this study, two consecutive datasets of same survey were available. Response propensities were estimated to the older dataset using logistic regression model. Then the propensities were imputed to the newer dataset. The result was that the imputed response propensities predicted response behavior quite well.
Keywords: response propensity, response rate

1 Introduction

Many surveys are carried out annually. The implementation of the surveys and response behavior remain quite similar in consecutive years. If a survey has conducted a number of times previously, it may be possible to determine different optimal designs for different subgroups on the basis of the past experience (Tourangeau et al. 2017). However, it is good to check several rounds if available and look forward how regular response rates are (Laaksonen 2016). Schouten et al. (2017) present adaptive survey design which offers several methods for data-driven tailoring of data collection.

One possibility to utilize a previous survey is to estimate response propensities before the data collection of the survey, using the dataset of the previous year. For the data collection, it may be beneficial to know estimated response propensities in advance. If response propensities are known, the data collection can be designed in a new way. The response propensities may be utilized to tailor the data collection.

This study was carried out using Statistics Finland's data. Two datasets were available: datasets of European Social Survey from rounds 7 (2014) and round 8 (2016). In this paper, ESS7 means the dataset of round 7 and ESS8 means the dataset of round 8. Response propensities were estimated to the ESS7 using logistic regression model. After this, the propensities were imputed from the ESS7 to the ESS8. An interesting question was how the imputed propensities predict actual response rates. Were individuals with low response propensities often non-respondents? On the other hand, were individuals with high response propensities often respondents?

2 European Social Survey

ESS is a cross-national survey that has been conducted across Europe since 2001. Its target population consists of all residents 15 years or older who are residents of the country within private households. The ESS is conducted every two years using face-to-face interviews. In Finland, the sample size of ESS was 3 400 in the rounds 7 and 8. Response rate was 62.7% in the round 7 and 57.7% in the round 8. ESS's websites include more information about the ESS.

3 Methods and results

The sizes of the datasets were 3400, including respondents and non-respondents. Both datasets contained a binary response indicator (1 = respondent, 0 = non-respondent) and a lot of register variables. A response propensity model was fitted to the ESS7. The model was a logistic regression model where the dependent variable was the response indicator. Explanatory variables of the model were selected from the register variables. In the final model, the explanatory variables were municipality group, gender and interaction age group x education. These variables had a statistically significant effect on response. All the explanatory variables were classified variables. The model was made using SAS program.

After modeling, the response propensities were imputed from the ESS7 to the ESS8. The imputation method was a donor-recipient method based on the explanatory variables of the model. A donor in the ESS7 and a recipient in the ESS8 had same characteristics (the same values in the variables municipality group, gender and interaction age group x education). For example, suppose that a donor person in the ESS7 had response propensity 0.53 and following characteristics: municipality group = 2 (semi-urban municipalities), gender = 1 (male) and age group x education = 5 (30-44 years, no final examination). If the ESS8 included people who had exactly same values in these variables, these people got imputed response propensity 0.53.

The ESS8 were divided into groups according to imputed response propensities. Actual response rates in the six response propensity groups in the ESS8 are shown in Figure 1.

Figure 1. Response rates in the ESS8.

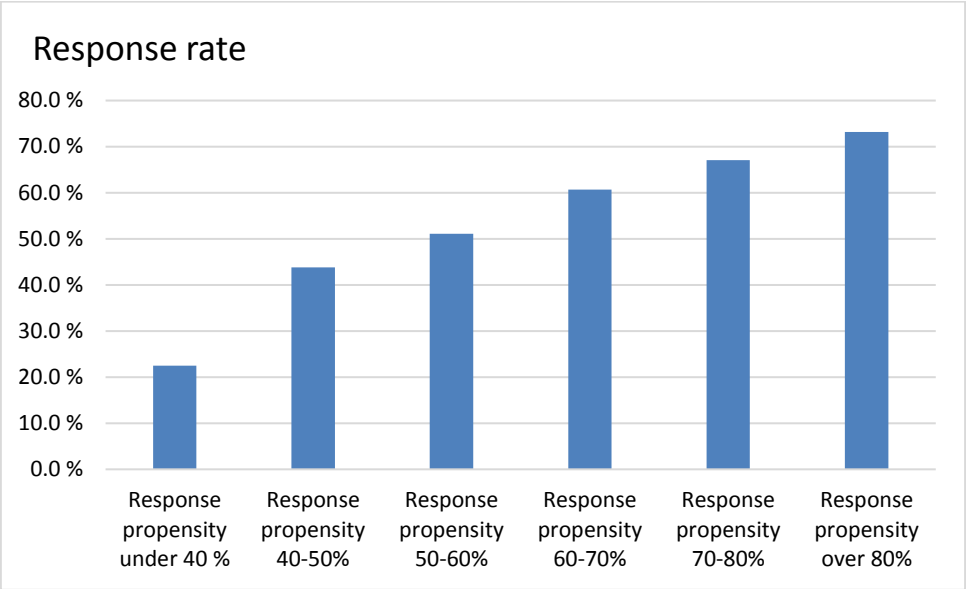


Figure 1 shows that imputed response propensities predict actual response rates quite well. The curve is rising, so there is a clear link between the imputed response propensities and the actual response rates. In a group where response propensities are under 40%, response rate is slightly more than 20%. In a group where response propensities are over 80%, response rate is more than 70%.

4 Conclusion

Using a good response propensity model, response behavior can be predicted before the data collection of the survey has begun, if the dataset of the previous survey is available. Predicting response rates is useful in surveys that are carried out annually, and where the sampling design and the implementation of the survey generally remains the same in consecutive years. When possible challenging respondents are known in advance, it is possible to plan the data collection in a new way and consider motivation letters or incentives for some respondents, for example. In telephone interview surveys, response propensities can determine the number of calls. The lower response propensity, the more contact attempts. It could be interesting to test this kind of responsive data collection which is based on the previous survey.

Utilizing a previous survey could also be beneficial when planning a new sampling design. Laaksonen (2016) presents a stratified sampling method which is based on the response rates of the previous survey. In that method, the strata are formed according to auxiliary variables. Sample sizes are bigger in groups where response rates have been low, and smaller in groups where response rates have been high in the previous survey. This kind of sampling method may lead to more representative set of respondents.

References

ESS's websites,

http://www.europeansocialsurvey.org/ (referenced 25.5.2018)

Laaksonen, S. (2016). Anticipation of unit nonresponse or not in the sampling designing From the point of view of the European Social Survey (ESS). International Workshop on Household Survey Nonresponse, 2016, Oslo.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive survey design*. Champman and Hall, New York.

Tourangeau, R., Brick, M., Lohr, S. and Li, J. (2017). Adaptive and Responsive Survey Designs: a Review and Assessment. *Journal of the Royal Statistical Society A* **180**, 203-223.

Effect of using Tobit and Heckit models in regression estimation for data with many zeros

Tetiana Ianevych[1]and Veronika Serhiienko[2]

[1]Taras Shevchenko National University of Kyiv, e-mail: yata452@univ.kiev.ua
[2]Taras Shevchenko National University of Kyiv, e-mail: nichka_2009@ukr.net

Abstract

In the work we examine the effect of using the Tobit and Heckit models as assisting for the generalized regression estimator in order to improve it for data containing many zero values.
Keywords: Tobit model, Heckit model, regression estimation, excess of zeros in data.

1 Introduction

It is rather frequent situation when the economic data, especially microeconomic data, contain observations where some variable of interest is equal to zero for a number of the observations in the data set. Such data have excess of zero values and this can lead to a number of econometric problems when using Ordinary Least Squares (OLS) to estimate the unknown parameters of a regression model. We faced with this problem when start to work with Ukrainian capital expenditure survey.

One of the models that widely used in such situations is the Tobit model introduced by Tobin in 1958. It is developed for the censored data. For the data, suffering from big number of zeros but not caused by censoring, another models can be used – the Heckit model. We examined the usefulness and accuracy of these models utilizing general linear regression estimator (GREG). For this we use Monte Carlo simulation method measuring the efficiency with the Absolute Relative Bias and the Relative Root Mean Square Error.

2 Models for data with excess of zeros

The key decision facing any researcher working with a data set containing zeros is the choice of the appropriate model. The following summarizes the key elements of such a decision. Suppose that the variable of interest is y_i, and there are a large number of zero values for y in a given data set. The first step is to determine why the zeros are present in the data. There can be two alternatives:

(1) the zeros appear as a result of censoring or

(2) the zeros represent a decision that the researcher has no control over for some reason.

The first alternative usually corresponds to Tobit model whereas the second one – to Heckit model. Let us consider them in details.

2.1 Tobit Model

The Tobit Model was introduced by Tobin in 1958. The Tobit model, also called a censored regression model, is designed to estimate linear relationships between variables when there is either left- or right-censoring in the dependent variable. Formally, it can be written as

$$y = \begin{cases} y^*, if\ \ y^* > \tau \\ \tau_y\ , if\ y^* \leq \tau \end{cases}$$

where $y_i^* = X_i\beta + e_i$, $e_i \sim N(0, \sigma^2)$. The most common choice is $\tau = \tau_y = 0$.

The coefficient of such model can be calculated using the Maximum Likelihood method. In R there is a function "tobit()" in the *AER* package developed for this task. The Tobit Model allows different generalizations. For more information see Humphreys (2013).

2.2 Heckit Model

This type of model is appropriate when $y_i = 0$ because of the non-observable response. It means that knowledge $y_i = 0$ is uninformative in estimating the determinants of the level of y_k We can formulate it starting from the "participation" equation

$$z_i^* = \omega_i\gamma + u_i$$

$$z_i = \begin{cases} 1, if\ \ z_i^* > 0 \\ 0, if\ z_i^* \leq 0 \end{cases}$$

and continuing with "consumption" equation

$$y_i = \begin{cases} x_i\beta + e_i, if\ z_i^* > 0 \\ \quad 0\ , if\ z_i^* \leq 0 \end{cases}$$

with errors $u_i \sim N(0,1)$ and $e_i \sim N(0, \sigma^2)$, to be correlated in general case $corr(u_i, e_i) = \rho$. This specific terminology comes from Jones (1989) who investigated cigarette consumption.

The coefficient of the Heckit model can be calculated using the Maximum Likelihood method or 2 step method, developed by Heckman (1976). In R you can use the function "selection()" inside the *sampleSelection* package developed for this task.

3 Analysing the simulated data

So, we want to investigate the efficiency of the general regression estimator based on Tobit and Heckit models comparing to the classical Horvitz-Thompson and regression estimator based on classical linear model.

Our first simulated population *U* consists of *N*=1000 elements for which we produce the values of y_i as follows

$$y^* = -2.35 + 1.6578 \cdot x + e, \text{ where } e \sim N(0,1) \text{ and}$$

$$y = \begin{cases} y^*, & y^* > 0 \\ 0, & y^* \le 0 \end{cases}.$$

The parameter of interest is the total $Y = \sum_{i \in U} y_i$. The underlying design is simple random sampling of size 100.

The main relative measure of efficiency for the estimators we used are:

the absolute relative bias

$$ARB = \left| \frac{1}{K} \sum_{k=1}^{K} \hat{Y}(s_k) - Y \right| / Y$$

and the relative root mean square error

$$RRMSE = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\hat{Y}(s_k) - Y)^2} / Y.$$

Making K=1000 Monte-Carlo simulations we obtained the results given in the Table 1.

Table 1: Comparison of estimators

	Horvitz-Thompson estimator (%)	**GREG estimator LM assisted (%)**	**GREG estimator Tobit assisted (%)**
ARB	0.1394116	2.764075	0.185612
RRMSE	9.2875309	7.200544	7.073001

For the second population *U* consisting of *N*=1000 elements we simulated the values of y_i as $z_i^* = 1 + \omega_i + x_i + u_i + e_i$, where $u_i \sim N(0,1)$, $e_i \sim N(0,0.6)$,

$$z_i = \begin{cases} 1, z_i^* > 0 \\ 0, z_i^* \le 0 \end{cases} \text{ and } \quad y_i = \begin{cases} 1 + x_i + u_i, & z_i^* > 0 \\ 0, & z_i^* \le 0 \end{cases}.$$

After K=1000 Monte-Carlo simulations we obtained the following results.

Table 2: Comparison of estimators

	Horvitz-Thompson estimator (%)	**GREG estimator LM assisted (%)**	**GREG estimator Heckit assisted (%)**
ARB	0.8497651	14.86922	13.33495
RRMSE	27.8450915	23.30287	22.53748

Conclusion

As we can see, usage of GREG estimator leads to biased but better results with regards to the accuracy. The usage of the Tobit and Heckit-based estimators improve the quality of GREG estimator with regard to both bias and mean square error if the underlying processes of zero-values appearing corresponds well with estimator. If it does not correspond the improvement can be lost. And the main useful thing is that all theses GREG estimators can be used for the small area estimation.

References

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics.* Vol.**4**, No.1, 23-39.

Heckman J. (1979) Sample selection bias as a specification error. *Econometrica*, Vol.**7,** No. 1, 153-161.

Humpreys, B.R. (2013). *Dealing With Zeros in Economic Data.* University of Alberta, https://pdfs.semanticscholar.org/35c3/8229c8f7393acffc93b4a83120661df1f02c.pdf .

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24-36.

Correlation between Logistics and Higher education in Latvia

Inguna Jurgelane-Kladava

Riga Tehnical University, e-mail: inguna.jurgelane-kaldava@rtu.lv

Abstract

The role of logistics in Latvian national economy is very important. The fourth part of services of Latvia forms field of logistics and transit. However, the support of government to higher education institutions who provides training programs for potential logistic specialists are minimal or does not exist. The author of the paper analyses is there any correlation between demand of logistics specialists, transit, development of logistics sector and support of the government to the higher education institutions who provides training programs for logistic specialists.
Keywords: logistics, higher education institution, employability

Introduction

To find out tendencies of transport and logistics field, author of the paper analyses statistical data.

Table 1: Freight transport carried by vehicles on roads in Latvia, year 1997–2017 (thsd. t) (author's calculation according data of csb.gov.lv)

Year	Domestic	Growth rate %	International	Growth rate %
1997	23500	X	1669	X
1999	31718	34.97	1683	0.84
2001	29992	-5.44	2307	37.08
2003	38665	28.92	3151	36.58
2005	46633	20.61	4892	55.25
2007	51533	10.51	8372	71.14
2009	31595	-38.69	6225	-25.65
2011	44787	41.75	9149	46.97
2013	50484	12.72	11674	27.60
2015	48517	-3.90	14052	20.37
2017	52189	7.57	15823	12.60

Applying calculation of chain growth rate, it can be seen that, the most significant increase of inland freight was in 2011 (41.75 %), but international in 2005 (55.25 %).

Table 2: Shipped by sea, cargo received at Latvian ports, year 1993–2017 (thsd. t)

Year	Shipped loads	Growth rate %	Cargo received	Growth rate %
1993	25765	X	1642	X
1995	36370	41.16	2615	59.26
1997	46695	28.39	3994	52.73
1999	45145	-3.32	3887	-2.68
2001	54372	20.44	2546	-34.50
2003	50918	-6.35	3837	50.71
2005	55890	9.76	4152	8.21
2007	55178	-1.27	7256	74.76
2009	57565	4.33	4152	-42.78
2011	61028	6.02	7793	87.69
2013	62350	2.17	8130	12.05
2015	62551	0.32	7019	-13.67
2017	54156	-13.42	7721	10.00

According the data in the table, it can be seen that, the most significant increase of shipped goods from ports of Latvia was in 1995 but increase of received cargo in ports of Latvia in 2011.

Table 3: Freight transport by rail in Latvia, year 1993–2017 (thsd. t)

Year	Domestic	Growth rate %	International	Growth rate %
1993	2736	x	27838	x
1995	3545	29.57	25295	-9.13
1997	2522	-28.86	38497	52.19
1999	1938	-23.16	31270	-18.77
2001	2011	3.77	35873	14.72
2003	2329	15.81	46026	28.30
2005	2633	13.05	52228	13.47
2007	2000	-24.04	50164	-3.95
2009	1299	-35.05	52380	4.42
2011	1193	-8.16	58192	11.10
2013	1178	-1.26	54653	-6.08
2015	1671	41.85	53974	-1.24
2017	1649	-1.32	42136	-21.93

Table 4: Data of University of Latvia on the program “E-business and logistics management system” 2010–2018

Study year	Budget places in the 1st year	Applications for budget places	Charge places	Applications for charge places	Part time applications
2010/2011	2	285	300	168	33
2011/2012	2	235	300	156	51
2012/2013	2	248	300	158	41
2013/2014	3	224	300	147	28
2014/2015	4	248	100	140	42
2015/2016	7	252	70	127	34
2016/2017	6	258	70	99	29
2017/2018	4	205	50	98	52
2018/2019	5	202	50	84	23

On average, for one budget place there are 62 applications.

Table 4: Data of Riga Technical University on the program “Business logistics” 2010–2018

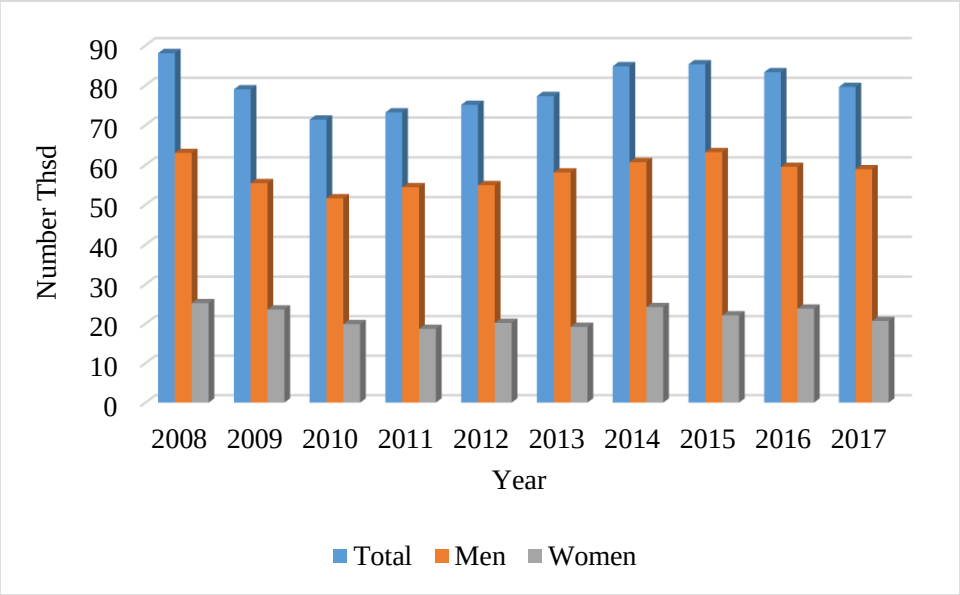
Study year	Budget places in the 1st year	Applications for budget places	Charge places	Applications for charge places	Part time applications
2010/2011	5	429	200	210	19
2011/2012	10	545	200	256	21
2012/2013	6	485	200	276	29
2013/2014	5	378	200	217	22
2014/2015	8	384	*	191	22
2015/2016	10	322	*	145	23
2016/2017	10	339	*	131	17
2017/2018	5	228	*	115	26
2018/2019	4	195	*	99	22

* Unlimited

On arithmetic mean, for one budget place there are 52 applications.

According the data of The State Employment Agency of Latvia (https://cvvp.nva.gov.lv/#/pub/vakances/saraksts#eyJvZmZzZXQiOjI1LCJsaW1pdCI6MjUsInBhZ2VZIjoxMDB9) , for example in 13.07.2018 employers were searching for 200 employees in the field of transportation and logistics. For comparison - on this date there were 25 vacancies in the field of production, 50 in the field of trade and marketing, 50 in the field of security and rescue services and 50 vacancies in the field of IT and telecommunication.

Figure 1: Employed in the transport and storage industry in Latvia 2008–2017 (Number thsd.)



The author's conclusion in the compilation and analysis of all the data that the state support in the preparation of new specialists does not correlate the transport and logistics sector with its development and demand for logistics specialists in Latvia.

References

Homepage of the State Employment Agency of the Republic of Latvia https://cvvp.nva.gov.lv/#/pub/vakances/saraksts#eyJvZmZzZXQiOjI1LCJsaW1pdCI6MjUsInBhZ2VZIjoxMDB9

Central Statistical Bureau of Latvia website, http://data.csb.gov.lv/pxweb/lv/Sociala/Sociala__ikgad__nodarb/?tablelist=true&rxid=, https://www.csb.gov.lv/statistikas-temas/metodologija/kravu-un-pasazieru-parvadajumi-38263.html

Latvija.lv website, https://www.latvija.lv/Epakalpojumi/EP190.

Handling missing data and errors in Estonian eHealth information system

Viktoria Kirpu[1] and Natalja Eigo[2]

[1]National Institute for Health Development of Estonia, e-mail: viktoria.kirpu@tai.ee
[2]National Institute for Health Development of Estonia, e-mail: natalja.eigo@tai.ee

Abstract

An overview about data loss and errors in Estonian eHealth information system is provided in this paper. Besides, some statistical imputation methods are described to solve these problems. At the end, possibilities of using additional information for improving data are discussed.
Keywords: non-response, loss, errors, imputation, additional information, biased assessment, unbiased assessment

1 Introduction

In Estonia health statistics are collected, processed, analysed and published by National Institute for Health Development (Estonian: *Tervise Arengu Instituut*, also called as NIHD). NIHD uses the eHealth information system or eHIS (Estonian: *Tervise infosüsteem*) as one data source for statistics. Unfortunately, the database concerned has deficiencies in the coverage and quality of the data. To validate eHIS data NIHD uses treatment invoices data provided by Estonian Health Insurance Fund (Estonian: *Eesti Haigekassa*, also called as EHIF). At the present EHIF has received more observations than the eHIS.

However if data has not been sent to the database, it is important to take this fact into account when computing statistics and to implement necessary statistical methods. Otherwise, incomplete data may lead to biased estimates that do not correspond to the population.

1.1 eHIS data

The eHealth information system or eHIS was created in 2008 and is managed and developed now by Health and Welfare Information Systems Centre (Estonian: *Tervise- ja Heaolu Infosüsteemide Keskus*, also called as TEHIK). eHIS an important database which is a part of the state health information system (Estonian eHealth Foundation, n.d.). Health care providers oblige to provide epicrisis and other medical documents to eHIS (Riigi Teataja I, 2018). This system's data among other functionalities is used for keeping records of state of health and for producing health statistics (National Institute for Health Development, 2017).

1.2 Estonian Health Insurance Fund data

The most important task of Estonian Health Insurance Fund is to organise health insurance in order to enable health insurance benefits for insured persons. In addition, the task of Health

Insurance Fund is to assist with preparing standards of treatment and treatment guidelines, motivate health care providers to develop quality of health services, organise the performance of international agreements concerning health insurance; participate in planning of health care. (Estonian Health Insurance Fund, n.d.) Estonian Health Insurance Fund also collects documentation about invoices for treatment cases from facilities providing health care services.

