EESTI
STATISTIKA

ISBN 978-9985-74-630-1

# BALTIC-NORDIC-UKRAINIAN WORKSHOP ON SURVEY STATISTICS 2022

August 23-26, 2022, Tartu, Estonia

Delta Centre, Narva Str. 18

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS
International Statistical Institute

Nordic Council of Ministers

EESTI
STATISTIKA

UNIVERSITY OF TARTU
1632

UNIVERSITY OF HELSINKI

# Baltic-Nordic-Ukrainian workshop on survey statistics 2022.

## Organizers

- Baltic-Nordic-Ukrainian (BNU) Network on Survey Statistics
- University of Tartu
- Statistics Estonia
- Estonian Statistical Society
- University of Helsinki

## Sponsors

- International Association of Survey Statisticians (IASS)
- Nordic Council of Ministers
- University of Tartu
- Statistics Estonia
- University of Helsinki

## Main themes

Nonprobability surveys, data integration, population statistics, applications in survey and official statistics

## Keynote Speakers

- Jean-François Beaumont (Statistics Canada)
- María del Mar Rueda (University of Granada)
- Carl-Erik Särndal (Sweden)
- Li-Chun Zhang (University of Southampton; Statistics Norway; University of Oslo)

Celebrating the 30th Anniversary of two milestone sources in survey statistics, "Model-Assisted Survey Sampling" by Carl-Erik Särndal, Bengt Swensson and Jan Wretman (Springer, 1992) and "Calibration Estimators in Survey Sampling" by Jean-Claude Deville and Carl-Erik Särndal (JASA 1992).

## Baltic-Nordic-Ukrainian (BNU) Network

The BNU network involves researchers, teachers, students and practicing statisticians in survey and official statistics from universities, national statistical agencies, research institutes and private organizations from Estonia, Finland, Latvia, Lithuania, Poland, Sweden and Ukraine.
https://wiki.helsinki.fi/display/BNU/Home

## Programme Committee

Risto Lehtonen, Finland (Chair)
Maciej Beręsewicz, Poland
Andrius Čiginas, Lithuania
Danutė Krapavickaitė, Lithuania
Thomas Laitila, Sweden
Mārtiņš Liberts, Latvia
Kaja Sõstra, Estonia
Imbi Traat, Estonia
Maria Valaste, Finland
Olga Vasylyk, Ukraine

## Organizing Committee

Imbi Traat, Estonia (Chair)
Kaur Lumiste, Estonia
Kaja Sõstra, Estonia
Maria Valaste, Finland
Mare Vähi, Estonia

Compiler: Kaja Sõstra
Designer: Nele Lumiste

**ISBN 978-9985-74-630-1**

# Preface

Dear Participants of the BNU Workshop on Survey Statistics 2022!

The Workshop is the latest in the long sequence of scientific and educational events organised annually since 1997 by the Baltic–Nordic–Ukrainian (BNU) Network on Survey Statistics. Since the pandemic has recently prevented on-site gatherings, we have been practicing virtual opportunities instead. The current workshop is the first hybrid event of the network. Most participants are on site and we offer free Zoom online connection for all registered participations. A total of about 90 participants have registered, most on site. The online connection offers a wide access to the scientific sessions. In addition, it enables participation of our partners from war-time Ukraine.

We have selected recent advances in methodologies for nonprobability surveys, data integration and population statistics as the main themes of the workshop. As the keynote lecturers for these areas we have invited Jean-François Beaumont of Statistics Canada, María del Mar Rueda of University of Granada and Li-Chun Zhang of University of Southampton, Statistics Norway and University of Oslo. Several invited lectures and contributed presentations will supplement nicely their presentations. We have also arranged a PC training session where the techniques for nonprobability data can be practiced.

In all our events, design-based model-assisted methods, calibration and related methods have traditionally played an important role. There are several invited and contributed presentations in the programme on the new aspects of the methods and applications.

The year 2022 is the jubilee year of two milestone publications in survey and official statistics, *Model-Assisted Survey Sampling* of Carl-Erik Särndal, Bengt Swensson and Jan Wretman (Springer 1992), and *Calibration Estimators in Survey Sampling* by Carl-Erik Särndal and Jean-Claude Deville (JASA 1992). We are very honored to have Carl-Erik Särndal participate in the workshop. His title for the keynote lecture is “Progress in survey science, yesterday, today, tomorrow“. A round table discussion is arranged after his talk.