2 Problems related to non-response

Non-response occurs in the analysed database when documentation about treatment case has not been submitted to the Health Information System (eHIS).

Lack of data does not only cause a loss of the necessary information and a reduction of the capacity of the study[1], but it causes bias in the estimates assessments[2]. It is crucial to minimise the number of undiscovered lost observations i.e. number of treatment cases, concerning which documentation was not submitted to eHIS and the lack of which was not discovered during checking. Otherwise, statistical conclusions, for example the confidence interval, may be estimated incorrectly. It is necessary to have an unbiased estimate assessment[3] for high-quality statistics or the bias should be reduced as much as possible. The smaller the bias, the better statistical results reflect the actual situation.

For example, the emergency type of a treatment case is the one that usually is not submitted to eHIS by the doctors. If there is a situation, where doctors do not note down emergency treatment cases, then it gives the impression that there are few that kind of treatment cases in the country. In such a situation, we can be certain that the received statistics do not describe the reality and we have received biased estimates of assessments. In other common case biases also arise when whole epicrisis has not even been provided.

There is a strong believing that if the rate of response is high, it is not important to take into account the non-response. Statistics does not focus on the rate of response as an indicator, which reduces the bias caused by non-response, as the rate of response itself does not measure it. Unlike variance, the bias does not near zero when increasing the sample size (Shouten & Cobben, 2007; Särndal & Lundström, 2005). In order to reduce the bias caused by non-response it is vital to use the necessary methods of assessment.

3 Handling a data set without non-response

Let $U = \{1, ..., k, ..., N\}$ be the population of the size N and y_k value of variable Y. Then the total sum of variable Y is: $Y = y_1 + ... + y_N = \sum_{k=1}^{N} y_k$. (Estonian eHealth Foundation, n.d.) In this case, we can get the value of variable Y as the eHIS data set is complete i.e. the patient epicrisis of all treatment cases have been provided to eHIS and there are characteristic values for all observations.

4 Handling a data set with errors and non-response

There is almost always a non-response in empirical data. Data with missing values occures in two different ways: when the observation is missing (unit non-response) or when only part

[1] An indicator (probability), which assesses how important the received result is for statistical purposes (Aron & Aron, 1997).

[2] An assessment, the mean value of which differs from the true value of the assessed parameter by a certain systematic error.

[3] An assessment, the mean value of which equals the true value of the assessed parameter.

of the response is missing (item non-response) (Särndal & Lundström, 2005; Andridge & Little, 2010). In our case, unit non-response means that the patient epicrisis wasn't submitted to eHIS and item non-response means that the patient epicrisis has been submitted with incomplete information. Errors in eHIS can be divided into the following types:

- **random errors** (caused by inaccuracy of measuring or recording, generally have little effect on the result and are difficult to discover), for example the wrong month of birth of the patient;
- **systematic errors** (mainly caused by the inaccuracy of the instrument), for example when the doctor uses the same diagnosis to describe all illnesses of the patients;
- **gross errors** (the value of the characteristic is outside the area of possible values for the characteristic), for example a cervical cancer diagnosis for a male patient;
- **logical errors**, where the values of various characteristics are inconsistent, for example the date of discharging a person from the hospital is marked to be after the date of death.

Upon discovery of errors, they must be eliminated and treated as non-response if necessary.

A data set with non-response cannot be processed in the same way as data with no non-response. The main methods for handling data with non-response is **using additional information** and **imputation**[4].

5 Using additional information

In the case of eHIS data loss, NIHD uses EHIF data as an additional source to improve data.

Three types of additional information are distinguished in case of data sets with loss: *InfoU*, *InfoS*, *InfoUS* (Särndal & Lundström, 2005).

- **InfoU**
  For this type of information the assisting information is vector $x_k = x_k^*$, which is known for each $k \in N$. Additional information, which is the total sums vector $X^* = \sum_U x_k^*$, is known on the level of population U.
- **InfoS**
  For this type of information the assisting information is known within the existing data but not on the level of the population and the assisting vector is $x_k = x_k^\circ$.
- **InfoUS**
  For this type of information $x_k = \begin{pmatrix} x_k^* \\ x_k^\circ \end{pmatrix}$ is known on the level of the population as well as the sample.

Values found from an additional source can be used for the replacement of missing data in eHIS. It helps to reduce bias of estimates.

In order to "enrich" eHIS data NIHD uses data received from EHIF as *InfoU* additional information. NIHD presumes that the data in this certain database has no item non-responses. But not all needed information can be found in this data source. For remaining missing data imputation methods should be applied.

[4] A procedure, where the missing values of one or several variables are replaced by the estimates based on the existing data or some other information.

6 Data imputation

The imputation methods of values are divided into three groups:

- statistical prediction methods;
- getting values from responded similar objects and replacing for those who have not responded;
- expert opinion.

The first two groups are classified as statistical methods. The methods of the first group are based on the relationship between variables such as finding regression models. The methods of the second group can be called donor-based, because the imputed value is borrowed from some observed object, which is similar to the missing object. The methods of the third group largely depend on the expert's skills and knowledge.

Imputation methods can be further classified as deterministic, i.e. when repeating the imputation procedure one always archieve at exactly the same result, or random, where different imputed values are received when repeating the procedure. Regression imputation is an example of a deterministic method. When we however impute the value of a randomly selected similar donor, it is a random method that is often used for Hot-Deck imputation.

It is important to take into account that imputed values are artificial - they are either construed according to some rule or the values of other responded objects. Therefore, imputed values always differ from the actual values of objects to some extent. It is expected that in case of imputation the estimates have a small dispersion and no bias or a small bias.

Generally imputation must be approached with caution as we are trying to produce reliable statistics from data we know is imprecise to a lesser or greater extent from the beginning. At the same time, it is often necessary to do so for practical reasons. There is no good reason for careful imputation to create more damage to assessments than other methods when producing statistics.

7 Summary

In order to produce statistics it is essential to know whether we are using a full dataset to analyze data, or one with losses. Different methods should be implemented for producing statistics according to this prior knowledge. In the case of data without losses, we can produce statistics immediately. In case of data with losses and/or errors, it is necessary to carry out data processing in advance. Various methods are used for this purpose. For example, missing values can be estimated by engaging additional information and imputation. Statistics mistakenly produced from data with losses will result biased estimates. By finding values for missing data, it is possible to reduce the bias or in some cases even eliminate it.

Therefore, in order to produce health statistics the quality of eHIS data must be checked, i.e. whether all necessary documents of treatment cases have been provided to the health information system.

In the current eHIS data quality control NIHD wishes to use data received from EHIF as InfoU additional information. Thanks to this, we would be able to find out how much data is not submitted to the health information system and how many errors are made when submitting data. According to this, it would be possible to use additional information and imputation for supplementing the health information system data set.

References

Andridge, R. R. & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* **78(1)**, 40 — 64.

Aron, A. & Aron, E. N. (1997). *Statistics for the behavioral and social sciences: A brief course*. NJ: Prentice Hall.

Särndal, C. & Lundström, S. (2005). *Estimation in surveys with nonresponse.* John Wiley and Sons: New-York.

Shouten, B. & Cobben, F. (2007). *R-indexes for comparison of different fieldwork strategies and data collection modules*. Voorburg-Heerlen: Statistics Netherland. Discussion paper 07002.

Estonian Health Insurance Fund (n.d.). Haigekassa organisatsioon. Last checked: 06.02.2018. https://www.haigekassa.ee/haigekassa/organisatsioon.

Estonian eHealth Foundation (n.d.). Tervise infosüsteem. Last checked: 06.02.2018. http://www.e-tervis.ee/index.php/et/eesti-etervise-sihtasutus/tervise-infosusteem.

National Institute for Health Development (2017). E-tervise infosüsteem. Last checked: 31.01.2018, http://www.e-tervis.ee/index.php/et/eesti-etervise-sihtasutus/tervise-infosusteem.

Riigi Teataja I (2018). Health Services Organisation Act[1], § 592. Last checked: 09.04.2018. https://www.riigiteataja.ee/akt/TTKS.

Multiple Imputation for Income

Seppo Laaksonen[1]

[1]University of Helsinki, e-mail: Seppo.Laaksonen@Helsinki.Fi

Abstract

Income is a demanding continuous variable in the case of missing data. One reason is that income of the non-respondents is often low although some high income people are not either good respondents. Imputation is in general a better method than weighting to solve the problem. The second point concerning income is that the average is not most interesting but income differences. This paper is focused on both these estimates using international data when we know true values. The missingness is created by the unknown person and hence we do not know its mechanism. The auxiliary variables available are not ideal leading to difficulties when implementing the imputation model. On the other hand, some imputation tasks are not working since they give negative incomes that are not correct at all.
Keywords: model-donor imputation, real-donor imputation, single vs multiple imputation

1 Introduction

Imputation is for replacing missing values with plausible ones. If this procedure has been done once, it is single imputation (SI). SI is a usual tool in statistical offices or other public survey institutes, in particular. However, SI can be performed several times as well. If this procedure is repeated a number of times and 'coordinated' well, the outcome is 'multiple imputation' (MI). What such a good coordination means, it is a special question? Rubin in his books (1987, 2004, 118-119) says that each imputation should be 'proper'. He also gives some rules for proper imputation but they are not necessarily easy to follow, or their implementation is not automatic. A big question here is how to repeat the imputation process well, that is, what is an appropriate Monte Carlo technique in order to get $L>1$ simulated versions for missing values?

Rubin (1996, 476, 2004, 75&77) also says that a theoretically fundamental form of MI is repeated imputation. Repeated imputations are draws from the posterior predictive distribution under a specific model that is a particular Bayesian model both for the data and the missing-data mechanism.

Several proper MI implementations are given in Rubin's books and in software packages (e.g. SAS and SPSS) using his book. He thus recommends that imputations should be created through a Bayesian process as follows: (i) specify a parametric model for the complete data, (ii) apply a prior distribution to the unknown model parameters, and (iii)

simulate *L* independent draws from the conditional distribution of the missing data given the observed data by Bayes' Theorem.

These Rubin's theoretical principles are one starting point of this paper. A good point is that MI is not difficult to apply since most types of estimates can be computed in a usual way (e.g. averages, quantiles, standard deviations and regression coefficients). The Rubin's framework also serves the formulas both for point estimates and for interval estimates. The point estimates are simply averages of *L* repeated complete-data estimates, and thus very logical. His interval estimates are not indisputably accepted. Björnstad (2007) gives a modified version for the second component of Rubin's formula. This leads to a larger confidence interval, as a function of the rate of imputed values. This is logical since Rubin's formula is without any explicit term of the imputation amount but his Bayesian rules might implicitly include the same; this is however difficult to recognize.

Björnstad (2007, 433) also invents a new term, non-Bayesian MI, since his imputation is not following a Bayesian process. This term 'non-Bayesian' is not used in ordinary imputation literature; it cannot be found 9 years alter from a book by Carpenter and Kenward (2013) that much follows Rubin's framework but they use the term 'frequentist'. We still use the term 'non-Bayesian,' since we cannot say whether it is equal to 'frequentist.'

Björnstad motivates his approach also from the practical points of view saying that in national statistical institutes the methods used for imputing for nonresponse very seldom if ever satisfy the requirement of being "proper." Moreover, Muñoz and Rueda (2009) say that several statistical agencies seem to prefer single imputation, mainly due to operational difficulties in maintaining multiple complete data sets, especially in large-scale surveys. We agree with these views. Since a non-Bayesian approach also leads to single imputation, that is commonly used if anything has been imputed, a conclusion could be that MI cannot be applied using a non-Bayesian framework. We do not agree with this argument. Consequently, we have over years applied non-Bayesian tools both for single and multiple imputation, although most often for single imputation. This paper first summaries our approach to imputation.

This approach first makes attempts to impute the missing values once. That is, the focus is first on single imputation. Correspondingly, the main target in imputations is to succeed in such estimates that are most important in each case. Since it is hard to impute correctly individual values, it is more relevant to try to get least unbiased estimates for some key estimates. Since we here concentrate on a continuous variable, that is, income, two types of estimates are of a special importance. One is income average and the other is income distribution, respectively. Income distribution can be measured by various indicators such as quantiles or Gini coefficient, but the coefficient of variation is here considered to be simple enough to indicate well income differences between people.

Rubin's approach can be implemented in various ways. We do not develop any own implementation but take advantage of the two existing implementations. These are derived from two general software packages, SAS and SPSS. We assume that their MI procedures follow a Bayesian process since there are such references in their manuals.

We thus use the term ‘Bayesian MI’ for application of SAS and SPSS. Respectively, our own imputation framework is called ‘Non-Bayesian MI.’

2 Imputation framework

In order to succeed in imputation, good auxiliary data or covariates are needed. In the case of lacking covariates, simple methods based on observed values only can be applied. But if there are covariates both for the respondents and for the non-respondents, ‘proper’ imputation methods can be used. In this case, the imputation framework (cf. Laaksonen 2016) includes the two core stages:

(i) Construction and implementing of *the imputation model*

(ii) Imputation itself or *imputation task*.

These two terms are also used by Rubin (2004) but these are integrated well together in our framework. An imputation model can be implemented using a smart knowledge of the imputation team or it can be estimated from the same data set or from a similar data set from an earlier survey or a parallel survey of another population. If the model is estimated from the same data set, it is expected that this replacer behaves more surely well in imputations. Hence we estimate the parameters of the imputation model from the same data set.

There are the two alternatives as a dependent variable in an imputation model. It is either (a) ‘*the variable being imputed’* or (b) ‘*the binary response indicator of the variable being imputed*.’ The same auxiliary variables can be used in both models. Naturally, the estimations that are needed in the next step are derived from the different data sets, from the respondents for the model (a) and from both the respondents and the non-respondents for the model (b). The covariates need to be completely observed to compute the predicted values for the stage (ii).

The imputed values themselves can also be determined by the two options: (i) they are calculated using the imputation model or (ii) they are borrowed from the units with the observed values using the imputation model as well. The previous option is called ‘*model-donor*’ imputation, and the second is ‘*real-donor’* imputation, respectively. The latter one is often called ‘hot deck’ but this term is not clear in all cases. Terms for the previous ones are often such that the model and the task are confused. For example, model imputation or regression imputation is not clear since these are referring to imputation model but the second step, imputation task, is not specified.

If a real-donor method is applied, an appropriate criterion and a valid technology to select a donor is needed. The natural criterion is to select an as a similar real-donor (observed value) as possible. This may be based on a kind of nearness metrics. If a clear criterion exists, it is good to select the nearest or another from the neighborhood. If any valid criterion does not exist, a random selection from the neighborhood can be used. This thus means that all units with observations are as close to each other within the neighborhood that can be called ‘an imputation cell,’

In our approach, the predicted values of either the model (a) or the model (b) are used as the nearness metrics, leading to real-donor methods. We focus on multiple imputation and hence we impute everything 10 times and calculate their average as the point estimate. The variance estimate is the sum of the between variance and the within variance. Rubin's formula does not include the response rate meaning the variance is smaller than in the case of Björnstad's formula.

Our framework thus is non-Bayesian and so we simply add the noise term to the predicted values. We test two types of the noise term using random numbers: (i) normally distributed residuals, (ii) normally distributed standard errors. We test several imputation models: (i) linear regression, (ii) log-linear regression, (iii) logistic regression, (iv) probit regression, (v) log-log regression (LL), (vi) complementary log-log regression (CLL).

SPSS and SAS use their methods and we simply apply them but we test two imputation models: (i) linear regression, and (ii) log-linear egression. They thus are Bayesian.

3 Empirical examples

The number of missing values or the imputation size is 3133 (out of 10000)) that is fairly realistic. The data set consists of a quite good number of covariates which all except age are categorical. The age was however categorized. The full list with the number of categories that is used in all imputation models is as follows: gender (2), five-year age group (11), marriage (2), civil status (2), education level (4), region (12), Internet at home or not (2), socio-economic status (4), unemployed or not (2), children or not (2). As seen any of these covariates is not well predicting yearly income (R-square of the linear regression model is about 40%).

Model-donor methods

The linear regression model is easy to apply for model-donor imputation but it does not give excellent results due to many negative values. Table 1 gives the results.

Table 1. Negative values of model-donor methods (NB = Non-Bayesian, B = Bayesian)

Method	Negative values, %
Using residuals NB	8.5
Using standard errors NB	0.3
SPSS B	16.8
SAS B	16.6

We find that all methods give negative values but Bayesian methods much more. Hence we do not use more model-donor methods but go to real-donor methods. We have explained already the basics of non-Bayesian methods but do not go details as far as Bayesian methods are concerned. Both SPSS and SAS have the method called 'Predictive mean matching methods' that always give observed values, thus not negative.

Real-donor methods

Table 2 presents the results. They are ordered by the imputation model applied. The last four methods are for binary regressions where are symmetric (probit, logit) and asymmetric link functions (CLL and LL). We find that log-linear regression is worst but it is not easy to know the reason. All imputed averages seem to be too big but the CV's almost always too small. Some imputation methods are however fairly good as far as income differences are concerned. One general conclusion could be that the imputations are leading to reduce the bias but not enough, concerning averages especially.

Table 2. Averages and coefficients of variation of yearly income and standard errors by Rubin and Björnstad

			Ranking		Mean ranking	Standard error of the mean	
Method	**Average**	**CV**	**Average**	**CV**		**Rubin**	**Björnstad**
Linear regression NB	46178	66.3	8	7	7.5	692	729
Linear regression SAS	45121	68.0	3	3	3	896	1017
Linear regression SPSS	45471	66.2	5	8	6.5	710	757
Log regression NB	46722	65.2	10	10	10	772	846
Log regression SAS	46034	66.7	7	5	6	864	973
Log regression SPSS	46179	66.1	9	9	9	692	728
Logit regression NB	45468	67.7	4	1	2.5	845	950
Probit regression NB	44785	67.9	1	2	1.5	754	822
CLL regression NB	44898	67.3	2	4	3	915	1047
LL regression NB	45493	66.4	6	6	6	864	976
True value	43531	67.7					

The major part of the standard error is derived from the within variance (from 59% to 79%). This is one reason that the differences between Rubin's and Björnstad's standard errors respectively are not big. They vary fairly much by methods. If the standard error is big, it is easier to get the result that covers the true value. On the other hand, a small standard error is often good. The reader can make his/her interpretation what method is best and which standard error formula. I prefer the probit regression NB.

References

Björnstad, J. (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics*, 433–452.

Carpenter, J. and Kenward, M. (2013): *Multiple Imputation and its Application*. Wiley & Sons

Laaksonen, S. (2016). A new framework for multiple imputation and applications to a binary variable. *Model Assisted Statistics and Applications*, 11.3, IOS Press. http://content.iospress.com/journals/model-assisted-statistics-and-applications/11/2

Muñoz, J.F. and Rueda, M.M. (2009). New imputation methods for missing dada using quantiles. *Journal of Computational and Applied Mathematics* 232, 305-317.

Rubin, D. (1987/2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library Edition.

Rubin, D. (1996). Multiple Imputation After 18+ Years. *Journal of American Statistical Association*, 473-489.

Population Estimation Beyond 2021

Mārtiņš Liberts[1]

[1]Central Statistical Bureau of Latvia, e-mail: martins.liberts@csb.gov.lv

Abstract

Population statistics in Latvia are produced using register/model based methodology since 2012. Precision evaluation of register/model based statistics is an ongoing process. The paper summarise the activities done for precision evaluation so far. The current register/model based methodology has worked so far and we plan to use the same methodology for Census 2021. However, a long-term aim is to develop an alternative methodology for population statistics.
Keywords: Population estimation, Population census, register/model based statistics

1 Introduction

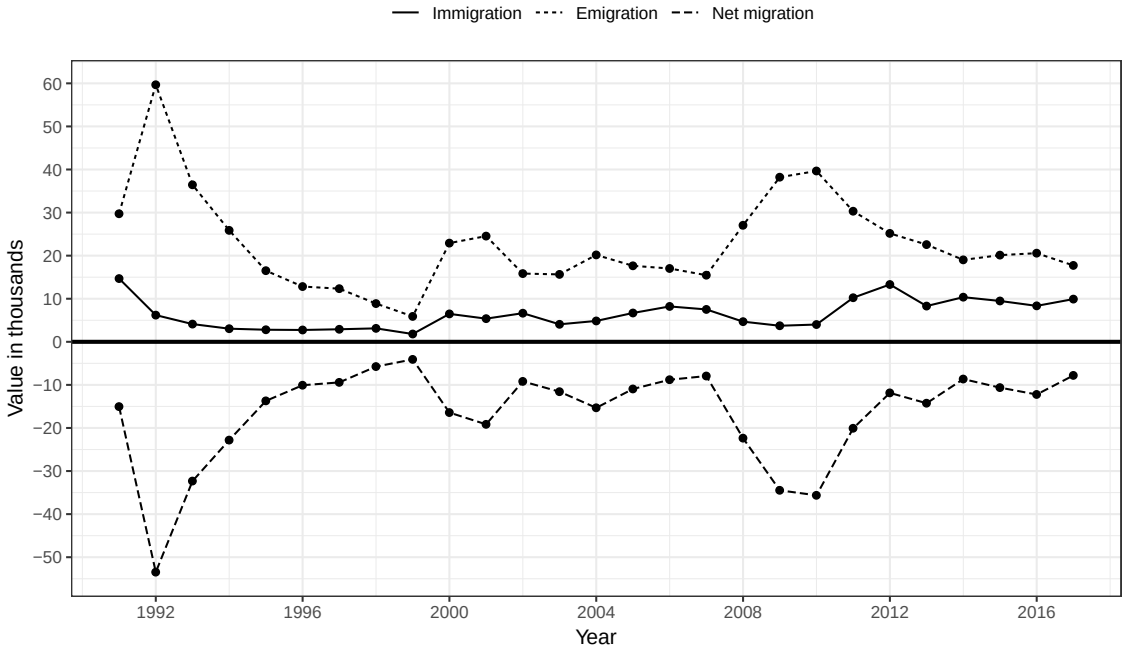
The population statistics of Latvia are produced using register/model based methodology since 2012 (Vaļkovska, 2012; Vaļkovska *et al.*, 2014; Aināre *et al.*, 2017). Even though there is Population Register in Latvia, it is not possible to produce population statistics using register based methodology only. The main reason being the significant over–coverage of the resident population in the register data.