Abstracts of all presentations are collected in this Proceedings, which is freely available on the web site of the BNU network.

The workshop is organized by the Baltic–Nordic–Ukrainian Network on Survey Statistics in cooperation with University of Tartu, University of Helsinki, Statistics Estonia and Estonian Statistical Society.

We are thankful to the International Association of Survey Statisticians (IASS) for sponsoring this event. Support of the Nordic Council of Ministers, University of Tartu, Statistics Estonia and University of Helsinki is greatly appreciated.

We wish to all participants fruitful sessions, new knowledge and useful contacts during the workshop.

On behalf of the organizers,

Risto Lehtonen, University of Helsinki

Imbi Traat, University of Tartu

## Content

## Programme

## Keynote papers



## Invited papers



## Contributed papers

| Scientific and Social Programme | | |
|---|---|---|
| **Monday 22 August** | | |
| Venue | | |
| | | Arrival |
| Delta Centre<br>Delta lobby<br>Narva Str. 18 | 16:00−19:30 | Registration desk open |
| Meeting point:<br>Delta lobby | 17:00 | Guided city walk -- Colours and tales of „Soup Town“, an oldest cozy slum near the very centre of Tartu |
| Delta Café | 19:30 | Welcome Party |
| **Tuesday 23 August** | | |
| Delta Centre<br>Room 1021 | 8:00−9:00<br>9:00−9:15 | Registration desk open<br>**Opening**<br>Imbi Traat (University of Tartu) Chair of Organizing Committee |
| Room 1021 | 9:15−10:00 | **Session 1** Invited lecture Chair Imbi Traat<br>Helle Visk (University of Tartu): Households and dwellings for register-based census: a graph-based approach |
| | 10:00−10:30 | Refreshments |
| Room 1021 | 10:30−12:00 | **Session 2** Contributed papers Chair Milda Šličkutė-Šeštokienė<br>Kristi Lehto (Statistics Estonia), Imbi Traat (University of Tartu): Mixed-mode Census survey in Estonia<br>Discussant: Iryna Rozora (Taras Shevchenko National Univesity of Kyiv) (online)<br>Evija Dundure (University of Latvia), Biruta Sloka (University of Latvia): Organisation of survey on financial literacy aspects related to voluntary pension savings and challenges in the realisation of the survey<br>Discussant: Anastasiia Volkova (University of Helsinki)<br>Kaja Sõstra (Statistics Estonia): Compilation of activity status and employment variables in Estonian register-based census<br>Discussant: Tomas Rudys (Statistics Lithuania) |
| Room 1021 | 12:00−12:45 | **Session 3** Invited lecture Chair Andrius Čiginas<br>Henri Luomaranta (Statistics Finland): Nowcasting with STS surveys: the Machine Learning approach |
| | 12:45−14:00 | Lunch |

| Scientific and Social Programme | | |
|---|---|---|
| **Tuesday 23 August (contd.)** | | |
| Room 1021 | 14:00−14:45 | **Session 4** Invited lecture Chair Maria Valaste<br>Baiba Zukula (Statistics Latvia), Jānis Jukāms (Statistics Latvia): The Population Census 2021 in Latvia |
| | 14:45−15:00 | Break |
| Room 1021 | 15:00−16:30 | **Session 5** Keynote lecture Chair Risto Lehtonen<br>Jean-François Beaumont (Statistics Canada): Making inferences from non-probability samples through data integration (online) |
| | 16:30−17:00 | Refreshments |
| Room 1021 | 17:00−17:45 | **Session 6** Invited lecture Chair Vilma Nekrašaitė-Liegė<br>Luis Castro-Martín (Andalusian School of Public Health), María del Mar Rueda (University of Granada), Carmen Sánchez-Cantalejo (Andalusian School of Public Health), Ramón Ferri-García (speaker) (University of Granada), Jorge Hidalgo Calderón (University of Granada), Andrés Cabrera (Andalusian School of Public Health): Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey |
| Room 1021 | 17:45−18:15 | Milda Šličkutė-Šeštokienė (Statistics Lithuania): Register-based Census, Lithuania<br>Discussant: Jānis Jukāms (Central Statistical Bureau of Latvia) |
| **Wednesday 24 August** | | |
| Room 1021 | 9:00−9:45 | **Session 7** Keynote lecture Chair Danutė Krapavickaitė<br>María del Mar Rueda (University of Granada), Ramón Ferri-García (speaker) (University of Granada): Non-probability surveys: a revision of methods for inference |
| | 9:45−10:15 | Refreshments |

| Scientific and Social Programme | | |
|---|---|---|
| **Wednesday 24 August (contd.)** | | |
| Room 1021 | 10:15−11:45 | **Session 8** Contributed papers Chair Krista Lagus<br><br>Ieva Burakauskaitė (Statistics Lithuania), Andrius Čiginas (Statistics Lithuania and Vilnius University): Non-probability sample integration in the survey of Lithuanian census<br><br>Discussant: Markus Gintas Šova (Office for National Statistics, UK) (online)<br><br>Vilma Nekrašaitė-Liegė (Statistics Lithuania and Vilnius Gediminas Technical University), Andrius Čiginas (Statistics Lithuania and Vilnius University), Danutė Krapavickaitė (Vilnius Gediminas Technical University): Usage of non-probability sample and scraped data to estimate proportions<br><br>Discussant: Tetiana Ianevych (Taras Shevchenko National University of Kyiv) (online)<br><br>Andrii Dzhoha (Taras Shevchenko National University of Kyiv), Iryna Rozora (Taras Shevchenko National Univesity of Kyiv): Sequential resource allocation under multi-armed bandit model with online clustering as side information (online)<br><br>Discussant: Helle Visk (University of Tartu) |
| Room 1021 | 11:45−12:30 | **Session 9** Invited lecture Chair Tomas Rudys<br><br>Beatriz Cobo (University of Granada), María del Mar Rueda (University of Granada), S. Pasadas (Institute for Advanced Social Studies, Spain), Luis Castro-Martín (Andalusian School of Public Health), Ramón Ferri-García (University of Granada): Estimation methods for integrating probability and non-probability survey samples (online) |
| | 12:30−13:45 | GROUP PHOTO<br>Lunch |
| Delta Centre Room 2004 | 13:45−15:45 | **Session 10** PC training session<br><br>Instructor: Luis Castro-Martín (University of Granada) Estimation methods for integrating probability and non-probability survey samples<br><br>NonProbEst The R package NonProbEst for estimation in non-probability surveys by M. Rueda, R. Ferri-García, L. Castro |
| | 15:45−16:15 | Refreshments |
| Meeting point: ERM lobby | 16:45 | Guided tour: ERM -- The Estonian National Museum |

| Scientific and Social Programme | | |
|---|---|---|
| **Thursday 25 August** | | |
| Room 1021 | 9:00−9:45 | **Session 11** Invited lecture Chair Baiba Zukula<br>Oleksandr Gladun (Ptoukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine): Censuses in Ukraine: past and perspective |
| | 9:45−10:15 | Refreshments |
| Room 1021 | 10:15−11:45 | **Session 12A** Contributed papers Chair Ieva Burakauskaitė<br>Ruslana Moskotina (Taras Shevchenko National University of Kyiv), Mykola Sydorov (Taras Shevchenko National University of Kyiv): The relationship between the number of reminders and the proportion of full responses in online surveys (online)<br>Discussant: Maria Litova, (University of Helsinki)<br>Ovidijus Baškauskas (Vilnius Gediminas Technical University), Danutė Krapavickaitė (Vilnius Gediminas Technical University): Estimation of Income Inequality Indicators<br>Discussant: Evija Dundure (University of Latvia)<br>Yana Bondarenko (Oles Honchar Dnipro National University): Statistical Analysis with Missing Data (online)<br>Discussant: Andris Fisenko (Bank of Latvia) |
| Room 1007 | 10:15−11:45 | **Session 12B** Contributed papers Chair Kaja Sõstra<br>Biruta Sloka (University of Latvia): Teaching aspects on surveys, on questionnaire design, on pilot survey, on sample selection and data collection using QuestionPro and obtained data analysis with SPSS<br>Discussant: Katja Laine (University of Helsinki)<br>Ilze Brante (University of Latvia) and Biruta Sloka (University of Latvia): Organisation and realisation of survey by different stakeholders (entrepreneurs, educators, students and public administrators) on work-based learning<br>Discussant: Tuukka Oikarinen (University of Helsinki)<br>Rita Vanaga (University of Latvia) and Biruta Sloka (University of Latvia): Challenges for real survey development and organisation<br>Discussant: Anastasiia Volkova (University of Helsinki) |