The difference between the register population and the census population was 155 thousand (7 % from the census population) in 2011 (Aināre *et al.*, 2017, p. 1). The cause of this difference is twofold. Firstly there is still important emigration flow from Latvia ongoing with recent peak value in 2009–2010 when emigration reached almost 40 thousand emigrants per year (see Figure 1) which is 2 % from the total population. Secondly there is lack of incentives for emigrants to provide correct information to the Population Register about the place of residence. As the result for many emigrants the declared country of residence is still Latvia.

We can describe the problem as a statistical classification problem. The task is to classify all individuals from the register population into two classes – *de facto* residents and *de facto* non–residents of Latvia. We have solved the problem using a logistic regression model. The dependent variable in the model is a binary variable denoting *de facto* residents with `1` and *de facto* non–residents with `0` and independent variables are different binary variables describing individuals. The model is estimated using 2011 data, where dependent variable is taken from Census 2011 and independent variables are created from several administrative data sources corresponding as close as possible to the year 2011. Each following year the same set of independent variables is created and the model is applied to classify all individuals from the register population into residents and non–residents (Aināre *et al.*, 2017).

The estimation methodology including the model was developed during 2011–2013. The first estimates using the new methodology were published in mid 2013 (the results for 2012 were revised according to the new methodology). The original model with some adjustments (by adding or removing some dependent variables) has been used since then.

Figure 1: Long-term migration



Source: Central Statistical Bureau of Latvia

2 Precision evaluation

Precision evaluation of register/model based statistics is not straightforward. Usually some external data are necessary for the precision evaluation. Currently three approaches are used for the precision evaluation:

- Micro data of population statistics are linked with survey data or other administrative data not used for modelling. It is possible to detect persons which are residents of Latvia but which we have not included in the frame for population statistics. This approach is applied yearly. The range of errors is 1 % to 2 %. See Aināre *et al.* (2017) for more details. Unfortunately we can detect only one side of errors using this approach. We can not detect persons which we have included in the frame for population statistics but who are not resident of Latvia using this approach.
- Methodology to estimate model errors is under development now. The task is to estimate variability of model estimates (assuming the model coefficients are estimated). Bootstrap methodology is being applied for this task.
- We have organised two large scale sample survey – Micro census 2015 and International migration survey 2017–2018. The main aim of those surveys is precision evaluation of the population statistics.

2.1 Micro census 2015

Micro census was organised in 2015. The aim of Micro census was the precision evaluation of population statistics. Micro census was done as an independent sample survey. The target population of Micro census was all private dwellings of Latvia. The sample size was 14,996 dwellings. Two-stage cluster sampling design was used (Liberts, 2017).

Table 1: Estimates of population – total and by gender

Gender	$\hat{\theta}_P$	$\hat{\theta}_M$	$\hat{\theta}_P - \hat{\theta}_M$	$\hat{\text{me}}\left(\hat{\theta}_M\right)$	p–value
Total	1 949 510	1 912 299	37 211	28 837	**0.011**
Males	892 394	864 970	27 424	16 500	**0.001**
Females	1 057 116	1 047 328	9 788	19 211	0.318

The main aim of the data collection process was to list all residents of sampled dwelling. Gender and age of residents was recorded. This information allows to get estimates of resident population size in breakdown by gender, age, and regions.

The total over–coverage rate (weighted) was 21.1 %. It is quite high if compared with other sample surveys were dwelling sample is used. This was expected because the population frame was created including completely all private dwellings from different data sources. This was done to reduce under–coverage risk as much as possible. Most of over–coverage cases were because of unoccupied dwellings (72.8 %). The potentially unoccupied dwellings are excluded usually from the population frame for other surveys.

The total response rate (weighted) was 93.5 %. This is very high if compared to other usual surveys. It was possible to achieve so high response rate because of two reasons: questionnaire was very short and proxy interviews (with neighbours or local municipality) were allowed. So we can hope to have potentially low non–response bias.

The results of Micro census were compared with population statistics (excluding population of institutional dwellings). Since Micro census was carried out as a sample survey – the results of Micro census have sampling errors. It was taken into account when comparing the population statistics and the results of Micro census. Comparison was made with the help of hypothesis testing.

Analysing the total population, we can conclude, that the difference between the population statistics and Micro census of the population is 37 thousand (1.9 %), which is statistically significant, because the margin of error is 29 thousand (relative margin of error is 1.5 %). Micro census indicates that the total population is overestimated. The analysis of the results split by gender shows that the total number of men in the population statistics is overestimated. The estimates of number of women do not have statistically significant difference. See Table 1 where $\hat{\theta}_P$ is the estimate of a population parameter using the current methodology, $\hat{\theta}_M$ is the estimate of a population parameter using Micro census data, $\hat{\text{me}}\left(\hat{\theta}_M\right)$ is the estimate of margin of error for $\hat{\theta}_M$, and "p–value" is p–value from hypothesis testing (equality of $\hat{\theta}_P$ and θ is tested assuming $\hat{\theta}_M$ is an unbiased estimate of θ).

Micro census results were rated as very valuable source of information for precision evaluation of population statistics. Some significant differences have been found between the current population statistics and the estimates from Micro census, however most of the differences are explainable and understandable. The results of this evaluation task show the direction of necessary improvements for the current methodology.

2.2 International migration survey 2017–2018

There was trial to estimate long-term international migration flows using Micro census. Unfortunately this trial was not successful. The main reason of failure was measurement errors. Micro census was done as one-wave survey. The field work of Micro census was organised during the 4th quarter of 2015. Respondents were asked to list residents of a sampled dwelling at three time points: 2015-01-01, 2015-09-01, and 2016-01-01. The listing of residents on 2015-09-01 was used for the population estimates as this listing was the closest

to the fieldwork period.

The listings of residents on 2015-01-01 (*past*) and 2016-01-01 (*future*) were used for migration estimation. Unfortunately the data collected about those time points were influenced by measurement errors. It was not possible to use Micro census data for reliable migration estimates.

Decision was made to organise International migration survey as a two-wave sample survey. The sample of 20,000 dwellings was drawn. The survey strategy is to monitor the sample dwellings in two time points. The task is to list the residents of sampled dwellings on two time points, namely 2017-12-01 and 2018-10-01. It will be possible to estimate international migration by comparing those lists (birth, death and internal migration should be excluded).

The data collection for the 1st wave has been done already. It was done from December 2017 till March 2018. The data processing is in process now. The data collection for the 2nd wave will be done at the last quarter of 2018.

3 Census 2021 and beyond

The current register/model based methodology for the estimation of population statistics has worked quite well. We have made precision evaluation of population statistics using Micro census in 2015. Precision evaluation of international migration statistics will be done using the results of International migration survey 2017–2018. The current plan is to use the same methodology also for Census 2021. So Census 2021 will be done as register/model based census in Latvia. However we have observed some drawback of the current methodology.

Firstly, the model used for population classification has been estimated using the data from Census 2011. So the question is – how long we can use this model? How to detect a time point when model fails to predict the current population? We do not have answers for those questions. But it is clear that it will not be possible to use the current model forever.

Secondly, the classification model works quite well for population size estimates. Unfortunately it fails to get good international migration estimates directly. The solution is to use external migration data and to estimate total emigration separately. Finally the results of model (probabilities) are adjusted to be in line with the external migration estimates. See Ainārе *et al.* (2017) for more details.

The long-term aim is to develop different methodology which would deal with those two drawbacks mentioned. There have been some attempts to achieve this aim.

We have tried to replace Census 2011 data with the data from a recent large scale sample survey data (for example, Labour Force Survey). This would allow to estimate the model using more recent data.

Another attempt was to replace supervise classification model (logistic regression) with unsupervised classification methods (for example, clustering). In this case it would be possible to discard Census 2011 data from the estimation phase.

Unfortunately none of those attempts have resulted with something reasonable. Work is in progress.

4 Conclusions

Population statistics in Latvia are produced using register/model based methodology since 2012. The same methodology will be used for Census 2021. The precision evaluation of register/model based statistics is not straightforward. Several approaches has been used for precision estimation. The current methodology has worked so far. However it is clear that

we will need to develop an alternative methodology. The main reason being that the current model is estimated using Census 2011 data. Census 2011 data becomes more and more outdated by time.

References

Ainãre, I., Liberts, M., Zukula, B., Šulca, S., Vaļkovska, J., Opermanis, B., Jurševskis, A., Lece, K. & Ceriņa, A. (2017). Method used to produce population statistics. Methodological report, Central Statistical Bureau of Latvia, Riga, Latvia. https://www.csb.gov.lv/sites/default/files/data/EN/demstat_metodologija_eng.pdf.

Liberts, M. (2017). Methodology and results of the micro census in Latvia. In *Baltic-Nordic-Ukrainian workshop on survey statistics theory and methodology*. Statistics Lithuania, Vilnius, Lithuania, pp. 58–63. http://vilniusworkshop2017.vgtu.lt/wp-content/uploads/2017/07/Workshop-Proceedings-2017.pdf.

Vaļkovska, J. (2012). The number of Latvian residents estimation via logistic regression. In *Workshop of Baltic-Nordic-Ukrainian network on survey statistics*. Central Statistical Bureau of Latvia, University of Latvia, Riga, Latvia, pp. 198–202. http://home.lu.lv/~pm90015/workshop2012/papers/W2012_Poster_VALKOVSKA_JELENA.pdf.

Vaļkovska, J., Liberts, M. & Zukula, B. (2014). The estimation of population in Latvia. In *Workshop of BNU network on survey statistics*. Statistics Estonia, Tallinn, Estonia. https://www.stat.ee/public/yritused/BNU/Valkovska_Liberts_Zukula.pdf.

Coherence studies in time series

Vytautas Pankūnas[1], Julija Janeiko[2] and Danutė Krapavickaitė[3]

[1]Vilnius Gediminas Technical University, e-mail: vytautas.pankunas@stud.vgtu.lt
[2]Vilnius Gediminas Technical University, e-mail: julija.janeiko@stud.vgtu.lt
[3]Vilnius Gediminas Technical University, e-mail: danute.krapavickaite@vgtu.lt

Abstract

The aim of this paper is to present a way to measure strength of a relationship between the two time series by a coefficient of coherence. A definition of the coherence coefficient is given and an example of its application is provided.

1 Introduction

A dictionary says that coherence means integration of diverse elements, relationships, or values. One of the principles in official statistics is coherence of statistical information in the sense of possibility to combine it together. A measure of coherence in official statistics has to show a degree to which statistical results arising from different statistical processes can be combined together. In the case of the time series coherence shows the degree to which different time series reflect the same phenomenon in economy.

2 A concept of coherence

A concept of coherence may be met in the different fields of science, like physics, geophysics, classical time series and elsewhere.

We follow the definition of the coherence as it is presented in Stoffer & Shumway (2006) and Wei (2006). Let we have two finite stationary time series $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$. Define their cross-covariance function

$$\gamma_{xy}(h) = \mathrm{E}\big((x_{t+h} - \mu_x)(y_t - \mu_y)\big) \quad \text{estimated by} \quad \frac{1}{n}\sum_{t=1}^{n}\big((x_{t+h} - \bar{x})(y_t - \bar{y})\big),$$

$\mu_x = Ex_t$, $\mu_y = Ey_t$, $\bar{x} = 1/n\sum_{i=1}^{n} x_i$, $\bar{y} = 1/n\sum_{i=1}^{n} y_i$, $t = 1, 2, ..., n$, $h = 0, 1, 2, ..., n$, $x_{t+n} = x_t$. The follwing representation is applied to the cross-covariance function:

$$\gamma_{xy}(h) = \sum_{k=-m_h}^{[n/2]} c_k \mathrm{e}^{i\omega_k t}, \quad m_h = \begin{cases} [n/2], & n \text{ is odd}, \\ [n/2]+1, & n \text{ is even} \end{cases}$$

with the frequencies $\omega_k = 2\pi k/n$. This representation is called Fourier representation, and frequencies $\omega_k = 2\pi k/n$ are called Fourier frequencies. The Fourier coefficients c_k (cross-spectrum) are given by the formula

$$c_k = \frac{1}{n}\sum_{h=1}^{n}\gamma_{xy}(h)\mathrm{e}^{-i\omega_k t}, \quad k = -m_h, -m_h+1, ..., 0, 1, ..., [n/2].$$

The cross-spectrum is generally a complex-valued function.

The energy associated with the cross-covariance sequence $\gamma_{xy}(h)$ is defined by $\sum_{h=1}^{n} \gamma_{xy}(h)$. The energy of $\gamma_{xy}(h)$ per unit time is called the power of the sequence:

$$Power = \frac{1}{n} \sum_{h=1}^{n} \gamma_{xy}(h) = \sum_{k=-[m_h]}^{[n/2]} c_k^2.$$

The quantities $f_0 = c_0^2$, $f_{[n/2]} = |c_{[n/2]}|^2$, $f_k = |c_{-k}|^2 + |c_k|^2 = 2|c_k|^2$, $k = 1, 2, ..., [n/2 - 1]$, which are obtained from the cross-covariance function $\gamma_{xy}(h)$ Fourier representation at the k-th frequency $\omega_k = 2\pi k/n$, are interpreted as the contribution of this frequency to the total power. The quantity f_k plotted as a function of ω_k is called a periodogram.

An important example of the application of the cross-spectrum is the problem of predicting an output series y_t from some input series x_t through a linear filter relation. A measure of strength of such a relation is the squared coherence function (or coherence coefficient), defined as

$$\mathrm{coh}_{xy}^2(\omega_k) = \frac{|f_{xy}(\omega_k)|^2}{f_{xx}(\omega_k) f_{yy}(\omega_k)} \tag{1}$$

where $f_{xx}(\omega_k)$ and $f_{yy}(\omega_k)$ are the individual spectra of the series x_t and y_t, respectively; $f_{xy}(\omega_k) = f_k$.

Another fact worth mentioning is that squared coherence coefficient is related with the conventional squared Pearson correlation coefficient in the form

$$\rho_{xy}^2 = \frac{\sigma_{yx}^2}{\sigma_x^2 \sigma_y^2}, \tag{2}$$

where σ_x^2 and σ_y^2 are variances of random variables x and y and $\sigma_{yx} = \sigma_{xy}$ is their covariance.

3 A coherence coefficient in practice

Data sets of quarterly aggregated statistics from different statistical surveys and administrative data sources in 2008-2017 are used for a case study. They are presented in Figure 1.

- Statistics Lithuania, Labor Force Survey data: variable for a number of employed (LFE); a number of unemployed (LFU) (in thousands).
- Statistics Lithuania, Labour remuneration survey data: number of employees (Emp); resource for remuneration (RRS) (in millions of Euros).
- Labour Exchange office data: number for registered unemployment (EU) (in thousands).
- Administrative data of the Social insurance institution Sodra: enterprise remuneration, from which taxes are payed (RSI) (in millions of Euros).

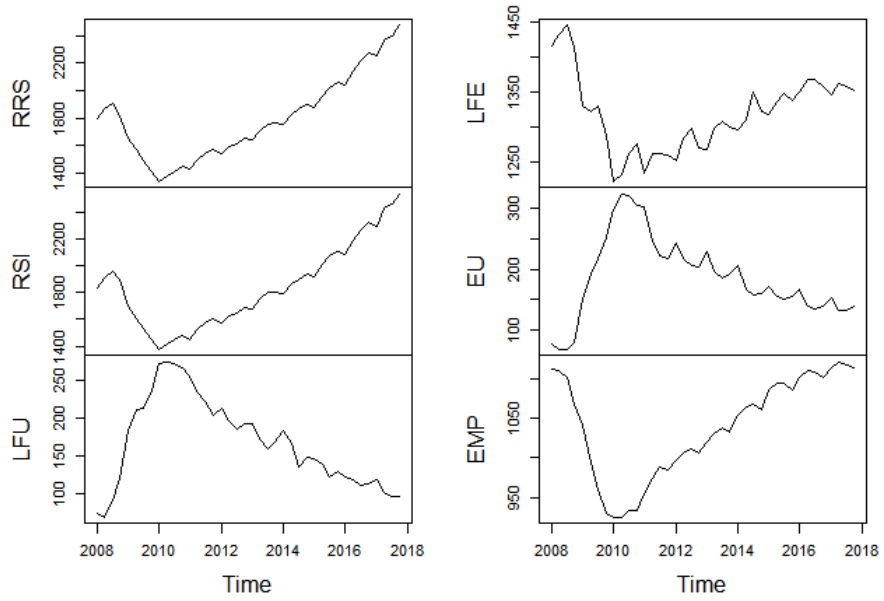


Figure 1: Graphical view of the quarterly data in 2008-2017

The common tendencies of change are observed despite the time series are generated by the different statistical processes. Some periodic fluctuations, possibly yearly, come to notice.

The coherence coefficients between the estimate for a number of unemployed LFU and enterprise remuneration RSI by Sodra are calculated for frequencies $\omega = 2\pi k/n$, $k = 1, 2, ..., [n/2]$. We consider this to be the best example to illustrate coherence as these two time series are generated by the different processes and they behave completely contrariwise: when LFU increases, RSI decreases and vice versa.

Before approaching to the coherence of RSI and LFU, let us draw the periodograms (Figure 2) for each *differentiated* series separately to find out which frequencies $\omega = \omega_k/2\pi$ have the highest contribution to the power of differentiated time series $\gamma_{xy}(h)$. The `R` package *astsa* is used to draw the periodograms and to calculate the coherence coefficients.

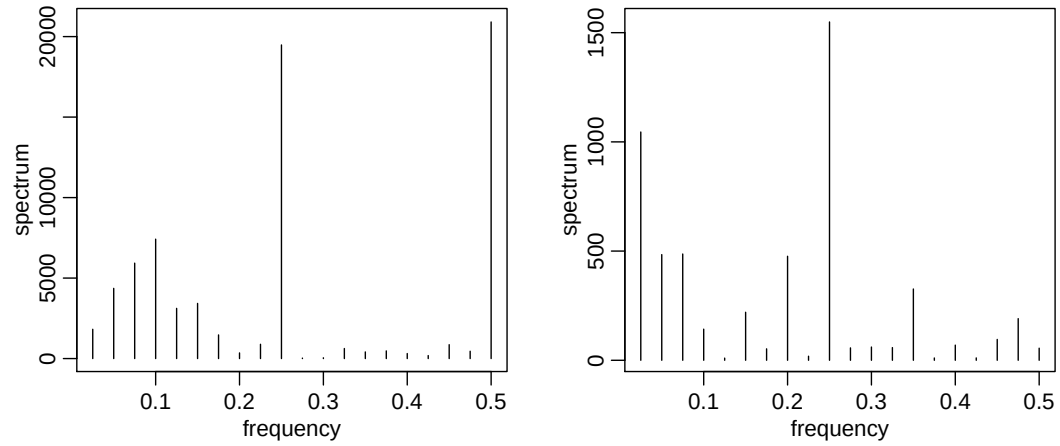


Figure 2: Periodograms of LFU (left) and RSI (right)

In the context of the study data the frequency $\omega = 0.25 = \omega_{10}/2\pi$ with the corresponding period $T = 1/\omega = 4$ – four quarters is the most important. We

pay attention to the frequencies $\omega = 0.025 = \omega_1/2\pi$ (period 40 quarters) and $\omega = 0.125 = \omega_5/2\pi$ (period 8 quarters) as well.

Now we can calculate and portray the coherence between LFU and RSI in Figure 3.

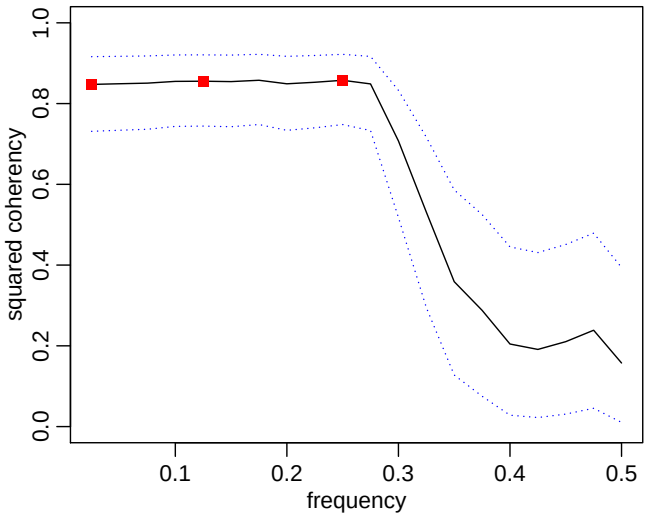


Figure 3: Coherence between LFU and RSI

The numeric values of the coherence coefficients and squared correlation coefficient between LFU and RSI are presented in Table 1.

Table 1: Coherence coefficients and squared correlation coefficient

ω	$\mathrm{coh}^2_{xy}(\omega)$
0.025	0.8475
0.125	0.8554
0.25	0.8577
$\rho(x, y)$	-0.8624

4 Conclusion

The simulation study shows that both kinds of indicators: squared correlation coefficient (2) and coherence coefficients (1) suit well for assessment of the linear dependency or similarity between the time series. At the moment it is hard to say anything about the preferences of some of them.

References

Shumway R. H. and Stoffer D. S. (2006), *Time Series Analysis and Its Applications*, Springer Texts in Statistics, DOI = 10.1007/978-3-319-52452-8.

Stoffer, D. (2017), *astsa: Applied Statistical Time Series Analysis*, R package version 1.8, `https://CRAN.R-project.org/package=astsa`.

Wei W. W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Pearson Education, `https://books.google.lt/books?id=aY0QAQAAIAAJ`.

Case study: The effect of text message reminder on survey nonresponse

Oona Pentala-Nikulainen[12]

[1]Helsinki University
[2]National Institute for Health and Welfare Finland, e-mail: oona.pentala-nikulainen@thl.fi

Abstract

In 2017 the National Institute for Health and Welfare Finland started the National FinSote Survey which enables monitoring the changes occurring in the population's well-being and health by different population groups and regions. It is a mail and online survey where we also tested some new ways of approach respondents in order to achieve a better response rate and to find methods for applying adaptive data collection methods in the future data collection rounds. This paper is a small case study of how sending text message reminders affected the response rate in the test group of 20 to 54 year old respondents. The study shows that sending a text message reminder was effective among some age groups but it was not a very cost-effective procedure considering the whole age group.

Keywords: BNU2018, nonresponse, response rate, approaching respondents, adaptive data collection

1 Introduction

The National FinSote Survey is a study of health, well-being and service use among Finnish adult population. It enables monitoring the changes occurring in the population's well-being and health by different population groups and regions. The study also produces follow-up and evaluation data on how well the service needs of the population are met as well as the views of the population on the social and health care service system, and the availability, quality and use of services.

The FinSote Survey was conducted for the first time in the fall of 2017 and the data collection period ended in April 2018. The sample size was 59 400 of adults 20 years old and upwards. The final response rate was 46% with big differences between different age groups; 20-to-54-year-olds 28%, 55-to-74-year-olds 58% and 75 years and older 57% (Pentala-Nikulainen *et al.*, 2018).