| Scientific and Social Programme | | |
|---|---|---|
| **Thursday 25 August (contd.)** | | |
| Room 1021 | 11:45−12:30 | **Session 13** Invited lecture Chair Vilma Nekrašaitė-Liegė<br>Jacek Białek (University of Lodz): Scanner data – advantages, challenges, and their processing in the *PriceIndices* package (online) |
| | 12:30−13:45 | Lunch |
| Room 1021 | 13:45−14:15 | EXTRA: Li-Chun Zhang (University of Southampton; Statistics Norway; University of Oslo): Introduction to the book "Graph Sampling" |
| Room 1021 | 14:15−15:00 | **Session 14** Invited lecture Chair Jānis Jukāms<br>Andrius Čiginas (Vilnius University): On design-based small area estimation |
| | 15:00−15:30 | Refreshments |
| Room 1021 | 15:30−16:15 | **Session 15** Keynote lecture Chair Imbi Traat<br>Carl-Erik Särndal (Sweden): Progress in survey science, yesterday, today, tomorrow |
| Room 1021 | 16:15−16:30 | Carl-Erik Särndal: Tribute to close associates no longer with us: Jean-Claude Deville, Jan Wretman |
| | 16:30−16:45 | Break |
| Room 1021 | 16:45−18:00 | **Session 16** Round Table<br>Topic: "Model-Assisted Estimation and Calibration Methods in Survey and Official Statistics Yesterday, Today and in the Future"<br>Moderator Risto Lehtonen<br>Panelists: Carl-Erik Särndal, Li-Chun Zhang, Imbi Traat, Danutė Krapavickaitė, Kaja Sõstra, Andrius Čiginas<br>The year 2022 is the 30th Anniversary of two milestone sources in survey statistics, *Model-Assisted Survey Sampling* by Carl-Erik Särndal, Bengt Swensson and Jan Wretman (Springer, 1992) and *Calibration Estimators in Survey Sampling* by Jean-Claude Deville and Carl-Erik Särndal (JASA, 1992).<br>The session is organized to acknowledge the major impact of these approaches on survey statistics methodology, and especially on official statistics practice. The aim is to share experiences and future views on the methodologies in the various operating environments of the panelists and the audience. |
| | 19:00−21:00 | Farewell Party – Restoran Vilde ja Vine |

| Scientific and Social Programme | | |
|---|---|---|
| **Friday 26 August** | | |
| Room 1021 | 9:00–9:45 | **Session 17** Keynote lecture Chair Biruta Sloka<br>Li-Chun Zhang (University of Southampton; Statistics Norway; University of Oslo): Transforming population statistics: From census to fractional counting |
| | 9:45−10:15 | Refreshments |
| Room 1021 | 10:15−11:45 | **Session 18A** Contributed papers Chair Risto Lehtonen<br>Donatas Šlevinskas (Vilnius Gediminas Technical University): StatVillage homework<br>Discussant: Ilze Brante (University of Latvia)<br>Sigita Purona-Sida (University of Latvia, Central Statistical Bureau of Latvia): Challenges and solutions to maintaining survey response rates in Social Statistics<br>Discussant: Milda Šličkutė-Šeštokienė (Statistics Lithuania)<br>Risto Lehtonen (University of Helsinki), Ari Veijanen (University of Helsinki): Model calibration and MRP methods for small area estimation: an empirical comparison<br>Discussant: Ovidijus Baškauskas (Vilnius Gediminas Technical University) |
| Room 1007 | 10:15−11:15 | **Session 18B** Contributed papers Chair Maria Valaste<br>Andris Fisenko (Latvijas Banka): Using administrative data to clarify and adjust economic survey data: The case of Latvian HFCS<br>Discussant: Olga Vasylyk (National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”) (online)<br>Katja Laine (University of Helsinki), Maria Litova (University of Helsinki), Tuukka Oikarinen (University of Helsinki): Crowdsourcing social wellbeing non-probability survey<br>Discussant: Rita Vanaga (University of Latvia) |
| Room 1021 | 11:45−12:00 | Closing |

# MAKING INFERENCES FROM NON-PROBABILITY SAMPLES THROUGH DATA INTEGRATION

**Jean-François Beaumont**[1]

[1] Statistics Canada, Canada
e-mail: jean-francois.beaumont@statcan.gc.ca

## Abstract

For several decades, national statistical agencies around the world have been using probability surveys as their preferred tool to meet information needs about a population of interest. In the last few years, there has been a wind of change and other data sources are being increasingly explored. Five key factors are behind this trend: the decline in response rates in probability surveys, the high cost of data collection, the increased burden on respondents, the desire for access to "real-time" statistics, and the proliferation of non-probability data sources.