Information was collected by mail and online questionnaire. The questionnaire form was available in Finnish, Swedish, Russian and English. The participants were contacted in total 4 times by sending them a paper form or a reminder letter. With every contact there were also instructions on how to respond online.

The FinSote Survey also works as one of institute's the pilot surveys in testing new methods for reaching participants in population surveys. In the future we are interested in applying adaptive data collection methods in survey settings and are currently experimenting different ways of activating especially young participants to take part in mail-based questionnaire surveys. This case study is part of that experiment. Reminding people about the importance of their participation in a survey can increase response rates, minimise bias in the data, and reduce the need to approach an excessive number of business (Dillman, 2007).

2 Methods

During the data collection we tested sending a reminder via text message to a test group of 7991 20-to-54-year old nonrespondents. The text reminder was an extra contact between the second and the third contact time approximately 4 months after the first contact. We will later on refer to this group as a test group. The control group consisted of 9585 20-to-54-year old nonrespondents for whom it was not possible to obtain an up-to-date mobile phone number. The reasons for the phone number not being available were usually the use of pre-paid subscriptions or having a mobile phone subscription under a new address or different name. The test group and the control group differed very little age and gender wise; in the test group there were slightly more older nonrespondents than in the control group.

The text message consisted of two parts. The first part was
"Hi! You received an invitation to take part in FinSote Survey earlier by mail. Please respond at www.thl.fi/finsote/answer. Your respondent code is xxxx. You will receive the password in a separate message."
and the second part was
"Your password concerning FinSote Survey is xxxx. Thank You in advance! More information at www.finsote.fi or p. 0800 97730. Please do not respond to this text."

The text messages were sent by programmed interface and we had a confirmation of each text which was sent to the participants. All the 15892 text were successfully delivered to the participants.

3 Results

The text messages did not result in higher number of contacts or refusals. There was only one reported case where the person who informed us of his refusal mentioned the text message as a reason so we assume that it was a suitable way to contact the test group.

When comparing the test and control groups' overall actualised response rates, it seems that the test group had a little higher response rate than the control group (12.7% vs 10.7%). However the effect was statistically significant only in some age groups, 25-to-29-year-olds and 45-to-49-year olds (figure 1 and table 1). There was also a difference between genders; the effect of text reminder was significant only among men (table 1).

Figure 1: Actualised response rates of the control and test groups by age

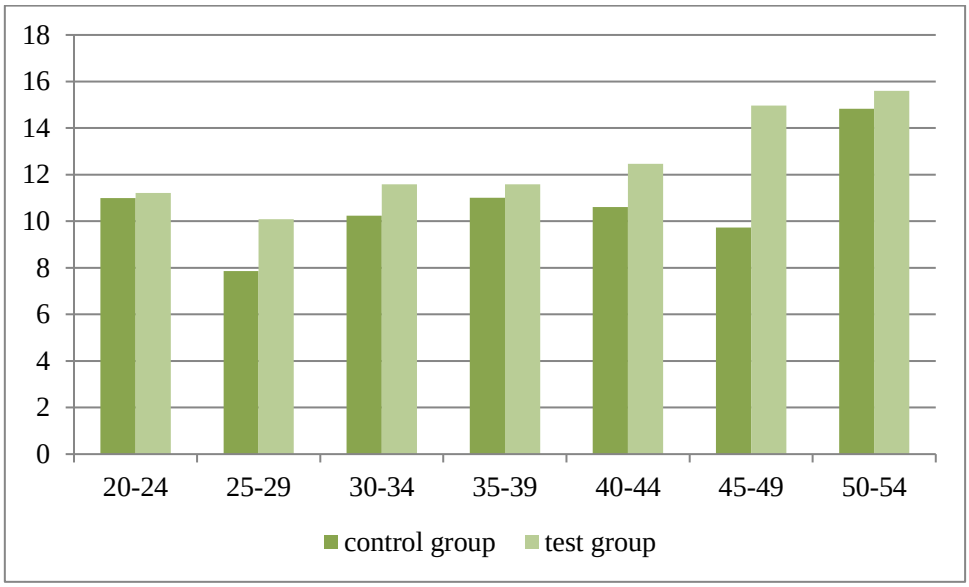


Table 1: Actualised response rates of the control and test groups by gender and age and p value for the difference between control and test group

Group	Response rate	P value
Men		
control	8.1	
test	11.2	<0.0001
Women		
control	13.5	
test	14.6	0.14
age group 20-24		
control	11.0	
test	11.2	0.87
age group 25-29		
control	7.9	
test	10.1	0.05
age group 30-34		
control	10.2	
test	11.6	0.027
age group 35-39		
control	11.0	
test	11.6	0.65
age group 40-44		
control	10.6	
test	12.5	0.16
age group 45-49		
control	9.7	
test	15.0	0.0001
age group 50-54		
control	14.8	
test	15.6	0.6

4 Discussion and conclusions

The text message reminder had a positive effect on response rates especially among men and 25-to-29-year olds and 45-to-49-year olds. Some of the effect in the older age group can be explained by the slight differences in the age distributions between the test and control groups.

Although the positive effect of text message reminder was significant in the test group the overall response rate still remained very low on both groups (table 1). This means that the text reminder might not have been effective enough for this age group and more action should be taken in activating these participants.

Sending a text reminder was an easy way and did not require a lot of work or time resources since the text sending application and interface had already been deployed in our institute. Most of the monetary resources went into requiring the mobile phone numbers for the nonrespondents from the phone company.

Sending a text reminder to a large group was not very cost-effective; at the end we calculated that the text reminder brought us about 170 participants who would have otherwise been nonrespondents. That is not a big group but in this era of t nonresponse large every participant counts. The future studies should include testing whether sending the text at an earlier point or even twice during the data collection period would increase the positive effect.

References

Pentala-Nikulainen O, Koskela T, Parikka S, Kilpeläinen H, Koskenniemi T, Aalto A-M, Muuri A, Koskinen S & Lounamaa A. (2018). *The basic results and methods of the National FinSote Survey 2017-2018*. Internet publication https://thl.fi/en/web/thlfi-en/research-and-expertwork/population-studies/national-finsote-survey

Dillman, D. A. (2007). *Mail and Internet Surveys: The tailored design method (2nd ed.)*. New Jersey: John Wiley & Sons, Inc.

Impact Factors Modeling of Households Deposit Dollarization in Ukraine

Nataliia Versal[1] and Iryna Rozora[2]

[1]Department of Insurance, Banking and Risk-Management, Taras Shevchenko National University of Kyiv, e-mail: nataliia_versal@univ.kiev.ua
[2]Department of Applied Statistics, , Taras Shevchenko National University of Kyiv,
e-mail: irozora@bigmir.net

Abstract

In the paper the households deposit dollarization (HDD) in Ukraine is considered. We determine the significance of factors influencing the change in the level of HDD. The data for modeling are taken from the site of the National Bank of Ukraine and the State Statistics Service of Ukraine for the period from January 2006 to December 2017.
Keywords: Deposit Dollarization, Devaluation, Interest rates, Exchange rates

1 Introduction

When we are looking into deposit dollarization, there is one simple truth to understand: there is an objective reality (i.e. in countries with unstable economies and weak local currencies) where economic agents consider foreign currency to be a reliable asset worth of investing money in (Duffy, Nikitin & Smith, 2006). Main reasons for deposit dollarization in emerging markets can be explained by hysteresis or ratchet effect due to high inflation, exchange rate volatility, interest rates volatility, etc. (Mongardini and Mueller, 1999, Honohan and Shi, 2002, Brown and Stix, 2014); currency risk premium (Honohan and Shi, 2002, Palley, 2003); money flow from abroad (Basso, Calvo-Gonzalez and Jurgilas, 2011; Versal and Stavytskyy, 2016); currency competition in value storing etc. Thus, dollarization of deposits inevitably appears in banking systems of emerging markets.

From the literature review, it becomes obvious that there are many factors that influence the decision of households to keep savings in one currency or another. At the same time, we decided to dwell on a more narrow issue. This is a problem of deposits keeping by households in foreign currency in banks. In particular, this question is interesting from the point of view of the structure of household deposits in local and foreign currencies in emerging markets.

Thus, the goal of our study is to determine the significance of factors influencing the

change in the level of household deposits dollarization (HDD) in Ukraine. In this regard, the most interesting is the model proposed by Neanidis and Savva, 2009. This model has the following form:

$$\Delta DD_{it} = \alpha_0 + \beta_1 \cdot erf_{it} + \beta_2 \cdot mbf_{it} + \beta_3 \cdot ec_{it} + \sum_{j=1}^{m} \gamma_j \cdot X_{j,it} + \varepsilon_{it}$$

According to this model, the dollarization of deposits depends on such factors as the exchange rate (erf), the monetary base (mbf), the error correction term related to the size of the desired dollarization (ec), and also such control variables as the interest rate differential, the rate of inflation, an index of asymmetry of exchange rate movements, an index of exchange rate intervention etc. It should also be noted that the proposed model allows assessing several countries at once.

In turn, we propose an alternative model based on the Neanidis and Savva (2009) model.

2 Methodology and data

We've changed the model of Neanidis and Savva in two main directions. Firstly, we've changed the approach to the calculation of HDD. In the classic version, HDD is calculated as the ratio of deposits in foreign currency to the total volume of deposits in both currencies. At the same time, the work of Versal, Stavytskyy (2016) explains the problems of this approach under sharp devaluation of the local currency. Accordingly, we propose to calculate HDD as a ratio of deposits in foreign currency to deposits in local currency. A similar approach is also used nowadays by IMF research staff (Mwase and Kumah, 2015). In Fig. 1, 2, the difference in approaches is obvious. In Fig. 1, which shows the results of the classic HDD calculation, its growth is evident with the devaluation of the local currency. In turn, a completely different HDD trend, if we exclude the impact of the exchange rate.

Figure 1: Classic HDD

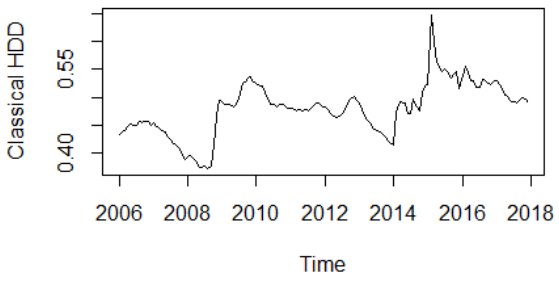
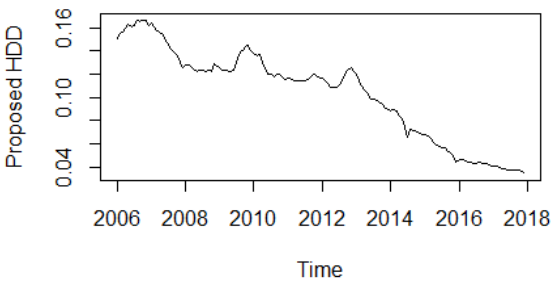


Figure 2: Proposed HDD



Secondly, we've reduced the number of variables, which makes the model applicable to one country. In particular, we've left in the model the most significant factors: the exchange rate, the monetary base, the difference in interest rates on deposits in local and foreign currencies, as well as the level of inflation. In addition, we added another important factor that affects the level of deposits in the banking system - wages.

The effect of the exchange rate on HDD may be different depending on the approach chosen for its calculation. If we take into account the classical approach, it is obvious that a direct link will manifest itself. At the same time, from an economic point of view, the opposite effect also possible. Thus, the growth of the foreign exchange rate for an import-dependent country can with high probability result in an increase in the prices of goods, which means that the households will be less able to save in any currency, i.e. the level of deposits in both foreign and local currency may decline. Another important factor is the peculiarity of deposit guarantee schemes. Ukraine is a vivid example. In Ukraine, the maximum amount of guarantee is expressed in local currency and is currently UAH 200 thousand, i.e. it's about USD 7700. Before the devaluation, the amount of guarantee covered USD 25 thousand dollars. This means that a part of the households has lost significant amounts of foreign currency deposits during the liquidation of banks in the last crisis. In tested models, we use the rate of growth of the exchange rate (*gerf*).

The impact of the monetary base on the HDD can also be ambiguous. On the one hand, an increase in the supply of money can stimulate savings in foreign currency; on the other hand, if an increase in the supply of money leads to inflation, the opposite impact is also possible. In tested models, we use the growth rate of the monetary base (*gmbf*).

The effect of the difference in interest rates on deposits in national currency (*dir*) and in foreign currency would seem to be exclusively inverse, i.e. the greater the difference, the smaller the HDD. In fact, it is not so obvious. In particular, the factor of trust plays a key role in emerging markets. Throughout the history of independent Ukraine, we can single out a few periods when the level of people's confidence in the local currency was high. This means that even if a very high interest rate is set for deposits in the local currency, the households may prefer deposits with a low interest rate, but in hard currency.

The level of inflation (CPI) can also have both a positive and negative impact on the HDD. This is explained by the fact that inflation expectations can push the population to store savings in foreign currency. At the same time, if the rate of inflation is very high, it can completely "eat up" income, which means that only a very small part will be converted into a hard currency.

In turn, such a factor as wage growth, on the one hand, can lead to the growth of HDD, if the rate of inflation is stable or less than the growth of wages. On the other hand, it can be the other way around. In tested models, we use the growth rate of wages (*gw*).

In this regard, we will test the following models:

$$HDD_i = \alpha_0 + \beta_1 \cdot gerf_i + \beta_2 \cdot gmbf_i + \beta_3 \cdot dir_i + \beta_4 \cdot CPI_i + \beta_5 \cdot gw_i + \varepsilon_i$$

Data for modeling are taken from the site of the National Bank of Ukraine and the State Statistics Service of Ukraine for the period from January 2006 to December 2017. Data are monthly.

2.1 Results

The results of estimated parameters are presented in table 1.

Table 1: Estimated values

Parameter	Estim. value	Std. Error	t value
α_0	0.1617318	0.0050433	32.069
β_1	-0.0706502	0.0423588	- 1.668
β_2	0.0866216	0.0895068	0.968
β_3	-0.0082190	0.0006089	-13.499
β_4	-0.0024468	0.0012103	-2.022
β_5	-0.0613168	0.0425139	-1.442

Residual standard error is equal to 0.02512 with 137 degrees of freedom. Multiple R-squared coefficient is $R^2 = 0.6042$,Adjusted R-squared is $R^2_{ad} = 0.5898$. F-statistic with 5 and 137 degree of freedom equals 41.83 and p-value: $<< 2.2 \cdot 10^{-16}$. Therefore, we can conclude that there is no reason to reject proposed model and the connection is significant.

References

Basso, H. S., Calvo-Gonzalez, O., Jurgilas, M. (2011). "Financial dollarization: The role of foreign-owned banks and interest rates", *Journal of Banking & Finance*, No. 35, 794-806.

Brown, M. Stix, H. (2015). "The Euroization of Bank Deposits in Eastern Europe", *Economic Policy*, Volume 30, Issue 81, 95–139.

Duffy, J., Nikitin, M., Smith, R.T. (2006). "Dollarization traps", *Journal of Money, Credit and Banking*, No. 38, 2073-2097.

Mongardini, J., Mueller, J. (1999). "Ratchet Effects in Currency Substitution: An Application to the Kyrgyz Republic", *IMF Working Paper*, No. WP/99/102.

Mwase N. and Francis Y. Kumah (2015). "Revisiting the Concept of Dollarization: The Global Financial Crisis and Dollarization in Low-Income Countries", *IMF Working Paper WP/15/12* Retrieved from: https://www.imf.org/external/pubs/ft/wp/2015/wp1512.pdf

Honohan, P., Shi, A. (2002). "Deposit dollarization and the financial sector in emerging economies", *World Bank Working Paper*, No. 2748.

Neanidis, K. C., Savva, C.S. (2009). “Financial dollarization: Short-run determinants in transition economies”, *Journal of Banking & Finance*, No. 33, 1860-1873.

Ozsoz, E. (2009). “Evaluating the Effects of Deposit Dollarization in Financial Intermediation in Transition Economies”, *Eastern European Economics*, Vol. 47, No. 4, 5-24.

Palley, T.I. (2003). “The economics of exchange rates and the dollarization debate: the case against extremes”, *International Journal of Political Economy*, Vol. 33, No. 1, 61-82.

Versal, N., Stavytskyy, A. (2016). “Trends in dollarization of Ukrainian banking sector”, *Economy and Forecasting*, No. 4, 106 – 117.

Brexit: challenges to estimate UK population

Natalia Rozora

Nielsen, e-mail: natalia.rozora@nielsen.com

Abstract

After UK's vote in 2016, UK will leave EU is on March 2019. Brexit changes not only Custom union, but also Immigrants systems. That changes not only future migration flow, but also put a lot of uncertainty to current UK immigrants. Non-existence of unique identity database for UK population even more complicates the population estimation after Brexit with switching to register-based population census. Considering growing homelessness population part (including hidden) put demand for more sophisticated approaches including simulation modelling under different scenarios of Brexit.
Keywords: Population census, Brexit, register-based population census, homelessness

References

Office for National Statistics. *https://www.ons.gov.uk/*

Suzanne Fitzpatrick, Hal Pawson, Glen Bramley, Steve Wilcox, Beth Watts & Jenny Wood. The homelessness monitor: England 2018. *Institute for Social Policy, Environment and Real Estate (I-SPHERE), Heriot-Watt University; City Futures Research Centre, University of New South Wales. April 2018*

Lavallée, P. & Rivest, L.-P. (2012). Capture–recapture sampling and indirect sampling. *Journal of Official Statistics* **28**, 1 - 27.

Särndal, C., Swensson, B. & Wretman, J. (2003). *Model assisted survey sampling.* Springer Verlag.

Tourism Incomes and Expenditures Surveys in Belarus

Natallia Sakovich

Belarus State Economic University, e-mail: sakovich-n@rambler.ru

Abstract

The main information sources of Tourism Statistics are considered. They include: 1) tourism industry enterprises, entities reports; 2) Households Sample Survey; 3) Labour Force Survey. For each of them the purposes, content, sample design, data collection mode, possible ways of surveys improvement are analysed
Keywords: tourism, incomes, expenses, satellite account, sample surveys

1 Introduction

The national System of Tourism Statistics (STS) as past of the National System of Statistics (NSS) is viewed as basic framework for coordination of statistical information on tourism as produced by all types of stakeholders. The STS is a set of interconnected statistical components comprising:

- statistical sources;
- data derived from those sources: statistical drawn from surveys, administrative records, statistic of the more synthetic nature, such as those integrated into and derived from tourism statistical activity, and the related data derived from adjacent statistical areas, like the balance of Payments and the National accounts;
- the specific tools and instruments used at certain stages of process (concepts, definitions, databases), Households surveys, tourism industry entities surveys are the main methods of tourism expenditures and income estimation.

2 Household Sample Survey

Households (HH) Sample Survey is conducted since January, 1995. Its main purpose is to get the information about the welfare of all population and particular demographic groups, detailed income and expenditure data. Main components of the survey are: baseline interview, four-quarterly interviews, four two-week interviews, which HH receives every quarter. More than 10 000 variables are investigated in the survey.

Survey object is households. Survey is carried out in all country regions and separately

in Minsk. Annually the survey covers 0,2 % or 6000 HH. In this survey three-stage probabilistic territorial sampling is used: 1) at the first step sampling units are cities and village councils; 2) at the second step – local-polling districts in city and data of the soviet account in village councils; 3) at the third – HH.

The methodology of weighing and extrapolation data on a general population is based on assignment of each finite unit (HH) the corresponding weight (B_i):

$$B_i = \frac{1}{p_1 \cdot p_2 \cdot p_3} \quad (1)$$

where p_1 - the probability of selecting each city and rural soviet; p_2- the probability of selecting each polling district in cities, zones and rural soviets; p_3 - the probability of selecting each household within the polling station or zone.

Base HH weights are corrected on uninhabited apartments and non-responses by using mathematical methods.

The sample program assumes filling in daily and quarterly questionnaires: expenses on food and nonfood, payment of services etc. Specifical questions on tourism:

- Did you make tourism trips?
- What were general parameters of your tourism trip (purpose; domestic, inbound or outbound; number of households members; durations)?
- Did you have tourism expenditures during reference quarter?
- What kind of the tourism expenditures did you have (package travel, package holidays and package tours; accommodation; food and drinks; local transport; international transport; recreations, culture and sporting activities; shopping; others)?

Household-based data provide significant types of information that the other cannot suitably supply. Current population characteristics, for example, are obtained either from household surveys.

3 Labour Force Survey

In Belarus Labour Force Survey (LFS) is conducted since 2012. One of the purposes is: to obtain statistics on employed by kind of activity, including tourism. Frequency of the results: quarterly and annual.

Sample frame is based on the 2009 Census and includes: set of cities in each region; set of village councils in each region; census enumeration districts in each selected city; villages in each selected village council; the households in each selected unit.

Sample size. The size of the sample is perhaps the most important parameter of the sample design, as it affects the precision, cost and duration of the survey more than any other factors. Used key indicator is the real unemployment rate. Target groups are economically active population (rural, urban, by regions, 5-year groups). Calculation

results by different variants have shown that required sample size is 26-29 thousand of households, or in average – 28 thousand. Without taking into account non-responses sample size is 22 thousand. Sample fraction is 0,6% HH.

Sample design. Survey object is the private households in urban and rural areas for each region, resident persons aged 15-74 years. Territorial three-stage sample is used. Primary unit – city or village council; secondary unit – census enumeration district or village; final sampling unit – household. At each stage units are selected with systematic selection with the probability, that is proportional population number or household number. Variables for stratification are: administrative districts, urban/rural.

Weighting procedure is connected with HH's weights and individual's weights. HH weights are calculated as inverse of overall sample probabilities. There are 25 census enumeration districts in cities and 16 village councils (zones). Individual's weights are based on iterative weighting:

Iteration I: a) weights are calculated by sex, five-year groups design; b) the first correction coefficient (k_1) is calculated; weighted variables are: region, sex, rural/urban; c) the second correction coefficient (k_2) is calculated; variables are: region, sex, 11 five-years groups. Weight is equal within each region, five-year groups in urban or rural area.

Iteration II: final individual weights for each five-year group:

$$K_i = B_b \cdot k_1 \cdot k_2 \tag{3}$$

where: $B_b = \frac{S_j}{s_j}$; $k_1 = \frac{S_t}{S_E}$; $k_2 = \frac{S_{jt}}{S_{E2}}$; S_j, s_j – population size in j-th sex-age group based on the result of the Census and survey; S_t – population size in t-th group by rural (urban), sex (on the Census data); S_E – extrapolated population size in t-th group (by B_b); S_{jt} – population size in j-th sex-age rural (urban) group; S_{E2}– extrapolated population size in jt-th group (by B_b and k_1).

Household-based surveys definitions of employment comprises paid workers, self-employed persons and contributing unpaid family workers who worked at least one hour or more during the reference period. The household survey provides information on the work status of the population without duplication (employed, unemployed or not in the labour force). Employed persons holding more than one job are counted only once. The household survey measures the earnings of paid workers in all occupations and industries in both the private and public sectors. The LFS covers hours worked by employees and self-employed persons. However, the LFS only estimates the employment within the borders of a given country and does not usually capture for example cross bolder workers or foreign seasonally workers.

4 Establishment-based Survey

Tourism activity is a complex, demand driven phenomenon. The tourism sector reflects this complexity by classifying comprehensive but fragmented set of industries of

tourism this complexity poses challenges for many domains within official statistics as it requires a fine level of disaggregation of activity with more details than usually produced. The typical tourism products are: accommodation of visitors, services of public catering entities, air transport, transport, used to cross land borders (railways, other public transport by land, private transport by land, pedestrians), river transport, rent of transport, tourism industry entities, culture services, sporting and others.

There are the following groups of primary data sources in labour statistics, business-statistics, statistics by kind of activity: 1) Establishment-level data (measuring labour demand); 2) Administrative records, such as: employment office registers, social security files, tax records, etc.

An Establishment is an enterprise, or part of an enterprise, that is situated in a single location and in which only a single productive activity accounts for most of value added. Establishment surveys are establishment censuses and establishment-based sample surveys, including small business surveys. The priority is given to the continuous reporting.

There are a large variety of establishment surveys, each designed to obtain specific information: production, export, employment, average earnings, etc. Detailed industrial classifications are much more reliably derived from establishment reports.

5 Conclusions

The experience of households survey and establishment surveys has shown following:

- main problems are: small sample localization; non-responses (20-30%); building of regional subsamples; the usage of methods combination to extrapolate each indicator from questionnaire becomes inadequate in some cases; different weighting schemes; structural parameters of employment estimation (for LFS);

- data from two sources (Households surveys and establishment-based surveys) differ from each other because of variations in definitions and coverage, source of information, methods of collections, estimating procedures. Sampling variability and response errors are additional reasons for discrepancies;

- additional specific tourism surveys are planned to use (informal tourism sector, telephone surveys, tourism employment).

References

Bokun, N (2013). Sample Survey of Households in Belarus: state and perspectives. *Statistics in transition, Warsaw*, 110-121

International Recommendation for Tourism Statistics 2008. Complicative Guide (2010)

Metodologicheskie polozheniya po postroeniyu vspomogatelnogo scheta turizma Respubliki Belarus (2017). Minsk, Belstat.

Sample Weights Calibration with Aim to Reduce the Estimation Bias Due to Under Coverage of the Well-Off Population

Volodymyr Sarioglo[1] and Nataliia Romanchuk[2]

[1] Ptoukha Institute for Demography and Social Studies, e-mail: sarioglo@idss.org.ua
[2] Ptoukha Institute for Demography and Social Studies, e-mail: romanchuk_nataliia@ukr.net

Abstract

In modern household sample surveys many indicators are estimated with significant bias due to the unwillingness of households to answer some questions, and undercoverage of the well-to-do population strata. When there is no access to personalized register data, an effective approach to mitigating these problems can be calibration of survey design weights with the use of relevant external information.

In the State Household Living Conditions Survey (HLCS) provided by the State Statistics Service of Ukraine (SSSU) a complex calibration procedure is tested in 2015 – 2017 in order to reduce the bias of income, income differentiation and income related estimates. Main data sources for this procedure are data on household disposable income from the National Accounting System (NAS) and data from the Tax Administration (TA) on wages and salary distribution. Received results testify to potential efficiency of such approach for increase of HLCS basic indicators reliability.

Keywords: calibration, estimation bias, reliability

1 Introduction

One of the serious problems that official statistics have to deal is reliable estimation of real income of households and their members. This problem is especially relevant for countries with a high level of shadow and informal economy, which is, in particular, Ukraine. In such conditions assessment of household ability-to-pay for utilities and services, level of social support programs targeting, household tax burden, and other issues become very problematic. Accordingly, the efficiency and effectiveness of socio-economic and fiscal policies are reduced.

Data on household incomes derived by modern household sample surveys are characterized by such disadvantages as underestimation of income due to the unwillingness of households to answer questions about the level and sources of income, and inadequate coverage of the well-to-do population strata due to their refusal to participate in surveys. Over the past decade these problems have become much worse

which negatively affects the reliability of the direct estimation of many important indicators by the results of sample surveys.

One of the main approaches to overcome these problems is the use of additional information (auxiliary in relation to the survey data). The use of such information is possible at different survey stages and depends on the research objectives, available auxiliary information, its quality and compatibility with the main source of data, etc.

Calibration of sample design weights is one of such approaches used at the indicator estimation stage (Deville, J.-C. and Särndal, C.-E. 1992). Calibration allows you to take into account available reliable auxiliary information in the indicator estimates and to provide analysis using full survey data set. In the HLCS during 2015 – 2017 a complex calibration procedure with a view to reduce possible biases in estimates of indicators due to households refuses to participate in the survey was tested. This procedure use demographic data, data from the NAS and the TA.

2 Methodological approach and results

The calibration procedure is carried out in 3 main stages: preparatory, and two calibration stages. At the preparatory stage data from additional sources are prepared for use in the calibration procedure. Numbers of population by sex-age groups and regions, numbers of households by regions and type of area are calculated. This information is used for calibration in the HLCS more than ten years. The new auxiliary information is the percentile distribution of the TA data on wages and salary by regions and data from NAS on disposable income by regions. For this information at the preparatory stage some data harmonization procedures are implemented: for instance, in the NAS data the amount of imputed rent is excluded from the amount of disposable income in every region; in the TA data the amounts of social contribution, income tax and military tax are excluded from the total amount of wages and salary in percentiles.

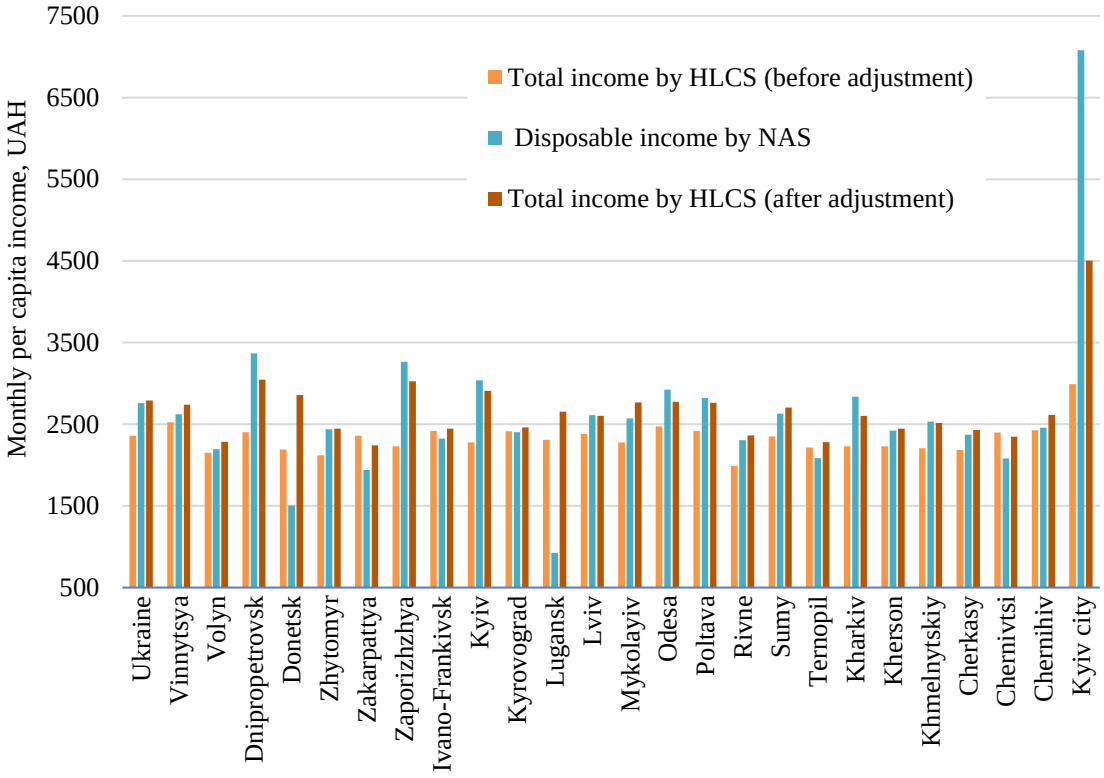
At the first stage of calibration of the HLCS sample weights the prepared auxiliary information on population by strata (regions and area type), household numbers by strata, and age - sex structure of population is used (SSSU 2011).

At the second stage of calibration information on regional distribution of disposable income and decile distribution of population by TA data is used (SSSU 2014). Wherein only TA data for decile groups in which number of people by TA data is higher than by HLCS estimates is considered. For all regions these are 8 -10 or 9, 10 decile groups only. Also numbers of population and households by regions and type of area are taken into account.

As it can be seen from the data presented on the Fig. 1, results of adjustment (calibration) are more significant for regions with higher household disposable income – Dnipropetrovsk, Zaporzhzhya, Kyiv city. It should be noted that in these regions the amount of wages and salary for highest decile groups by TA data is also higher. For some western regions – Zakarpattya, Ivano-Frankivsk, Ternopil, Chernivtsi – the

disposable income is lower than income, estimated by HLCS, but TA data nevertheless lead to a correction of income upwards.

Figure 1: Monthly per capita income by region of Ukraine, 2016



In the Table 1 some differences in estimates of household expenditures before and after weights calibration are presented. As it can be seen for some groups of expenditures the differences are quite significant.

It should be noted that in practice full compliance of HLCS adjusted estimates and auxiliary data is generally not achieved. This is due to restrictions on the minimal quality level of calibrated weights – maximum and minimum values, correlations with the design weights etc.

As it follows from the obtained results, some direct estimates of the HLCS can be substantially biased. Accordingly, their reliability in reality can be much lower than estimated without taking this fact into account. In our opinion the proposed calibration scheme can significantly improve the reliability of the HLCS results.

Table 1: Differences in estimates of household expenditures
before and after weights calibration

	Per household (UAH)		***% of total consumption expenditures***	
	before adjustment	*after adjustment*	*before adjustment*	*after adjustment*
Consumption expenditures				
food and non-alcoholic beverages	2852.69	3055.22	53.6	51.5
alcoholic beverages and tobacco	168.14	185.31	3.10	3.20
clothing and footwear	314.65	367.47	5.9	6.2
housing, water, gas, electricity and other fuels	917.53	999.77	17.2	16.8
furnishings, household equipment and routine maintenance of the house	97.16	116.10	1.8	2.0
transport	205.33	270.94	3.9	4.6
recreation and culture	80.63	106.40	1.5	1.8
restaurants and hotels	125.04	169.04	2.3	2.9
miscellaneous goods and services	142.06	165.86	2.7	2.8
Non consumption expenditures				
help relatives, other people	186.74	224.53	3.2	3.4
purchase of real estate, bank deposits, construction, overhaul	141.73	271.11	2.5	4.2
GINI index (by the total income)			0.220	0.234
Number of persons with equivalent per capita total income lower than the minimum subsistence level, %			51.1	35.2

References

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

Methodology of calculating the sampling weights for grossing up of the results of the State Household Living Conditions Sample Survey on the population. – Kyiv, State Statistics Service of Ukraine (SSSU), 2011 (in Ukrainian). Available at: http://www.ukrstat.gov.ua/metod_polog/metod_doc/2006/521/metod.htm (Accessed: 18 May 2018).

Methodology of adjustment of household's living condition indicators with the purpose of taking into account income and expenditures of the well-to-do population. – Kyiv, State Statistics Service of Ukraine (SSSU), 2014 (in Ukrainian). Available at: http://www.ukrstat.gov.ua/metod_polog/metod_doc/2014/415/metod_pol_koryg_zhrn.zip (Accessed: 18 May 2018).

Calibration of Register Based Census Data

Elvijs Siliņš[1]

[1]Central Statistical Bureau of Latvia, e-mail: silinselvijs@gmail.com

Abstract

The aim of this work was to get detailed statistical information about indicators of economic activity. Regularly obtain information about economic activity of inhabitants of Latvia is one of the Labour Force Survey (LFS) targets. Lack of LFS is that the obtained estimations are credible only in large domains but not in small population subgroups. Since Census 2021 is coming, once in a year Central Statistical Bureau of Latvia prepare register based census data where also is available information about economic activity. Lack of this data source is that it is not always possible to get all necessary information only from administrative data sources and these data could be biased because of measurement errors. Approach in this paper is to use the strengths of both data sources to get the desired result. Thus, it was evaluated which variable estimates from LFS is precise enough and in the next step register based census data were calibrated to these results. Calibration was done in program R using function 'calib' from package 'sampling'. More detailed methodology and results will be presented in workshop.
Keywords: register based census data, Labour Force Survey, calibration, detailed statistics.

References

R Core Team (2018). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ .

Yves Tillé and Alina Matei (2016). Sampling: *Survey sampling*. R package version 2.8. https://CRAN.R-project.org/package=sampling

Reporting tool for annual change studies by using survey data

Alina Sinisalo[1] and Arto Latukka[2]

[1] Natural Resources Institute Finland, e-mail: alina.sinisalo@luke.fi
[2] Natural Resources Institute Finland, e-mail: arto.latukka@luke.fi

Abstract

The economic results of Finnish agricultural holdings are published in Economy Doctor online system. The objective of this project was to develop Economy Doctor online service towards more versatile reporting system. A new feature was added with automated calculation routine to report year-to-year differences by classes selected by user, maintaining data confidentially, to promote further utilization and to produce more concrete help for researchers, decision makers and the public audience.
Keywords: FADN, change, reporting

1 Background and objectives

The Farm Accountancy Data Network (FADN) is an instrument for evaluating the income of agricultural holdings and the impacts of the Common Agricultural Policy. FADN data is collected every year from a sample of the agricultural holdings in the European Union. Natural Resources Institute Finland (Luke) is responsible of organizing the delivery of survey results to the EU from Finland.

Economy Doctor is a reporting service (www.luke.fi/economydoctor) for publishing time series of business activities and income of Finnish agricultural holdings. Due to data confidentially decrees the results are published at average level by region, economic size and type of holding. Every year microeconomic data is collected from approximately 900 voluntary agricultural holdings. The observations in sample are suitably weighted by using weight factors calculated individually for each farm and, further, the data are used to describe the situation of all Finnish farms.

Year-to-year changes are often of interest when studying economic performance and comparing the previous results present situation. To calculate the changes it may be needed to gather information from various sources and then subtract the results.

We developed a new feature in Economy Doctor internet reporting service, which runs analyses to estimate for example how each kind of income or cost affect the observed

profitability and financial statement. A routine programmed in Economy Doctor calculates for any economic report the differences between selected years of interest.

The objective of this project was to implement automated calculation routines for more effective utilization of sample survey results without compromising data privacy. The overall target is to develop Economy Doctor reporting service to produce more concrete help for researchers, decision makers and the public audience.

2 Reporting tool and results

Agriculture and horticulture service in Economy Doctor is founded on flexible reporting system and user interface, which allow extensive possibilities to examine the economic results of Finnish agricultural holdings. The results are based on profitability bookkeeping farm system administrated by Statistical services in Luke.

In online service it is not possible to get results of individual farms. All results are average figures from at least five farms. The results are shown as rounded figures, which is part of data confidentially policy.

The reporting routine collects from database the holdings matching the selected criteria and calculates a report of interest. The routine includes a weighting procedure, utilizing Farm Structure Survey (FSS), such that the results are representative by type of holding and economic size of those taken into examination. Weight factors are calculated individually for each farm based on the fact how much there are on each area different farms in regards of type of holding and economic size. The reports show the number of bookkeeping farms by classes and the rounded number of all farms.

A new feature to calculate the differences in Economy Doctor Agriculture and horticulture services was programmed. User can get reports between any selected years by classes of interest. The system produces automatically the differences reported in table format. Before the result table presented to user the system calculates first weighted annual averages and further differences between years. It is not obligatory to study consecutive years, but any period of change can be reported. The results are calculated in real time by automated calculation routine.

Example printout is given in Table 1 indicating the year-to-year changes for the period 2010–2015. In addition to years and type of holding, it is possible to optionally select two other classification variables.

Table 1: Differences for the dairy farms' production costs in 2010–2015. Partial printout from Economy Doctor Agriculture and horticulture service.

Production Costs	Dairy Farms				
	2011_2010	2012_2011	2013_2012	2014_2013	2015_2014
Farms represented	9 520	8 950	8 430	8 040	7 680
Farms in sample	330<n<340	320<n<330	310<n<320	300<n<310	280<n<290
Arable land	3,1	4,5	0,5	3,6	3,8
Livestock Units	1,6	4,2	1	3,2	3,1
PRODUCTION COSTS	**18 056**	**39 255**	**8 036**	**14 275**	**6 868**
Material costs	**8 175**	**9 668**	**4 246**	**1 893**	**1 813**
Fertilizer. Lime	504	1 627	162	1 006	-486
Other crop production costs	776	914	226	583	528
Fuel and lubricants	1 465	1 784	70	-286	-170
Electricity	710	481	66	-13	381
Forage costs	4 721	4 862	3 723	603	1 560
Farm use	**5 052**	**9 153**	**1 131**	**-107**	**1 559**
Livestock costs	**388**	**2 328**	**78**	**1 113**	**809**
Livestock purchasing	-313	603	-329	-154	32
Other livestock costs	701	1 725	407	1 267	777
Machinery cost	**3 287**	**3 924**	**2 332**	**3 893**	**411**
Depreciation of machines	1 169	1 235	674	919	934
Other machinery costs	2 118	2 688	1 658	2 975	-523
Buildings costs	**594**	**1 603**	**898**	**2 018**	**-423**
Depreciation of Buildings	287	1 373	1 042	746	595
Other buildings costs	308	230	-145	1 272	-1 018

The aim is to develop Economy Doctor online portal even more user-friendly and to offer more versatile possibilities to study key indicators and economic results based on the need of users. Future work is to add visual reporting environment.

References

European Commission (2018). *Farm Accounting Data Network An A to Z of methodology.* Available at: http://ec.europa.eu/agriculture/rica/pdf/site_en.pdf

Natural Resources Institute Finland (2018). *Economy Doctor* online portal, available at: http://www.luke.fi/economydoctor

Sinisalo, A. ja Latukka, A. (2018). *Tunnuslukujen muutosten tutkiminen Taloustohtorissa*. In: Toim. Tuula Puhakainen ja Mikko Hakojärvi. Maataloustieteen Päivät 2018, 10.–11.1.2018, Viikki, Helsinki: esitelmä- ja posteritiivistelmät. Suomen maataloustieteellisen seuran tiedote no 34: p. 163. Available at: http://www.smts.fi/sites/smts.fi/files/MTP2018_Abstraktikirja.pdf

Preparation for the register-based census

Milda Šličkutė-Šeštokienė[1]

[1]Statistics Lithuania, e-mail: milda.slickute@stat.gov.lt

Abstract

Statistics Lithuania, like other National Statistics Institutes, is constantly moving towards a wider usage of administrative sources. Administrative sources help to spare the costs as well as to improve the quality of the results. In Statistics Lithuania 43 percent of the published results are based on administrative sources.

Administrative sources were also widely used for Population Census 2011, but only as auxiliary information. In 2021 Population Census will be for the first time completely register-based, all the micro data will be obtained by linking number of administrative and statistical sources, no fieldwork will be carried out.

Keywords: census, register-based, administrative sources.

References

Wallgren A. and Wallgren B. (2014): Register-Based Statistics: Administrative Data for Statistical Purposes. John Wiley & Sons, Ltd

The Local Pivotal Method and its Application on StatVillage Data

Diana Sokurova[1]

[1]University of Tartu, e-mail: diana.sokurova12@gmail.com

Abstract

The purpose of this paper is to give an overview of the local pivotal method and compare it with other well-known sampling methods, applied on the real data. In the theoretical part detailed description of the local pivotal method is given. In the practical part, Monte Carlo simulation is conducted to find out which sampling method gives better estimation for data coming from a hypothetical village StatVillage, which is based on real data.
Keywords: survey sampling, sample survey theory, statistical estimation, pivotal method, local pivotal method

1 Introduction

The aim of a probabilistic survey sampling is to find out the strategy and estimator-function that leads to best estimate of the population's parameter of interest. Very popular sampling methods are simple random sampling, for which all objects in population have the equal inclusion probabilities, stratified sampling, for which population is divided into strata by determined criteria and after that some sampling method is applied in every stratum separately. The stratified sampling method, with simple random sampling applied in each stratum, is called stratified random sampling. The method, where systematic sampling is applied in each stratum, is called systematic stratified sampling. systematic sampling is based on the selection of elements from an ordered sampling frame.

Here the novel sampling method is presented, called the local pivotal method. The local pivotal method is special case of the general pivotal method, what was introduced in (Deville & Tillé, 1998). The local pivotal method was created to achieve spatially balanced sample.(Grafström *et al.*, 2012)

2 Pivotal Method

The pivotal method allows using the unequal inclusion probabilities as the initial ones. Method updates the initial inclusion probabilities interactively so that in each step the inclusion probabilities are recalculated until they becomes equal 0 or 1. The value of 1 means that corresponded unit is in sample, the value of 0 - not in a sample.

To describe the updating rule,some notation have to be introduced. We denote possibly updated inclusion probabilities with π_i', and that the unit i is finished if $\pi_i' = 0$ or $\pi_i' = 1$. Once a unit is finished, it is not used in algorithm again. (Deville & Tillé, 1998)

Algorithm 1. Pivotal Method

1. Choose two units i ja j randomly, $i, j \in U$, where U is the finite population.
2. Update the vector of their inclusion probabilities by the following updating rule.
   (a) If $\pi_i + \pi_j < 1$, then
   $$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j), \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0), \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j}. \end{cases}$$
   (b) If $\pi_i + \pi_j \geq 1$, then
   $$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1), \text{ with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1), \text{ with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j}. \end{cases}$$
3. Repeat algorithm until all units are finished(i.e equal to 0 or 1).

Because of the randomly choosing procedure, this method is also called random pivotal method.

3 Local Pivotal Method

The local pivotal methods update the inclusion probabilities according to the updating rule described above, but for two nearby units at each step. There is two different ways to choose the two nearby units i and j. At the first way, it is required that two units are the nearest neighbors to each other (Algorithm 2). At the second way it is sufficed that only one of units is the nearest neighbor to another (Algoritm 3). (Grafström *et al.*, 2012)

Algorithm 2. Local Pivotal Method I

1. Randomly choose one unit i.
2. Choose unit j, a nearest neighbor to i. If two or more units have the same distance to i, then randomly choose between them with equal probability.
3. If j has i as its nearest neighbor, then update the inclusion probabilities according to the updating rule. Otherwise go to 1.
4. If all units are finished, then stop. Otherwise go to 1.