In this presentation, I will provide a brief overview of the history of probability surveys and explain why there is a wind of change. Non-probability surveys are not a panacea. They typically suffer from selection/coverage bias and may be fraught with measurement errors. I will illustrate the selection bias through data of an online volunteer-based survey and two probability surveys conducted by Statistics Canada.

The main question that will be addressed in this presentation is: How to leverage data from a non-probability source while preserving a valid statistical inference framework and an acceptable quality? Approaches that address this question typically involve the integration of data from probability and non-probability sources. I will review some data integration methods, including dual frame weighting (e.g., Kim and Tam, 2021), statistical matching (e.g., Rivers, 2007), inverse probability weighting (e.g., Chen, Li and Wu, 2020) and small area estimation (e.g., Rao and Molina, 2015). I will discuss the characteristics of each approach, including their benefits and limitations, and present a few empirical results. I will conclude with some additional thoughts on the future of probability and non-probability surveys. A significant portion of this presentation is based on Beaumont (2020) and empirical results in Beaumont, Bosa, Brennan, Charlebois and Chu (2022).

**Keywords:** calibration, dual frame weighting, inverse probability weighting, small area estimation, statistical matching.

## References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1-28.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2022). Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. Presentation at the 2022 Morris Hansen Memorial Lecture, https://washstat.org/hansen/2022Beaumont.pdf, March 1st, 2022.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Kim, J. K., and Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89, 382-401.

Rao, J.N.K., and Molina, I. (2015). *Small area estimation*. Second Edition, Wiley, Hoboken, NJ.

Rivers, D. (2007). Sampling from web surveys. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

# NON-PROBABILITY SURVEYS: A REVISION OF METHODS FOR INFERENCE

**M. Rueda[1] and R. Ferri-García[2] and L. Castro-Martín[3]**

[1] University of Granada, Spain
e-mail: mrueda@ugr.es

[2] University of Granada, Spain
e-mail: rferri@ugr.es

[3] Andalusian School of Public Health, Spain
e-mail: luiscastro193@gmail.com

## Abstract

Since the early XX$^{th}$ century, probability samples have been the standard procedure for the obtention of information from a population of interest, in those cases where a census was not feasible. However, the decreasing response rates and growing costs of traditional survey methods, which can guarantee probability sampling schemes to a certain extent, have favored the rise of non-probability samples to obtain information from a population of interest. Non-probability samples are often obtained from online surveys or using procedures aimed to collect large amounts of passive data. Despite their lower cost and immediacy, these samples entail a number of drawbacks, especially regarding their selection bias. This bias can be mitigated using design-based and model-based methods developed in literature. Design-based methods, such as Propensity Score Adjustment and Kernel Weighting, aim to estimate the probability of an individual of the population of being included in the non-probability sample, and use them to obtain weights or to match individuals with similar characteristics in an available probability sample. Model-based methods, such as Statistical Matching, also known as Mass Imputation, model-assisted or model-calibrated estimators, aim to predict the value of the target variable in a probability sample or a complete census of the population where the target variable has not been measured. In this session, we describe and compare the available methods for inference in non-probability samples, and explain how Machine Learning techniques could boost these methods. Finally, we give some recommendations on further research lines regarding estimation from non-probability samples.

**Keywords:** Non-probability sampling, Kernel Weighting, Propensity Score Adjustment, Statistical Matching, model-based estimators.

# ESTIMATING IN NON-PROBABILITY SURVEYS WITH R

**M. Rueda[1] and R. Ferri-García[2] and L. Castro-Martín[3]**

[1] University of Granada, Spain
e-mail: mrueda@ugr.es

[2] University of Granada, Spain
e-mail: rferri@ugr.es

[3] Andalusian School of Public Health, Spain
e-mail: luiscastro193@gmail.com

## Abstract

The convenience of non-probability surveys, often in the form of online surveys, is widely known. They allow for an easy, cheap and efficient way of recollecting data. However, researchers are also aware of the important bias problems which are associated with these kinds of methodologies. A wide variety of methods have been proposed in the last years in order to reduce the bias which such surveys imply. In practice, however, the application of those methods may become a difficult task. This course will show how to correctly apply them with R using the NonProbEst package, which includes state-of-the-art techniques and machine learning models. It includes basic concepts as well as important considerations in order to obtain optimal results, such as the application of hyperparameter optimization processes. A diverse set of alternatives will be covered so it can be adapted to any context, including Propensity Score Adjusting (also known as Propensity Weighting), Statistical Matching and Model Based approaches (also known as Mass Imputation), Model Assisted, Model Calibrated and variance estimation via Jackknife. Even though previous knowledge of the programming language R is required, the contents are presented in a clear and easy to apply way.