Algorithm 3. Local Pivotal Method II

1. Randomly choose one unit i.
2. Choose unit j, a nearest neighbor to i. If two or more units have the same distance to i, then randomly choose between them with equal probability.
3. Update the inclusion probabilities for the units i and j according to the updating rule.
4. If all units are finished, then stop. Otherwise go to 1.

4 Simulation

To compare Local Pivotal Method according with other sampling methods on the data from StatVillage (Schwarz, 1997) Monte-Carlo simulation was used. For each sampling method, 1000 samples were drawn and 1000 estimates were found. Then Monte-Carlo mean and Monte-Carlo standard error was calculated by following formulas:

$$E_{MC}(\hat{t}) = \frac{1}{1000}\sum_{k=1}^{1000}\hat{t}_k,$$

$$\sqrt{V_{MC}(\hat{t})} = \sqrt{\frac{1}{999}\sum_{k=1}^{1000}(\hat{t}_k, -E_{MC}(\hat{t}))^2},$$

where $\hat{t}_k$ stands for the estimate total t in the simulation step k, $k = 1, \ldots, 1000$.

Simulations was done for two study variables: continuous variable household month income (moninch) , and discrete variable household size (hhsize). The block and the house numbers from address of household and number of income recipients were taken as the auxiliary information.

4.1 Results of Simulation

The result of Monte-Carlo simulation for continious variable are listed in the table 1

Table 1: Estimates and standard errors of variable *moninch*

Actual value	4 843 695	
Selection method	$E_{MC}(\hat{t})$	$\sqrt{V_{MC}(\hat{t})}$
Simple random sampling	4 838 280	152 718.57
Stratified random sampling	4 842 568	139 327.05
Systematic stratified sampling	4 845 962	58 380.02
Random pivotal method	4 844 994	156 132.98
Local pivotal method I	4 843 181	50 190.02
Local pivotal method II	4 844 111	49 183.08

Based on the table 1, the most accurate estimates are get from the systematic stratified sampling and both local pivotal methods where the difference in standard errors in local pivotal methods is small.

Below is given results for discrete variable for the various methods of selection.

Table 2: Estimates and standard errors of variable *hhsize*

Actual value	3 000	
Selection method	$E_{MC}(\hat{t})$	$\sqrt{V_{MC}(\hat{t})}$
Simple random sampling	2 999.64	73.22
Stratified random sampling	2 997.93	54.13
Systematic stratified sampling	3 002.18	51.15
Random pivotal method	3 001.38	71.17
Local pivotal method I	3 000.8	60.19
Local pivotal method II	3 001.44	58.42

Based on the table 2 , the most accurate estimate for standard error is derived from a systematic stratified sampling. Both local pivotal methods did not give the best or worst estimate, but the local pivotal method II is a bit more accurate than the local pivotal I.

In conclusion, in the case of a continuous variable, both local pivotal methods provide more accurate estimates, and in this case, it is recommended to use local pivotal method II because it is more accurate and faster in execution. In the case of a discrete variable, each of the local pivotal methods did not provide any better estimates, and in this case, it is advisable to use a systematic stratified selection.

References

Deville, J.-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89–101.

Grafström, A., Lundström, N. L. P. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics* **68**, 514–520.

Schwarz, C. (1997). Statvillage: An on-line, www-accessible, hypothetical city based on real data for use in an introductory class in survey sampling. *Journal of Statistics Education* **5**.

Sample models in monitoring survey UniDOS.

Mykola Sydorov[1] and Oleksiy Sereda[2]

[1]Taras Shevchenko National University of Kyiv, e-mail: myksyd@knu.ua
[2]Taras Shevchenko National University of Kyiv, e-mail: as_sereda@knu.ua

Abstract

The paper presents models of sampling in the 9th wave of the monitoring survey UniDOS 2013. The problems in sample construction comprise inaccurate and incomplete information about the population, hard access to some groups of students and small size of strata.
Keywords: Multilevel sampling, strata

1 Introduction

Since 2009 the Faculty of Sociology of Taras Shevchenko National University of Kyiv conducts a monitoring survey of students. The survey examines issues of motivation for obtaining higher education, plans for the future after graduation, expectations of students regarding the labour market and further employment, the attitudes to the educational process at the university faculties, etc. Every year we have the problem with constructing an optimal sample model. In this paper, we present sample models of poll in the spring of 2013.

2 General Population

The general population consists of full-time students of 17 faculties and institutes of the Taras Shevchenko National University of Kyiv, which is 17484 students. Part-time students was not included in the population, because, education is not their main activity and the time of stay at the university is considerably limited compared with the students of full-time education, as well as their limited stay and communication in the community. At the beginning of the survey we did not have all relevant data on the composition of groups and courses in the faculties of KNU, as based upon estimates for September 2013. The structure of the population is given in Appendix 1. The biggest problem was the quality of information for the first year students. Also the first year can be considered a special category for which the questionnaire is different from other years of study questionnaire because of first year students can't be asked about the quality of education and other aspects of the university because they studied just for several weeks.

Thus, the sample was calculated only for 2+ year students, of which 13658, excluding

those absent because of filed practice. Based on the purpose and objectives of the research, the research team was interested in the opinion of the representatives of all faculties and institutes, therefore, at this stage, a continuous selection is applied.

3 Sampling

To achieve the goals and objectives, as well as to implement the methodological plan of the study, we used multilevel sample selection scheme.

1st level - continuous selection of faculties, volume - in proportion to the number of students at the faculty: faculty - stratum.

Level 2 - stratification at the year of study (bachelor 2+, masters – all) - in proportion to the number of students available for each year: every year of study - stratum

Level 3 - nested method of selection, "nest" corresponds to the selected groups (group) of a year of study at the faculty.

Level 4 - random selection of respondents in each selected group (using two-colours cards).

The choice of such a model is due to the inability to obtain lists of students from all faculties and institutes, which would allow the use of the random selection model. At the same time, students are distributed at auditorium time by groups, in which the number and composition of students at each faculty are approximately the same. Thus, we have only 2 variables to describe the population: the approximate number of students in the faculties and the approximate number of students in the groups. We do not even have the distribution of students by gender.

The next step was to calculate the sample size. Based on the formula for simple random selection, for the general population 13658 we should have a sample of 374 respondents for a sampling error of 5% with a confidence level of 0.95. But, if we make conclusions about each faculty separately, then the sample size should be calculated for every faculty separately and we will receive a slightly larger number (Appendix 2). The total number of all respondents in all faculties in this case is 4105. This significantly exceeds the client's ability, which is about 1200 questionnaires.

Next, we propose two approaches to the formation of a sample population.

3.1. Sample approach #1

Since during the construction of a sample of 1,200 respondents in some departments there were not enough respondents in sample, for the adequate representation within the faculty, we decided to secure a minimum number of 50 respondents for each faculty / institute.

The sample was divided into 2 parts: proportional and additional. The proportional number was 1085 respondents and the additional - another 112 for those faculties,

where the number of respondents did not reach 50 (Appendix 3.). However, after the calculation of the size proportional to the courses selection at the faculties, it turned out that the size of additional selection should be 117 (Appendix 4). This is due to the rounding of the calculated numbers of students in the groups.

Thus, we obtained a sample that allows us to conduct a representative survey of students at the KNU. The sample size was 1202 respondents. Sample weights are given in Appendix 5.

3.2. Sample approach #2

Since the minimum number of respondents for each faculty must be at least 50 people, from the maximum sample size we can select 50 * 17 = 850 respondents. The residual of 1200 - 850 = 350 is distributed among the faculties, in proportion to the difference in between the number of students with the smallest faculty. The smallest number of students study at the Faculty of Sociology – 138 students available for surveying, therefore, to construct the proportions, we will subtract the number of students of each faculty from the number of faculty of sociology. Afterwards we calculate proportions. The general population, the proportion of the faculties and the estimated sample sizes by faculty are given in Appendix 6.

The next step was to calculate the sample size for each faculty and for each year of study. Due to rounding, the proportion has slightly changed and is shown in appendix 7. It also shows weight ratios.

To calculate this sample option, we used R (R Core Team, 2018) package and the surveyplanning package (Breidaks, Liberts, & Jukams, 2017).

It should be noted that during the field stage the first approach to the sample construction was used.

3.3. Final stage of sampling

The next step was random selection of groups, which was carried out for each year of the study of each faculty separately. We did not have information about the principle of grouping students into groups: in some departments, the division was based on an alphabet; in some - in separate groups there were students with higher grades, in other - students living in dormitories in one group, and local residents of Kyiv - up to other, etc. Therefore, the only approach to groups selection was random.

In order to select respondents in the group, the method of labelled cards was used as the only suitable method of randomization. The pre-interviewer receives a set of white cards and a set of white cards marked with a red square. The number of white cards was equal to the number of students presented in the group minus the number of respondents who should be interviewed in this group. The number of white cards marked with a red square is equal to the number of students to be interviewed in this group. The pile of cards with cards of both types (white and coloured) was well mixed and handed out to

the students.

At the stage of elaboration of the sampling design, several ways of respondent selection at the last stage (in groups) were proposed. Among the main ones were the following methods of selection:

- random selection of respondents by student lists. But the lists of students enrolled in the university are confidential information, so there are significant problems with the access;

- step-by-step selection of respondents directly in the classrooms. For this selection procedure, information is needed on the number of students in the group and the number of students to be interviewed. Based on this information, the interviewer calculates the step and makes selection. This selection procedure requires an interviewer to make mathematical calculations and takes a long time. This increases the possibility that the interviewer will make a mistake in the calculations, or in the process of calculating the step. There is a high probability that the interviewer will not be able to calculate the step at all and distribute the questionnaire to all interested persons.

The selection procedure with the use of cards was chosen because: the use of cards does not require serious mental calculations, as well as the availability of lists of students studying in this group.

- Working with multi-coloured cards and the selection procedure reminds a certain lottery game. Increased interest in the method of selection increases the motivation of the interviewer to follow all the requirements of this procedure.

- Red cards received by respondents are an important element in controlling the work of interviewers.

4. Conclusions

Both of the above procedures for sampling the population allow us to conduct a representative survey of students at the faculties of the KNU. The peculiarities of the design of the sample are related to the tasks set for the research team and the strict constraints on the customer's resources. In 2013, the 1st approach was implemented, although the 2nd one we consider simpler and more convenient. In addition, it has a smaller range of weights. The second method we plan to apply in the study, which will take place in September 2018.

References

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing,Vienna, Austria. URL https://www.R-project.org/

Breidaks J, Liberts M, Jukams J (2017). *_surveyplanning: Survey planning tools_. R package version 2.9*, <URL:https://csblatvia.github.io/surveyplanning/>

Appendices

Appendix 1.

Table # 1. Structure of the general population

Faculty / Institute	Number of students	B1 (Bach. degree stud. 1 year)	B2	B3	B4	S1 (specialist degree)	M1 (MA stud. 1 year)	M2
Educational and Scientific Centre "Institute of Biology"	933	196	172	129	151	21	133	131
Geography Faculty	1051	207	178	143	195	53	140	135
Geological Faculty	401	75	90	46	73	0	56	61
Faculty of Economics	1671	364	322	241	278	0	244	222
Historical Faculty	754	161	140	97	138	64	78	76
Faculty of Cybernetics	913	210	175	139	148	55	107	79
Mechanical and Mathematical Faculty	673	126	124	110	117	30	88	78
Faculty of Psychology	669	161	130	108	115	19	74	62
Radiophysics Faculty	750	142	157	163	126	18	78	66
Faculty of Sociology	347	84	72	44	63	18	33	33
Faculty of Physics	713	143	116	103	106	19	109	117
Philosophy Faculty	770	181	174	98	157	25	73	62
Chemical Faculty	481	91	88	73	83	20	66	60
Faculty of Law	2153	415	413	273	409	87	282	274
Institute of Journalism	1150	235	261	170	226	15	120	123
Institute of International Relations	1762	175	336	294	348	110	254	245
Institute of Philology	2293	612	510	476	25	43	305	322
Total	17484	3578	3458	2707	2758	597	2240	2146

In the table # 1, years in which the number of students indicated for 0 is those for which

there was no set ("Specialist" of the Geological and Economic Faculties). There were also years in which the students were in practice for the period of the survey, so they were difficult to access (they are highlighted in grey). For further calculations, their number was also counted as 0, because the ability to involve these students in participating in the survey is absent.

Appendix 2.

Table # 2. Calculation of the sample size for each faculty separately

Faculty / Institute	N of general population	n of sample
Educational and Scientific Centre "Institute of Biology"	737	253
Geography Faculty	844	264
Geological Faculty	326	176
Faculty of Economics	1307	297
Historical Faculty	593	233
Faculty of Cybernetics	703	248
Mechanical and Mathematical Faculty	547	226
Faculty of Psychology	508	219
Radiophysics Faculty	608	235
Faculty of Sociology	138	102
Faculty of Physics	570	229
Philosophy Faculty	589	233
Chemical Faculty	390	194
Faculty of Law	1738	315
Institute of Journalism	792	259
Institute of International Relations	1587	309
Institute of Philology	1681	313
Total	13658	4105

Appendix 3.

Table # 3. Calculation primary and additional selection in sample

Faculty / Institute	Total number for 2+	B2 (Bach. degree stud. 1 year)	B3	B4	S1 (specialist degree)	M1 (MA stud. 1 year)	M2	Number of additional selection
Educational and Scientific Centre "Institute of Biology"	59	14	10	12	2	11	10	0
Geography Faculty	67	14	11	15	4	11	11	0

Faculty / Institute	Total number for 2+	B2 (Bach. degree stud. 1 year)	B3	B4	S1 (specialist degree)	M1 (MA stud. 1 year)	M2	Number of additional selection
Geological Faculty	26	7	4	6	0	4	5	24
Faculty of Economics	104	26	19	22	0	19	18	0
Historical Faculty	47	11	8	11	5	6	6	3
Faculty of Cybernetics	56	14	11	12	4	9	6	0
Mechanical and Mathematical Faculty	43	10	9	9	2	7	6	7
Faculty of Psychology	40	10	9	9	2	6	5	10
Radiophysics Faculty	48	12	13	10	1	6	5	2
Faculty of Sociology	11	6	0	0	0	3	3	39
Faculty of Physics	45	9	8	8	2	9	9	5
Philosophy Faculty	47	14	8	12	2	6	5	3
Chemical Faculty	31	7	6	7	2	5	5	19
Faculty of Law	138	33	22	32	7	22	22	0
Institute of Journalism	63	21	14	18	1	10	0	0
Institute of International Relations	126	27	23	28	9	20	19	0
Institute of Philology	134	41	38	2	3	24	26	0
Total	1085							112

Appendix 4.

Table # 4. Final calculation of the sample in approach 1.

Faculty / Institute	Total number for 2+	B2 (Bach. degree stud. 1 year)	B3	B4	S1 (specialist degree)	M1 (MA stud. 1 year)	M2
Educational and Scientific Centre "Institute of Biology"	59	14	10	12	2	11	10

Faculty / Institute	Total number for 2+	B2 (Bach. degree stud. 1 year)	B3	B4	S1 (specialist degree)	M1 (MA stud. 1 year)	M2
Geography Faculty	66	14	11	15	4	11	11
Geological Faculty	50	14	7	11	0	9	9
Faculty of Economics	104	26	19	22	0	19	18
Historical Faculty	50	12	8	12	5	7	6
Faculty of Cybernetics	56	14	11	12	4	9	6
Mechanical and Mathematical Faculty	50	11	10	11	3	8	7
Faculty of Psychology	50	13	11	11	2	7	6
Radiophysics Faculty	52	13	13	10	5	6	5
Faculty of Sociology	50	26	0	0	0	12	12
Faculty of Physics	50	10	9	9	2	10	10
Philosophy Faculty	49	15	8	13	2	6	5
Chemical Faculty	50	11	9	11	3	8	8
Faculty of Law	138	33	22	32	7	22	22
Institute of Journalism	68	21	14	18	5	10	0
Institute of International Relations	126	27	23	28	9	20	19
Institute of Philology	134	41	38	2	3	24	26
Total	1202	315	223	229	56	199	180

Appendix 5.

Table # 5. Weights coefficients for the sample in approach 1.

Faculty / Institute	For the Faculty / Institute	B2	B3	B4	S1	M1	M2
Educational and Scientific Centre "Institute of Biology"	1.0993	1.0812	1.1353	1.1074	0.9241	1.0641	1.0812
Geography Faculty	1.1254	1.1189	1.1441	1.1441	1.1661	1.1201	1.1189
Geological Faculty	0.5738	0.5658	0.5783	0.584	1	0.5476	0.5658
Faculty of Economics	1.106	1.0899	1.1163	1.1121	1	1.1302	1.0899
Historical Faculty	1.0438	1.0267	1.0671	1.0121	1.1265	0.9806	1.0267
Faculty of Cybernetics	1.1048	1.1001	1.1121	1.0854	1.2101	1.0463	1.1001
Mechanical and Mathematical Faculty	0.9628	0.9921	0.9681	0.9361	0.8801	0.9681	0.9921

Faculty / Institute	For the Faculty / Institute	B2	B3	B4	S1	M1	M2
Faculty of Psychology	0.8942	0.8801	0.8641	0.9201	0.8361	0.9304	0.8801
Radiophysics Faculty	1.029	1.0629	1.1035	1.1089	0.3168	1.1441	1.0629
Faculty of Sociology	0.2429	0.2437	1	1	1	0.242	0.2437
Faculty of Physics	1.0033	1.0209	1.0072	1.0365	0.8361	0.9593	1.0209
Philosophy Faculty	1.0579	1.0209	1.0781	1.0629	1.1001	1.0708	1.0209
Chemical Faculty	0.6865	0.7041	0.7138	0.6641	0.5867	0.7261	0.7041
Faculty of Law	1.1084	1.1014	1.0921	1.1248	1.0938	1.1281	1.1014
Institute of Journalism	1.025	1.0938	1.0687	1.105	0.264	1.0561	1.0938
Institute of International Relations	1.1085	1.0952	1.125	1.0938	1.0756	1.1177	1.0952
Institute of Philology	1.104	1.0947	1.1024	1.1001	1.2614	1.1184	1.0947
MIN	**0.2429**	0.2437	0.5783	0.584	0.264	**0.242**	0.2437
MAX	**1.1254**	1.1189	1.1441	1.1441	**1.2101**	1.1441	1.1189

Appendix 6.

Table # 6. Calculation of the sample size for each faculty separately for approach 2

Faculty / Institute	N of general population	General – number of stud. sociology faculty	n of sample
Educational and Scientific Centre "Institute of Biology"	737	599	69
Geography Faculty	844	706	72
Geological Faculty	326	188	56
Faculty of Economics	1307	1169	86
Historical Faculty	593	455	64
Faculty of Cybernetics	703	565	67
Mechanical and Mathematical Faculty	547	409	63
Faculty of Psychology	508	370	61
Radiophysics Faculty	608	470	65
Faculty of Sociology	138	0	50
Faculty of Physics	570	432	63
Philosophy Faculty	589	451	64
Chemical Faculty	390	252	58
Faculty of Law	1738	1600	100
Institute of Journalism	792	654	70
Institute of International Relations	1587	1449	95
Institute of Philology	1681	1543	98
Total	13658	11312	1201

Appendix 7.

Table # 7. Sample in approach 2 by faculty and years.

Faculty / Institute	For the Faculty / Institute	B2	B3	B4	S1	M1	M2
Educational and Scientific Centre "Institute of Biology"	68	16	12	14	2	12	12
Geography Faculty	73	15	12	17	5	12	12
Geological Faculty	56	15	8	13	0	10	10
Faculty of Economics	86	21	16	18	0	16	15
Historical Faculty	63	15	10	15	7	8	8
Faculty of Cybernetics	67	17	13	14	5	10	8
Mechanical and Mathematical Faculty	62	14	13	13	3	10	9
Faculty of Psychology	61	16	13	14	2	9	7
Radiophysics Faculty	64	17	17	13	2	8	7
Faculty of Sociology	50	26	0	0	0	12	12
Faculty of Physics	63	13	11	12	2	12	13
Philosophy Faculty	65	19	11	17	3	8	7
Chemical Faculty	58	13	11	12	3	10	9
Faculty of Law	101	24	16	24	5	16	16
Institute of Journalism	70	23	15	20	1	11	0
Institute of International Relations	96	20	18	21	7	15	15
Institute of Philology	99	30	28	1	3	18	19
Total	1202	314	224	238	50	197	179

Min and max weights of coefficients for the sample in approach 2.

	Weights for Faculties	Weights for Years of study
min	0.8108	0.2216
max	1.4696	1.6618

Using Mobile Positioning for Improving the Quality of Register Data

Kaja Sõstra[1] and Kristi Lehto[2]

[1]Statistics Estonia, e-mail: kaja.sostra@stat.ee
[2]Statistics Estonia, e-mail: kristi.lehto@stat.ee

Abstract

The next population and housing census in Estonia in 2021 is planned to be register-based. One of the biggest challenges on the way to register based census in Estonia is the difference between registered and actual places of residence. By law, everybody is obliged to ensure that their correct usual address is entered in the Population Register, but there are different reasons why people don't do that. The difference between registered and actual places of residence affects the regional population statistics and all household and family characteristics.

To solve this problem Statistics Estonia (SE) carried out a pilot project for testing possibility to use mobile positioning data. The anchor points of the first and second place of residence were estimated based on the mobile positioning data. Anchor points and other auxiliary information was used to build a model for selecting the most probable place of residence from the set of addresses.
Keywords: register based census, mobile positioning, anchor point model

1 Introduction

Countries who have conducted register based population and housing census (PHC) are using information of place of usual residence from Population Register. Estonia is planning to conduct the next PHC based on registers. Estonian Population Register (PR) contains information of a place of usual residence. The problem is, that people don't live in their place of residence registered in PR in Estonia. Surveys performed by Statistics Estonia show that 20–25% of all permanent residents of Estonia have not registered their actual place of residence in PR and live at a different address. This problem has historical roots. At the beginning of Estonia reindependence in 1990s the registration of a place of residence in PR was voluntary. Now it is obligatory more than ten years but lots of people don't consider registration necessary (Äär, 2017). Today, there are several new reasons and factors causing false registration of place of residence in PR: school and kindergarten places, free public transport, several social benefits, wish to support different local government etc.

The difference between registered and actual places of residence affects the breakdown

of the regional population statistics and all household and family characteristics. (Tiit, Visk, Levenko, 2018).

2 Pilot study

2.1 Participation in pilot study

Statistics Estonia started pilot project for testing possibility to use mobile positioning data for register based census. The aim of pilot study was to test the feasibility of specifying actual place of residence with the help of mobile positioning data. Study was organised in cooperation with University of Tartu and Positium LBS. The pilot study consists of following steps:

- Statistics Estonia asked volunteers to participate in the pilot
- Set of potential addresses was created for each participant based on registers
- The home anchor points were estimated based on the mobile positioning data
- Anchor points and other auxiliary information was used to build a model for selecting the most probable place of residence from the set of addresses.

All participants of the study filled in digitally signed written consent to use their mobile positioning data. Together with consent, the following information was collected: participant's name and identification code, mobile number and mobile operator, the actual place of residence. Identification code was used for linking data from registers, mobile number and mobile operator were needed for obtaining mobile positioning data from operator. Participant's actual place of residence was collected to compare positioning data and register data with actual address.

All together 310 persons participated in the study, four participants gave two phone numbers. The phone numbers divided between three mobile operators: Telia (158) Elisa (103), Tele 2 (53). Mobile positioning data was received from Telia and Elisa. Participants with mobile numbers of operator Tele 2 were not included into next steps of the study.

2.2 Linking of addresses from registers

Set of potential addresses was created for each participant based on Population Register and Land Register. The following addresses were linked to the participant's data:

- Place of residence of participant from Population Register (PR)
- Addresses of participant's real estates from Land Register (LR)
- Place of residence of participant's spouse and children from PR
- Addresses of real estates of participant's spouse and children from LR
- Addresses of other relatives from PR and LR

Addresses from registers were compared with participant's actual place of residence. The actual place of residence coincided with PR address for 67% participants. For participants with different actual place of residence and PR address other addresses were compared with place of residence. Most likely these persons live on the address of own real estate registered in Land Register (26%), on the address of spoce's real estate (19%) or spouce's address in Population Register (14%). For about half participants living elsewhere than PR address it was possible to find their actual address from registers.

2.3 Calculation of anchor points

Agreements with mobile operators, mobile phone positioning data extraction and anchor points calculations were made by company Positium LBS. Mobile phone positioning data was received from two mobile operators (Telia and Elisa) for 261 phone numbers and 257 persons. Anchor points were calculated for 243 phone numbers (93%) and 240 persons. The mobile phone usage activity of some persons was too low for calculating anchor points.

Anchor points were calculated according to methodology developed by University of Tartu (Ahas et al, 2010). Anchor points are locations which person visits regularly. Calculation of anchor points consists of eight steps. The two regular cells that had the highest number of days with calls are selected for the calculation of home and work-time anchor points. Home anchor point has average starting time of daily calls after 17:00 and standard deviation of beginning time of calls greater than 0.175. Working time anchor point has average starting time of daily calls before 17:00 and standard deviation less than 0.175. If the home and work-time anchor points are located in the same network cell and cannot be separately identified then anchor point is determined as work-home anchor point.

For estimating the quality of anchor points, the home and work-home anchor points were compared with the actual place of residence of participant. Anchor point and actual place of residents were in the same county – 97%, in the same municipality – 87%, in the same settlement – 67%. The coordinates of the actual place of residence were inside of anchor point polygon for 82% of participants. Comparison shows that coincidence is high. Most of cases the distance is shorter than 10 km. There are only few cases with large discrepancy between the anchor point and the place of residence.

The accuracy of anchor points is at the level of the service area of a mobile antenna (a network cell). The diameter of the polygon of anchor point is negatively correlated with population density. There are more mobile antennae in the densely populated areas. Diameter of polygon of anchor point is about 1-1,5 km in urban areas and 10-15 km in rural areas.

3 Validation of the quality of place of residence in registers

For comparison and making decisions the polygon of anchor point and coordinates of register address were used. If polygon of home, second home or work-home anchor points included coordinates of register address then it was decided that register address is actual

address. If coordinates of register address were outside of all polygons then it was decided that this address is not actual address of participant. This decisions were compared with information about actual place of residence.

Results of validation of participants PR addresses are shown in table 1.

Table 1: Results of validation of PR addresses

		Decision made using mobile positioning data (home, work-home and second home anchor point)		
		PR address inside anchor polygon	**PR address outside anchor polygon**	**Total**
Coincidence of actual place of residence with PR address	**PR = Actual**	146	15	161
	PR ≠ Actual	16	63	79
	Total	162	78	240

Correct decision is made based on mobile positioning data for 146 + 63 = 210 persons (87%). Other register addresses described in chapter 2.2 were compared similarly and decision was made if other register address is actual address. The most probable place of residence was selected from the set of potential addresses.

4 Conclusions

One of the biggest obstacle on the way to register based census in Estonia is the problem of difference between registered and actual places of residence.

Feasibility study showed that mobile positioning data could be one solution for solving this problem. It is possible to use mobile positioning data for validation of the quality of registered addresses. One prerequisite for using mobile positioning data for register based census is solving legislation issues.

References

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M. (2010), Using mobile positioning data to model locations meaningful to users of mobile phone. *Journal of Urban Technology*, 17:1, 3-27

Sõstra, K., Lehto, K. Geospatial mobile data to increase the quality of usual place of residence. *Paper presented in European Conference on Quality of Official Statistics Q2018.*

Tiit, E.-M., Visk, H., Levenko, V. (2018), Partnership index. *Quarterly Bulletin of Statistics Estonia*. 1/2018, pp.29 - 43

Äär, H. (2017), Coincidence of actual place of residence with Population Register records. *Quarterly Bulletin of Statistics Estonia*. 1/2017, pp. 80–83.

Handling Nonsampling Errors— Case Salo

Maria Valaste[1,2]

[1]University of Helsinki, e-mail: maria.valaste@helsinki.fi
[2]The Social Insurance Institution of Finland (KELA), email: maria.valaste@kela.fi

Abstract

In real life nonsampling errors are almost inevitable. This paper concentrates on nonsampling errors. A research project "Sudden structural change – case study of Nokia-city Salo" from the Social Insurance Institution of Finland is discussed.
Keywords: survey, nonsampling error, nonresponse error, coverage error, measurement error

1 Introduction

Salo is a middle size town in South-Western Finland approximately 50 kilometers from Turku and 100 kilometers from Helsinki. Until 2012 the assembly factor of Nokia mobile phone company was situated in Salo. It employed more than 4000 persons and was the largest private employer is the area.

In the summer 2012 the factory was closed down causing Salo to become an area of sudden structural change. It has already since 2009, when some of the major subcontractors of Nokia were transferred to Asia, received millions of euros in order to minimize the negative effects of sudden structural change.

One aim in the research project is to follow the inhabitants of Salo and their well-being for several (approximately 10) years in order find out how they cope with the sudden structural change and its effects (Ylikännö & Kehusmaa, 2015). First baseline survey was conducted in spring 2013 and the second follow up survey in spring 2015. Currently the third survey is planned.

As expected, the survey data was not complete. E.g. the both surveys consisted nonresponse. This paper will focus on nonsampling errors.

2 Nonsampling errors

In a perfect case the variable of interest is measured on every unit in the sample without error, so that errors in the estimates occur only because just part of the population is included in the sample. Such errors are referred to as sampling errors. (Thompson, 2012). In real life nonsampling errors may also arise.

Groves (1989); Alwin (1991, 2007); de Leeuw *et al.* (2008); Groves *et al.* (2009) specify four sources of error in surveys: coverage error, sampling error, nonresponse error and measurement error. Most important types of nonsampling errors are nonresponse, coverage errors and measurement errors (de Leeuw *et al.*, 2008). Lehtonen & Pahkinen (2004) also adds to this list processing errors.

Nonresponse error

Nonresponse error occurs when some of the sampled units do not respond and when these units differ from those who do and in a way relevant to the study (de Leeuw *et al.*, 2008). There are two types of nonresponse in surveys: unit nonresponse and item nonresponse. Unit nonresponse is the failure to obtain any information from an sample unit. Item-nonresponse refers to the failure to obtain information for one or more questions in a survey, given that the other questions are completed. (de Leeuw *et al.*, 2008).

The methodologies for handling unit non-response and item non-response can differ but in both cases the reasons for missing values has to be investigated. Usually indicator variable is created for unit response or item response and missingness rates and descriptive statistics are computed.

Statistical weighting can be used to make the sample resemble the population with respect to some characteristics. E.g. post-stratification is a basic calibration method for to reduce the bias due to unit non-response. In order to create post-stratification weights an auxiliary information for specified subgroups of the population is required. The weights of the sample units is adjusted to match the totals within the specified subgroups. The subgroups are called post-strata, and the statistical adjustment procedure is called post-stratification.

Coverage error

In surveys two types of coverage errors may exist: undercoverage and overcoverage errors. An undercoverage error arises when some population elements are not included in the sampling frame. An overcoverage error is present when a unit from the target population appears more then once in the sampling frame. A good coverage of the frame population can guarantee a low coverage errors.

Measurement error

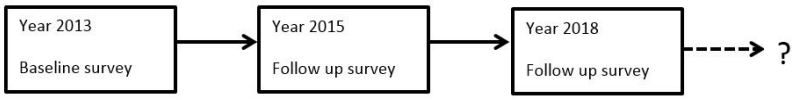
A measurement error is a lack of measurement precision due to weakness in the measurement instrument. Carefully planned and tested measurement instruments can reduce measurement errors.

3 Case Salo

First baseline survey was conducted in spring 2013 and the second follow up survey in spring 2015 (Figure 1). The data of the first survey study was gathered from the mailed questionnaire which was distributed in spring 2013. The questionnaires were distributed to everyone living in Salo and representing the following birth cohorts: 1961–1963, 1971–1973, 1981–1983, and 1991–1993. Of the study population, 2133 subjects completed and returned the questionnaire. The response rate was 29%. Subjects were asked to answer to the questions about their background, educational level and main type of activity, residency, willingness to relocate, use of services, health, social well-being and income. (Valaste, 2015)

The second follow-up survey in spring 2015 utilized the mailed questionnaire but also web survey. Target population was those who participated the first survey and also those who have moved to Salo after the first survey. 2287 subject completed the questionnaire. 1285 subjects participated in the baseline and follow up survey and 1002 subjects were new subjects. The response rate for the follow up survey was 29%.

Figure 1: Baseline survey and follow up surveys.



The frame population for both surveys was determined from the central population register. In baseline survey 51 respondents was not reached and 2 refused to answer. In the second survey 8 refused to answer.

In both surveys older cohorts were more active than the younger cohorts. Females responded more actively than males. Both baseline and follow up surveys included nonresponse. Post-stratification weights was constructed. Auxiliary information (gender and age group) were available and post-stratification weight was created for both surveys. Also a more sophisticated approaches was considered but unfortunately a limited information on the frame population was available.

4 Conclusion

Currently the third survey is planned. As earlier survey rounds, this also will have challenges especially nonresponse issues. How to improve the response rate? And in general, what is a lesson learned from the previous surveys?

References

Alwin, D. F. (1991). Research on survey quality. *Sociological Methods and Research* **20**, 3–29.

Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley, New Jersey.

de Leeuw, E. D., Hox, J. J. & Dillman, D. A. (2008). The cornerstones of survey research. In de Leeuw E. D., H. J. J. & D. D. A., eds., *The international handbook of survey methodology*. Erlbaum/Taylor & Francis, New York/London.

Groves, R. M. (1989). *Survey errors and survey costs*. John Wiley & Sons, New York.

Groves, R. M., Jr., F. J. F., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2009). *Survey methodology*. New Jersey: John Wiley & Sons.

Lehtonen, R. & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*. John Wiley & Sons, Chichester, 2nd edn.

Thompson, S. K. (2012). *Sampling*. New Jersey: John Wiley & Sons.

Valaste, M. (2015). Aineisto ja menetelmät. In M. Ylikännö & S. Kehusmaa, eds., *Muuttuva Salo. Kyselytutkimus äkillisen rakennemuutoksen alueen asukkaiden hyvinvoinnista*, chap. 3. Kela, Helsinki, pp. 13–16.

Ylikännö, M. & Kehusmaa, S. (2015). *Muuttuva Salo. Kyselytutkimus äkillisen rakennemuutoksen alueen asukkaiden hyvinvoinnista*. Kela.

Modelling of Survey Data

Olga Vasylyk[1] and Oksana Lagoda[2]

[1]Taras Shevchenko National University of Kyiv, Ukraine, e-mail: ovasylyk@univ.kiev.ua
[2]Kyiv National University of Technologies and Design, Ukraine, e-mail: oksala@ukr.net

Abstract

The fundamentals and assumptions of the model-based approach in sample surveys are presented, such methods as superpopulation modelling and Bayesian modelling are shortly described, and the outline of our study of modelling survey data is given.

Keywords: Bayesian modelling, model-based approach, sample surveys, survey data, superpopulation modelling

1 Introduction

Survey data may be viewed as the outcome of two random processes: the process generating the values in the finite population and the process selecting the sample data from the finite population values. There are three different approaches to sample design and analysis: the design-based approach, the model-based approach, and the design-based model-assisted approach. The advantages and disadvantages of all three approaches have been widely discussed in the literature in recent years.

Since we are interested in using models for survey data, we consider the model-based approach and the design-based model-assisted approach, and investigate how they are applied for solving different problem in sample surveys.

2 Model-based approach in survey sampling: main features

For a population U with N units, let Y = (y_1,…,y_N), where y_i is the set of survey variables for unit i, and let I = (I_1,…,I_N) denote a set of inclusion indicator variables, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included.

Model-based approach to survey sampling inference requires a model for the survey variables Y, which are treated as random (Little 2004). The model is then used to predict the nonsampled values of the population, and hence finite population quantities Q.(Y)

There are two major variants: superpopulation modeling and Bayesian modeling.

2.1 Superpopulation modeling

Analytic inference from survey data relates to the superpopulation model, but when the sample selection probabilities are correlated with the values of the model response variables even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process.

In superpopulation modeling the N population values of Y are assumed to be a random sample from a "superpopulation" and are assigned a probability distribution $p(Y|\theta)$ indexed by fixed parameters θ. Inferences are based on the joint distribution of Y and I.

2.2 Bayesian modeling

Bayesian modeling requires specification of a prior distribution p(Y) for the population values. Inferences for finite population quantities Q(Y) are based on the posterior predictive distribution $p(Y_{exc}|Y_{inc})$ of the nonsampled values Y_{exc}, given the sampled values Y_{inc}. In this case, model formulations do not involve the distribution for I, basing inferences only on the distribution of Y. This is justified when the sampling mechanism is "non-informative".

Sampling mechanism is said to be non-informative for a variable Y if the distribution of the sampled values of Y and the distribution of the non-sampled values of this variable are the same (Chambers, 2003). Or, in other words, the distribution of I given Y does not depend on the values of Y (Little, 2004).

2.3 Design-based model assisted approach in survey sampling

Design-based model-assisted approach attempts to combine the desirable features of design-based and model-based methods (Särndal *et al.,* 1992*)*.

3 Outline of the study

In our study of modelling survey data, the following main issues were considered:

1. Model-based approach in survey sampling

2. Population models

3. Design-based model assisted approach in survey sampling

4. Model-based and model-assisted estimation for domains and small areas

5. Model-based and model-assisted methods in dealing with nonresponses

6. Weighting and calibration

7. Modelling of complex survey data

8. Models for survey sampling with sensitive characteristics

As a result of the study, a textbook for Master students specializing in Statistics at Ukrainian universities will be prepared.

References

Chambers, R. (2003) *An introduction to model-based survey sampling*. Seminario international de estadistica en eusradi, No.42, 90 p.

Lehtonen, R. (2006) *The Role of Models in Model-Assisted and Model-Dependent Estimation for Domains and Small Areas.* In: Proceedings of the Workshop on Survey Sampling Theory and Methodology (Ventspils, Latvia) pp. 35-44.

Lehtonen, R. (2009) *Estimation for domains and small areas with design-based and model-based methods.* Lectures at the BNU Summer School on Survey Statistics (Kyiv, Ukraine).

Little, R. J. A. (2004) To model or not to model? Competing Modes of Inference for Finite Population Sampling. *The Journal of the Ameriacan Statistical Association*, Vol.99, No. 499. pp.546-556.

Montanari G.E. and Ranalli M.G. *Multiple and ridge model calibration for sample surveys.* Proceedings of the Workshop in Calibration and estimation in surveys, Ottawa, October 2007, Statistics Canada.

D. Pfeffermann (2010) Small Area Estimation: Basic Concepts, Models and Ongoing Research. *The Survey Statistician*, No.62, pp. 26-32.

D. Pfeffermann (2011) Modelling of complex syrvey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology,* Vol. 37, No. 2, pp. 115-136.

Rao, J.N.K. (2005) Interplay between sample survey theory and practice: an appraisal. *Survey Methodology*, Vol. 31, No. 2, pp. 117-138.

Särndal, C.-E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Särndal, C.-E., Lundstrom, S. (2005) *Estimation in Surveys with Nonresponse.* Wiley, 212 p.

Valliant, R., Dorfman, A.H. and Royall R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* John. Wiley & Sons, New York.

Vasylyk O., Ianevych T. Using models in sample surveys. *Applied Statistics. Actuarial and Financial Mathematics*, No. 1, pp. 72 – 79, - 2014 (Ukrainian)

Combining data from registers, surveys and 2011 Population and Housing Census to prepare database for 2021 register-based Population and Housing Census in Latvia

Pēteris Veģis

Central Statistical Bureau of Latvia, e-mail: peteris.vegis@csb.gov.lv

Abstract

In 2012 Cabinet of Ministers of Latvia made decision that 2021 Census will be based on administrative data or statistical sample survey data, if necessary. CSB of Latvia together with other interested institutions prepared a plan of activities for preparation and conduction of register-based Census in 2021 that was adopted by the Government in June 2015. In 2015 research was started on possibilities to use administrative data sources for Census needs and in parallel another activity to build a Social Statistics Data Warehouse, where all available administrative data will be stored, was started.

Paper will introduce with main results of research done mainly on indicators characterizing the economic activity and educational attainment of population and some future plans, including ongoing research on housing, household and family indicators.

Keywords: BNU2018, census, administrative data, register-based.

1 Introduction

Use of administrative data for the Population and Housing Census (hereinafter Census) needs was started in 2000 Census, when data from Population Register (hereinafter PR) was available. Address of usual residence was taken from PR and data comparisons were made with PR data.

Wider use of administrative data was organized in preparation, conduction and evaluation of results of 2011 Census. Three basic registers - PR, State Address Register and Real Estate State Cadastre Information System were used. Due to various reasons, e.g., non-response, unmet respondents or interviewer mistakes etc., during the Census it was not possible to obtain information on all persons registered with the PR. Therefore, to find out if the persons not surveyed can (cannot) be considered as the resident population of the Republic of Latvia, on 1 March 2011 the information of the State Revenue Service (hereinafter SRS), the State Employment Agency (hereinafter referred to as SEA), the State Social Insurance Agency (hereinafter SSIA), the National Health Service, the Ministry of Education and Science (hereinafter MES) and local government was used for estimates. This method was further developed for annual population estimates.

The preliminary research of administrative data sources for census needs was started with the Eurostat VIP.ADMIN grant project in 2015 when data on economic activity of population was checked. The next grant project was on educational attainment of population, but the third ongoing grant project is on household, family and housing data.

2 Economic activity of population

The following Census core topics in accordance with the Recommendations of the Conference of European Statisticians and Regulation (EC) No 763/2008 were studied and analysed - current activity status, main job, occupation, industry (branch of economic activity) and status in employment. Annual data are obtained since 01.01.2015.

Assessment of the availability and quality of administrative data in the databases of the SRS, SSIA, SEA, National Education Information System (hereinafter NEIS) of MES, as well as data received from Higher Education Establishments (hereinafter HEE) in comparison with the data from the Labour Force Survey (hereinafter LFS) and results of the 2011 Census was performed. In addition, some research on Farm Structure Survey data and data from the Rural Support Service and the Agricultural Data Centre was done.

2.1 Determination of the current activity status

Data necessary for the research of the economically active and not active population were available from several administrative data sources - information on taxpayers collected by the SRS, information on farm owners/users from the CSB Statistical Farm Register (hereinafter SFR), data from the SEA unemployed database, etc.

The research that was carried out formed a basis for the conclusion that the information included in administrative data sources can be used to acquire information on the economic activity status of persons necessary for the Programme of the 2021 Census. Work on the improvement of methodology that was started in 2015 is continued annually.

2.2 Determination of the economically active population

The CSB has access to administrative data sources – SRS data on tax payers and SEA data on unemployed persons – based on an interdepartmental agreement for acquiring information on the economical characteristics of the population foreseen in the Programme of the Census. Part of the required information (about farm owners/users) may be also found in the SFR of the CSB.

Some differences were identified between the criteria used by the SRS and the employment definition of the ILO. Information on working hours is not available for all employed persons since they are not recorded for certain employees. The database also includes persons under working age. It is not possible to determine whether a person has worked during the reference week for a part of employed persons as data for them are available only on quarterly or annual earnings. The database includes information on persons who have already ceased their working relationships during the reference period, as the SRS data reflect all tax payments, but they may also be made after the termination of employment relations. Data on sickness benefits are collected with a long-time lag.

As regards **determination of employment** based on administrative data only data related to remuneration for work are selected from SRS tables that also contain information on types of income other than wages and salaries. A person is employed if at least one of the tables reflects data on the respective person's hours worked, income, as well as a periodical leave, with maintaining one's job. Information about employment abroad is obtained from tables on income obtained abroad by a natural person (resident) and income obtained abroad by a natural person (seafarer) who is employed on a ship used for international transport.

The analysis carried out on **determination of main job** during the initial project shows that approximately 80% of Latvian employees earn their income only from one job, while 20% are employed in several jobs. Main part of these 20% have their record of work in one of jobs, if not than usual methodology was used - the hours worked, or the average monthly salary were compared between them. It was decided that November would be better reference time as December that is not typical in relation to economic activity because of Christmas and new Year.

The **occupational code of employed persons** has been included in the SRS database since 1 July 2013. Employers must submit information regarding the occupation of employees and the number of hours worked to the SRS. Overall, information on the occupation of employees corresponds with the ILO methodology and it covers about 83% of all employed persons. After comparing occupation codes at two-digit level, it can be concluded that administrative data often do not contain information on those occupations that are present in the sectors with the highest rates of shadow economy in Latvia (construction, trade, taxi services, etc.). Moreover, additional research should be carried out on those employed in agriculture, considering the seasonal nature of agricultural work. The compliance rate of occupation codes in LFS and SRS data at the two-digit level amounts to 61.5%, which is valued as satisfactory. The correspondence at the level of one-digit codes was higher - 68.9%. Development of methodology for obtaining the missing information on occupation of persons employed from regular statistical surveys performed by the CSB and imputation is done in 2017. Work is continued also considering the possibility to use data of various professional associations, etc.