**Keywords:** Non-probability sampling, Propensity Score Adjustment, Statistical Matching, model-based estimators, Propensity Weighting, Mass Imputation

# PROGRESS IN SURVEY SCIENCE AND PRACTICE YESTERDAY – TODAY – TOMORROW

**Carl-Erik Särndal**[1]

[1] e-mail: carl.sarndal@telia.com

## Abstract

To realize progress is a goal in every scientific field. This holds for survey science as well, in particular for its application in National Statistical Offices (NSO:s). In well over one hundred years of existence, the field can point to several examples of periods and circumstances that triggered significant progress.

At the present time, frequently mentioned challenges for progress are:

- nonresponse is a serious threat to the validity of sample survey statistics;
- high cost is affecting the traditional sample surveys, where probability sampling have been a prominent feature;
- use of low cost alternative data sources, big data and others, is tempting, but raises unresolved questions about quality and validity of survey statistics produced in such ways.

This presentation offers some personal thoughts on progress realized in the field – especially in the last fifty years - as well as thoughts on how this may prepare for future progress, in the light of the challenges mentioned.

**Keywords:** progress, valid statistics, high survey cost.

## References

Laudan, L. (1977) Progress and its Problems. Towards a Theory of Scientific Growth. Los Angeles: University of California Press.

# INTRODUCTION TO THE BOOK GRAPH SAMPLING

**Li-Chun Zhang** [1]

[1] University of Southampton, UK
e-mail: L.Zhang@soton.ac.uk

## Abstract

Finite population sampling has found numerous applications in the past century. The validity of sampling inference of real populations, hence its universal applicability, derives from the known sampling probabilities associated with the sample, "irrespectively of the unknown properties of the target population studied" (Neyman, 1934).

A valued graph is a more powerful representation, which allows one to incorporate the connections among the units in addition. The underlying structure is a graph given as a finite collection of nodes (for units/entities) and edges (for connections). Attaching measures to the nodes or edges or both yields a valued graph. Many technological, socio-economic and biological phenomena exhibit a graph structure that may be the central interest of study, or the edges may effectively provide access to those nodes that are the primary targets. Either way, graph sampling is a statistical approach to study real graphs. Just like finite population sampling, graph sampling is universally applicable based on exploring the variation over all possible subgraphs (i.e. sample graphs), which can be taken from the given population graph, according to a specified method of probability sampling.

On the one hand, graph sampling encompasses finite population sampling in the sense that, apart from element and cluster sampling, all the so-called "unconventional" techniques that make use of the connections between the relevant units can be more effectively studied as special cases of graph sampling, such as indirect, network, adaptive cluster, line-intercept or spatial sampling. On the other hand, graph sampling theory yields a rigorous approach to genuine graph problems, where the interest of estimation is given directly as graph parameters, allowing one to devise and make use of various probabilistic breadth- or depth-first non-exhaustive graph traversal algorithms.

The recently published book Graph Sampling establishes a theoretical framework that unifies the key elements in the pioneering works of Birnbaum & Sirken, O. Frank and S. Thompson, and develops new general sampling strategies including when one does not observe all the possible ways by which a given subgraph can be sampled. There is a wide range of application areas, such as social networks, internet, epidemiological and biomedical studies, graph-based machine learning, environmental and spatial statistics.

# INTRODUCTION TO THE BOOK GRAPH SAMPLING

**Li-Chun Zhang** [1]

[1] University of Southampton, UK
e-mail: L.Zhang@soton.ac.uk

## Abstract

Conducting a census is the oldest method of producing detailed population statistics. However, the high cost and--hence--low frequency of census has become increasingly difficult to justify, so that many countries are currently developing alternative ways to produce census-like population statistics annually. Here we shall focus on the problem of estimating population counts (with basic demographics) at detailed locations.

In the first part of the talk, we summarise the two alternative frameworks for producing census-like population statistics, where direct counting from the Central Population Register is either unavailable or deemed unreliable. First, the *Register Survey* approach is based on combining available population registers and *large* coverage surveys. Replacing the census enumeration list by an integrated register compiled from existing data is the key to cost reduction, as is the case in Israel 2008 and 2021. The second *Fractional Counting* approach aims to estimate the population counts directly in a manner that can be characterised as register-based. Starting from an extended population register (EPR) with negligible under-coverage errors, each EPR person is assigned, successively, a probability of belonging to the target population, such as practised in Estonia and Latvia, and a vector of probabilities of living at one of the known addresses. Not only can it further reduce the cost, but the framework enables a conceptual shift away from population statistics envisaged as the results of pigeonhole classification. For instance, instead of 'how many people have a permanent address at a given place', one can make statistics about 'how many people can be expected at the given place and, in addition, where else they can be expected'.

For any statistics that require model-based methods of estimation, whether the model is trained real-time or fixed in advance, there is a question of valid uncertainty assessment. In the second part of the talk, we provide an introduction to *design-based inference methods for model-based statistics*. A relevant starting point in the present context is that the register-based census-like statistics have sufficient quality, such that they could replace sample survey or census altogether in producing statistics, as is the case in the Nordic and Baltic countries. Nevertheless, the register-based statistics are not without errors, which motivates the central concept of *audit sampling inference* (Zhang, 2021), "Wherever the goal of survey sampling is to produce a point estimate of some target parameter of a given finite population, audit sampling aims not to estimate the target parameter itself, but some chosen accuracy measure of any given estimator of the target parameter, which may be potentially biased due to failure of the underlying assumptions or other favourable conditions that are necessary." In particular, we shall explain how the *total and individual errors* of either completely register-based statistics or sample-based model estimators can be evaluated with respect to the known audit sampling distribution.

# SCANNER DATA – ADVANTAGES, CHALLENGES, AND THEIR PROCESSING IN THE *PriceIndices* PACKAGE

**Jacek Białek** [1]

[1] University of Lodz, Poland

## Abstract

Scanner data can be obtained from a wide variety of retailers (supermarkets, home electronics, Internet shops, etc.) and provide information at the level of the barcode, i.e. the Global Trade Item Number (GTIN) or its European version: European Article Number (EAN). One of advantages of using scanner data in the Consumer Price Index (CPI) measurement is the fact that they contain complete transaction information, i.e. prices and quantities for every sold item. One of new challenges connected with scanner data is the choice of the index formula which should be able to reduce the chain drift bias and the substitution bias. The main purpose of the presentation is to discuss a broad spectrum of benefits and challenges related to the use of scanner data in measuring inflation.

An additional purpose of the work is to present the utility of the *PriceIndices* R-package in the field of analysing the dynamics of scanner prices. The presentation of this R package is divided into the following areas: scanner data preparing, data set characteristics, bilateral index calculations, multilateral index calculations, extensions of multilateral indices, aggregation of index results, and comparison of price indices. In particular, to demonstrate the package, the entire data processing (from the preparation of the row scanner data to the calculation of the price indices) and some empirical study will be performed.

**Keywords:** scanner data, scanner data classification, product matching, price indices, multilateral indices, *PriceIndices* package e.

# REWEIGHTING WITH MACHINE LEARNING TECHNIQUES IN PANEL SURVEYS. APPLICATION TO THE HEALTH CARE AND SOCIAL SURVEY

**L. Castro-Martín[1] and M. M. Rueda[2] and C. Sánchez-Cantalejo[3] and R. Ferri-García[4] and J. Hidalgo Calderón[5] and A. Cabrera[6]**

[1] Andalusian School of Public Health, Spain
e-mail: luiscastro193@gmail.com

[2] University of Granada, Spain
e-mail: mrueda@ugr.es

[3] Andalusian School of Public Health, Spain
e-mail: carmen.sanchezcantalejo.easp@juntadeandalucia.es

[4] University of Granada, Spain
e-mail: rferri@ugr.es

[5] University of Granada, Spain
e-mail: jorgehcal@ugr.es

[6] Andalusian School of Public Health, Spain
e-mail: andres.cabrera.easp@juntadeandalucia.es

## Abstract

The rapid evolution of COVID-19 required tools to perform a fast and efficient evaluation of the situation. Healthcare services around the world relied on surveys to fill those information gaps regarding the social, economic and health impacts of the disease. Those surveys enabled decision makers to take preventive and protective measures, especially among the most vulnerable population.