A 4-digit code (NACE Rev. 2) of **Industry (branch of economic activity)** is included in the CSB Statistical Business Register (hereinafter SBR). Industry (branch of economic activity) in the main job for employed persons was determined by combining the SRS data with the SBR data. One of the problems – municipalities in the SRS reports use only one general public administration activity code (NACE 8411), but various businesses are under the supervision thereof. To solve the problem, a methodology for adding appropriate NACE code for the persons employed by enterprises of local governments was developed. During the study, the information on the industries (at section level) obtained from the administrative data was compared with the LFS data (employed persons living in private households). It was possible to use the Section code from SBR or SRS data for 99.4% of employees that had the occupation code. The compliance rate between estimates and LFS accounted for 72.8% that is satisfactory.

It is possible to determine **the status in employment** for those persons who are employed and who have a specific main job. In Latvia, the data are available from two sources: information on taxpayers gathered by the SRS or SFR on owners or users of farms. It is necessary to continue studying the issues related to the employment status by providing special attention to the improvement of the methodology for determining the employers and self-employed status.

SEA collects information on registered **unemployed persons**, job seekers and persons having other status. The CSB and the SEA have concluded inter-departmental agreement on the data receipt. Unemployed persons (99.8%) and job seekers (0.2%) registered with the SEA met the ILO definition of unemployed, while persons having other status did not. Some differences are found with ILO definitions. Only persons who have registered with the SEA are included in the data base. A maximum age limit is defined for the registered unemployed persons. Work on imputation of unemployed persons is started as there is a difference in comparison with the LFS based estimates.

2.3 Determination of economically not active population

The CSB has access to SSIA data on persons receiving state pensions. A comparison with the LFS data and data from 2011 Census suggests that recipients of pensions from administrative data sources are set within the limits of the confidence interval. Information on capital income receivers was obtained from the SRS. Individual data provided by the MES and HEE are used to obtain data about students.

3 Educational attainment

The main source of data is NEIS of MES and HEE as up to 2017 there was no higher education register. In addition, other data sources of professional data bases (Register of Medical Persons and Medical Support Persons, Register of Sailors, data about education of lawyers, bailiffs, notaries and their assistants etc.). Data about higher education obtained abroad is available just partly. Therefore, additional question about it was included in 2015 Population Micro Census and in 2018 External Migration Panel Survey.

Results from collecting data about the highest completed level of education from different administrative data sources (NEIS of MES, HEE, etc.) show that only 1,4% of all population aged 15 and over have no information about the educational attainment. Whereas, for 64,2% of all population aged 15 and over the data source of educational attainment is the 2011 Census database.

The classification of levels of education for the 2021 Census has changed comparing with the classification used in 2011 Census and the result can be observed in data – 7,3% of persons aged 15 and over have obtained higher education, but the new classification and Commission Implementing Regulation (EU) 2017/543 requires more detailed levels: short-cycle tertiary education or bachelor's or equivalent level or master's or equivalent level.

After the data correction and imputation, comparison of results with other data from CSB surveys, analysis of the quality of the data sources will be continued, possibility to define priorities for the data sources will be studied to improve the quality of educational attainment indicator annually.

The accumulative education database on highest education level according to ISCED-A (2011) classification was made and data is used as preprint in regular statistical sample surveys. Work will be continued on maintaining and annual updating of this database that contains information on persons' education from all available data sources and on identifying other possible additional data sources from which information about the level of education of the population could be obtained.

4 Conclusions

Research done is a good base for successful register-based Census 2021 as regards topics of economic activity and educational attainment. Nevertheless, work should be continued to improve data quality and data estimation and imputation methods. Identifying of new data sources, analysing their quality and following of legislative changes in existing ones also should be continued. Research is now continued with checking of administrative data availability on household, family and housing indicators.

New circumstances that declared address of persons will be used should be considered. Existing annual population estimates methodology will be changed. The aim is to develop a methodology based on data from administrative registers but does not directly use the 2011 Census data, which ages each year and becomes less informative about the current situation. As for 2021 Census declared usual residence will be used some problems are identified. Children under the age of 15 without adults are declared to be living in the place of residence. The proportion of single parents, especially lone fathers, increases in the family nucleus. In addition, household concept will be changed from housekeeping concept to household-dwelling concept. Influence of this change should be evaluated, too.

As all annual data is stored in the Social Statistics Data Warehouse it is necessary to improve functionality of it.

References

Conference of European Statisticians *Recommendations for the 2020 Censuses of population and Housing*, United Nations, New York and Geneva, 2015

Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses, Official Journal of the European Union, 13.08.2008

Commission Implementing Regulation (EU) 2017/543 of 22 March 2017 laying down rules for the application of Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns, Official Journal of the European Union, 23.03.2017

Rethinking Sampling for UK Business Surveys

Markus Gintas Šova[1], Nathan Joseph Calvin Price[2] and Gareth G. James[3]

[1]Office for National Statistics, UK, e-mail: markus.sova@ons.gov.uk
[2]Office of the Chief Government Statistician, Zanzibar, e-mail: Nathan.Price.OCGS@gmail.com
[3]Office for National Statistics, UK, e-mail: gareth.james@ons.gov.uk

Abstract

The introduction of electronic data collection allows us to rethink how surveys are designed. We present two subsampling methods for business surveys to allow survey modularisation.
Keywords: business surveys, survey modularisation, PRN subsampling

1 Introduction

Traditionally, business surveys have consisted of paper questionnaires dispatched by post. The questionnaire has one or more (sometimes many) questions which the business completes by hand and then returns to the National Statistical Institute (NSI). The concept of *one* survey is thus closely associated with the design of *one* questionnaire.

More recently, the advent of electronic data collection over the internet has raised interesting possibilities. For example, there are some questions which occur in more than one survey, albeit sometimes with subtle variations. If these can be harmonised, then the surveys could be integrated into a single survey with some core questions asked of all selected businesses, and the remaining questions modularised and given to subsamples of the survey. Alternatively, if a survey has more than one question, electronic data collection would allow the survey to be *de-integrated*, with all selected businesses being asked only one question. This would be particularly useful for large strata of small businesses because it would result in a fairer short-term distribution of response burden.

This paper examines how subsampling can be applied to allow survey integration by modularisation. The following section briefly describes how the Office for National Statistics (ONS, UK's NSI) currently implements rotational sampling. Two subsampling methodologies are presented in section 3, with some concluding remarks in section 4.

2 PRN Sampling

Since 1994, ONS has used the Inter-Departmental Business Register (IDBR) as the sampling frame for most of its business surveys. The IDBR's key sampling methodology

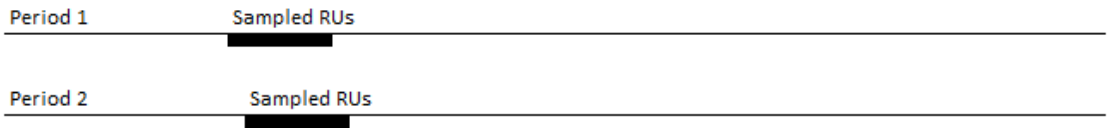
is stratified rotational sampling using permanent random numbers (PRNs). The IDBR's sampling units are called *reporting units* (RUs). Most RUs are enterprises, but some enterprises are split into two or more RUs for statistical purposes. When an RU is created on the IDBR, a PRN is generated whose value is permanently associated with the RU. From the point of view of a survey, each stratum consists of a set of RUs distributed along the PRN line (the set of all possible PRN values). For any stratum h a sample of size n_h is selected as the first n_h RUs on the PRN line whose PRN values are greater than or equal to a specified PRN start point, as shown in figure 1. Here the long thin line represents the PRN line, and the short thick line represents the selected sample which starts from the PRN start. If the PRN start is so large that there are insufficient RUs with a greater PRN value, the shortfall is made up for by selecting RUs from the beginning of the PRN line. Thus the PRN line is really a ring, but for clarity we shall continue to portray it as a line.

Figure 1: Representation of a PRN sample for one stratum

Because all the PRN values are independently generated from the same distribution, we have a simple random sample for the stratum.

After the sample has been selected, each stratum's PRN start is recalculated in preparation for the next survey period. This is done by moving the PRN start to the right so that a set number of RUs leave the sample. Over time the sample moves along the PRN line in a controlled way, as shown in figure 2. For further details see Ohlsson (1995).

Figure 2: Representation of a rotating PRN sample



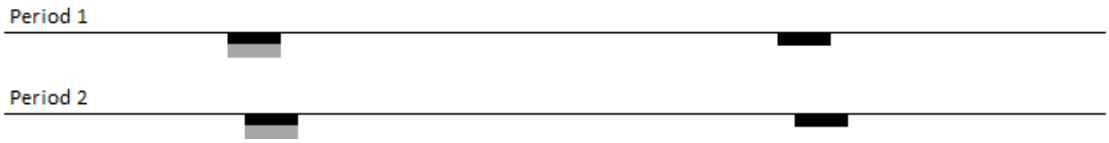
We note two important properties: Firstly, as the sample moves along the PRN line, every RU takes its turn at being sampled, ensuring a fair distribution of response burden over time. Secondly, the sample overlap between consecutive periods is controlled, giving a better variance of change than with independent sampling (see Lindblom, 2014).

3 Methodologies for Subsampling

ONS has started developing a new business register, called the *Statistical Business Register* (SBR), to replace the IDBR, giving the opportunity to specify one or more subsampling methodologies for survey modularisation. Ideally, we want the subsampling methodology to retain the key advantages of rotational sampling using PRNs, namely a fair distribution of burden over time and a controlled high overlap proportion between successive periods. James (2016) examines a number of subsampling methodologies. We present two of these as being particularly promising.

In figure 2, the sample is represented by a segment of the PRN line. Now consider several such segments of equal size, equally spaced along the PRN line. The set of these segments is the main sample. One (or more) of the segments is the subsample, whose RUs are to answer questions from a module. In principle, there can be as many distinct modules as there are segments. This method is more fully described and simulated by Price (2016). Figure 3 depicts how this Multiple Segment method works over time for a simple two segment example. Each segment is represented by a short thick black line. The subsample is represented by a thick grey line.

Figure 3: Representation of a rotating Multiple Segment sample and subsample



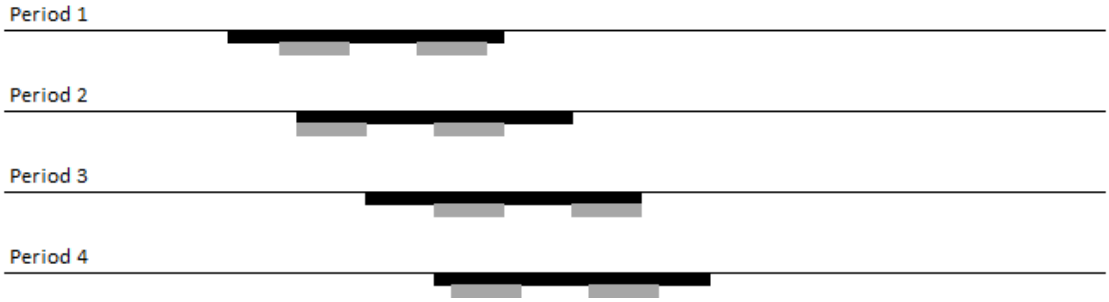
Each segment operates as a PRN sample. So the subsample has our desired properties of fair burden distribution and controlled high overlap. The equal sizes of the segments ensure that they move along the PRN line at the same rate (although with some sophistication this requirement can be eased). Otherwise, with segments moving at different speeds one segment will close in on another, resulting in RUs being rotated into the trailing segment only a few periods after being rotated out of the leading segment. Eventually the segments will collide, resulting in RUs being in the sample for twice as long as desired. The problems of encroaching and colliding segments can also be caused by births and deaths of RUs in the stratum. Therefore the Multiple Segment method requires the distances between segments to be monitored, and if they start getting too close then the rotation rates of individual segments may need to be temporarily amended. Price (2017a) offers and simulates solutions to this issue.

We now consider the main sample as a single segment on the PRN line, with a subsample represented by a subsegment. If this were moving at the same speed as the main sample then the overlap in the subsample between consecutive periods would be small, possibly even zero. Conversely, if the subsample were to move along the PRN sample at a slower rate than the main sample in order to maintain a high overlap, eventually the subsample would reach the start of the main sample and need to be repositioned at the end, resulting in a reduced subsample overlap at the time of the repositioning. This is shown in figure 4, where we see that there is a reduced subsample overlap between periods 2 and 3 and between periods 3 and 4. Furthermore, there are some RUs which do not get subsampled at all in this pass of the PRN line.

Suppose we now increase the number of subsamples whilst decreasing their size. Figure 5 shows two subsamples applied the scenario of figure 4. Only between periods 2 and 3 is the subsample overlap reduced. Further increasing the number of subsamples will lessen the reduction in the subsample overlap but will increase its frequency. We call this the *Stonehenge Method* due to it resembling a large stone being moved on wooden rollers.

Figure 4: The rotating subsample problem (the subsample moves more slowly)

Figure 5: Two rotating subsamples



4 Concluding Remarks

Of the two subsampling methods presented, that of Multiple Segments is elegant in its simplicity, with all subsamples having the desirable properties of a rotating PRN sample. However, the method requires monitoring with occasional intervention to ensure that no segment gets too close to another. The Stonehenge Method does not have this risk as it has only one segment. The subsample overlap is occasionally reduced, but this reduction can be lessened by having two or more subsamples. We have included both of these subsampling methods in the sampling specifications for the SBR.

References

James, G.G. (2016). *Options for sub-sampling.* ONS internal report.

Lindblom, A. (2014). On Precision in Estimates of Change over Time where Samples are Positively coordinated by Permanent Random Numbers. *Journal of Official Statistics* **30(4)**, 773 - 785.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers, in *Business Survey Methods*, eds. Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. & Kott, P.S. Wiley.

Price, N.J.C. (2016). *A PRN Sub-Sampling Specification.* ONS internal report.

Price, N.J.C. (2017a). *Multi-Segment Sampling: Making it Work.* ONS internal report.

Price, N.J.C. (2017b). *Overlap for GlassesCases-Subsampling.* ONS internal report.

Workshop Programme

Time	Monday	Tuesday	Wednesday	Thursday	Friday
	August 20	August 21	August 22	August 23	August 24
08.00 – 08.50		*Registration*			
08.50 – 09.00		*Opening*			
09.00 – 10.30		Anders Holmberg, Li Chun Zhang	Anders Holmberg, Li Chun Zhang	Anders Holmberg, Li Chun Zhang	Melike Oguz Alper (09.00 – 09.45) Natalia Bokun (09.50 – 10.35)
		Break	*Break*	*Break*	*Break*
10.50 – 12.50		Contributed Session I	Contributed Session III	Contributed Session IV	Contributed Session VI (10.55 – 11.40)
		Lunch	*Lunch*	*Lunch*	Closing of the workshop (11.45 – 13.00)
13.50 – 14.35		Carl-Erik Särndal	Manuela Lenk	Maciej Beręsewicz	
14.40 – 15.25		Carl-Erik Särndal	Danutė Krapavickaitė	Juris Breidaks	
		Break	*Break*	*Break*	
15.45 – 17.45		Contributed Session II	Special Contributed Session (15.45 – 16.15)	Contributed Session V	
17.00			City tour in Jelgava		
19.00	Workshop welcome party	BNU Steering Committee meeting		Workshop farewell party	

Session times are indicative here as the final workshop programme was not available when workshop proceedings were printed. Please look at the workshop programme available online or at the printed workshop programme for precise session times.

List of participants

Surname	Name	Country	Institution	e-mail
Ainâre	Ieva	Latvia	Central Statistical Bureau of Latvia	ieva.ainare@csb.gov.lv
Antanavičiūtė	Aušra	Lithuania	Statistics Lithuania	ausra.antanaviciute@stat.gov.lt
Balode	Ilze	Latvia	Ventspils University of Applied Sciences	ilze.balode@venta.lv
Bandarenka	Natallia	Belarus	State Institute of Management and Social Technologies of the Belarusian State University. Department of Financial Management.	bondnata@mail.ru
Behmane	Maranda	Latvia	Central Statistical Bureau of Latvia	maranda.behmane@csb.gov.lv
Beręsewicz	Maciej	Poland	Poznań University of Economics and Business	maciej.beresewicz@ue.poznan.pl
Bokun	Natallia	Belarus	Belarus State Economic University	nataliabokun@rambler.ru
Breidaks	Juris	Latvia	Central Statistical Bureau of Latvia	juris.breidaks@csb.gov.lv
Fisenko	Andris	Latvia	Bank of Latvia	Andris.Fisenko@bank.lv
Galdauskaitė	Dovilė	Lithuania	Statistics Lithuania	dovile.galdauskaite@stat.gov.lt
Gard	Rikard	Sweden	Statistics Sweden	Rikard.gard@gmail.com
Gobina	Inese	Latvia	Rīga Stradiņš University	Inese.Gobina@rsu.lv
Grēna	Zane	Latvia	Central Statistical Bureau of Latvia	Zane.Grena@csb.gov.lv
Halytsia	Yuliia	Ukraine	Taras Shevchenko National University of Kyiv	halytsya2013@gmail.com
Hellstrand	Julia	Finland	University of Helsinki/ Statistics Finland	julia.hellstrand@helsinki.fi
Holmberg	Anders	Norway	Statistics Norway	anders.holmberg@ssb.no
Honkala	Miika	Finland	Statistics Finland	miika.honkala@stat.fi
Ianevych	Tetiana	Ukraine	Taras Shevchenko National University of Kyiv	yata452@univ.kiev.ua
Janeiko	Julija	Lithuania	Vilnius Gediminas Technical University	julija.jan@gmail.com
Jozefa	Laura	Latvia	Central Statistical Bureau of Latvia	Laura.Jozefa@csb.gov.lv
Jukams	Janis	Latvia	Central Statistical Bureau of Latvia	janis.jukams@csb.gov.lv
Jurgelane-Kaldava	Inguna	Latvia	Riga Technical University	inguna.jurgelane@rtu.lv
Kantane	Inara	Latvia	Rīga Stradiņš University	inara.kantane@rsu.lv
Kirpu	Viktoria	Estonia	University of Tartu	viktoria.kirpu@gmail.com
Krapavickaitė	Danutė	Lithuania	Vilnius Gediminas Technical University	danute.krapavickaite@vgtu.lt
Laaksonen	Seppo	Finland	University of Helsinki	Seppo.Laaksonen@Helsinki.Fi
Laitila	Thomas	Sweden	Örebro University, Sweden	thomas.laitila@oru.se
Lapins	Janis	Latvia	Bank of Latvia	Janis.Lapins@bank.lv
Lenk	Manuela	Austria	Statistics Austria	manuela.lenk@statistik.gv.at
Liberts	Mārtiņš	Latvia	Central Statistical Bureau of Latvia	martins.liberts@csb.gov.lv

Surname	Name	Country	Institution	e-mail
Matveja	Zane	Latvia	Central Statistical Bureau of Latvia	zane.matveja@csb.gov.lv
Meldere	Sigita	Latvia	Central Statistical Bureau of Latvia	sigita.meldere@csb.gov.lv
Nikoluskina	Olesja	Latvia	Central Statistical Bureau of Latvia	olesja.nikoluskina@csb.gov.lv
Oguz-Alper	Melike	Norway	Statistics Norway	melike.oguz.alper@ssb.no
Pankūnas	Vytautas	Lithuania	Vilnius Gediminas Technical University	v.pankunas@gmail.com
Pavasare	Ruāna	Latvia	Central Statistical Bureau of Latvia	Ruana.Pavasare@csb.gov.lv
Pentala-Nikulainen	Oona	Finland	Helsinki University	oona.pentala@gmail.com
Pyankova	Anastasiya	Russia	Higher School of Economics	apyankova@hse.ru
Rozora	Iryna	Ukraine	Taras Shevchenko National University of Kyiv	irozora@bigmir.net
Rozora	Natalia	United Kingdom	Nielsen	rozora@ukr.net
Sakovich	Natallia	Belarus	Belarus State Economic University	Sakovich-n@rambler.ru
Sarioglo	Volodymyr	Ukraine	Ptoukha Institute for Demography and Social Studies	sarioglo@idss.org.ua
Särndal	Carl-Erik	Sweden	Statistics Sweden	carl.sarndal@telia.com
Serhiienko	Veronika	Ukraine	Taras Shevchenko National University of Kyiv	nichka_2009@ukr.net
Siliņš	Elvijs	Latvia	Central Statistical Bureau of Latvia	silinselvijs@gmail.com
Sinisalo	Alina	Finland	Natural Resources Institute Finland (Luke)	alina.sinisalo@luke.fi
Slickute-Sestokiene	Milda	Lithuania	Statistics Lithuania	milda.slickute@stat.gov.lt
Sloka	Biruta	Latvia	University of Latvia	Biruta.Sloka@lu.lv
Sokurova	Diana	Estonia	University of Tartu	diana.sokurova12@gmail.com
Sydorov	Mykola	Ukraine	Taras Shevchenko National University of Kyiv	ms123@ukr.net
Sydorova	Daryna	Ukraine	Taras Shevchenko National University of Kyiv	dashasydorova@gmail.com
Synohub	Svitlana	Ukraine	Nielsen Ukraine	Sveta.llana@gmail.com
Sõstra	Kaja	Estonia	Statistics Estonia	kaja.sostra@stat.ee
Šova	Markus Gintas	United Kingdom	Office for National Statistics	markus.sova@ons.gov.uk
Traat	Imbi	Estonia	University of Tartu	imbi.traat@ut.ee
Valaste	Maria	Finland	University of Helsinki	maria.valaste@helsinki.fi
Vasylyk	Olga	Ukraine	Taras Shevchenko National University of Kyiv	olva75@gmail.com
Vegis	Peteris	Latvia	Central Statistical Bureau of Latvia	Peteris.Vegis@csb.gov.lv
Voronova	Jelena	Latvia	Central Statistical Bureau of Latvia	jelena.voronova@csb.gov.lv
Zhang	Li-Chun	Norway	University of Southampton, Statistics Norway and University of Oslo	L.Zhang@soton.ac.uk
Zukula	Baiba	Latvia	Central Statistical Bureau of Latvia	baiba.zukula@csb.gov.lv

CENTRĀLĀ STATISTIKAS PĀRVALDE
CENTRAL STATISTICAL BUREAU OF LATVIA

Lāčplēša iela 1, Rīga, LV-1301, Latvia, Phone +371 67366850, fax +371 67830137
e-mail: csb@csb.gov.lv, http://www.csb.gov.lv

ISBN 978-9984-06-528-1 (PDF)
ISBN 978-9984-06-527-4 (Print)