The Health Care and Social Survey (ESSOC, by its initials in Spanish), whose protocol was developed in Sánchez-Cantalejo et al. (2021), is a research project that arose from the necessity of information previously described. Its objective is to provide reliable and specific information about the impact of COVID-19 in Andalusia over time in several variables that may be useful in decision-making for controlling the consequences of the pandemic. The survey covers the evolution of health, socioeconomic, psychosocial, behavioral, labor, environmental and clinical variables, both in the general population and among the most vulnerable ones. The study integrates data from various sources, based on surveys, population statistics and official registers (clinical, epidemiological and environmental). Regarding the surveys, they have followed an overlapping panel design, using different adjustment methods to account for non-response and attrition. These methods, including the role of Machine Learning algorithms in their application, will be developed in this session.

**Keywords:** COVID-19, panel surveys, overlapping design, non-response, Machine Learning

## References

Sánchez-Cantalejo, C., Rueda, M. M., Saez, M., Enrique, I., Ferri-García, R., De La Fuente, M., Castro-Martín, L., … & Cabrera-León, A. (2021) Impact of COVID-19 on the health of the general and more vulnerable population and its determinants: Health care and social survey-ESSOC, study protocol. *International Journal of Environmental Research and Public Health*, **18**, 8120.

# ON DESIGN-BASED SMALL AREA ESTIMATION

**A. Čiginas**[1,2]

[1] Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@stat.gov.lt

[2] Vilnius University, Lithuania
e-mail: andrius.ciginas@mif.vu.lt

## Abstract

Small area estimation methods are used in surveys, where sample sizes are too small to get reliable direct estimates of parameters in some population domains. We consider design-based linear combinations of direct and synthetic estimators of domain means and propose a two-step procedure to approach the optimal combination. We construct the mean square error estimator suitable for this and any other linear composition that estimates the optimal one. We consider also the case of small true domain proportions and propose a new design-based composite estimator to estimate them. We apply the constructed estimators to the data of the Lithuanian Labor Force Survey and the statistical survey of the Lithuanian census and compare them with empirical best linear unbiased predictors and some other composite estimators.

**Keywords:** composite estimator, synthetic estimator, mean square error, bias.

# Estimation methods for integrating probability and non-probability survey samples

**B. Cobo[1], M. Rueda[1], S. Pasadas[2], L. Castro[3] and R. Ferri[1]**

[1] University of Granada, Spain
e-mail: beacr@ugr.es, mrueda@ugr.es, rferri@ugr.es

[2] Institute for Advanced Social Studies, Spain
e-mail: spasadas@iesa.csic.es

[3] Andalusian School of Public Health, Spain
e-mail: luiscastro193@gmail.com

## Abstract

In recent years, different methods have been provided to combine information from multiple data sources. We focus on the case of probability and non-probability samples that share the same questionnaire, combining both to maximize the efficiency of the estimates with the help of machine learning methods. We develop a new estimation method to integrate data from probability and non-probability samples, we evaluate the efficiency of the resulting estimates by comparing them with other strategies that have been used before. The application of this method to the second wave of the Survey on the impact of the COVID-19 pandemic in Spain allows us to conclude that the estimation method we propose is the best option to reduce the biases observed in our data.

**Keywords:** non-probabilistic surveys, machine learning techniques, propensity score matching, survey sampling.

# CENSUSES IN UKRAINE: PAST AND PERSPECTIVE

**O. Gladun**

Ptoukha Institute for Demography and Social Studies
of the National Academy of Sciences of Ukraine, Ukraine
e-mail: gladun.ua@gmail.com

**Abstract**

The paper provides a brief overview of the history of population censuses in modern Ukraine before 1991. The situation with the population censuses after Ukraine gained independence and the peculiarities of conducting them in the future are described.

**Keywords:** population census, Ukraine.

## 1 Population censuses on the territory of modern Ukraine before 1991

Census data are a photograph of the demo-social state of the country. A comparison of census data over a certain period of time makes it possible to assess the direction and speed of the country's movement on various aspects of life. This applies not only to demographic aspects (births, migration, age and sex structure), but also social (marriage, family) and economic (accommodation on the territory, living conditions, sources of livelihoods, etc.). On the other hand, the history of the country determines the peculiarities of the census.

Throughout its history, the territory of modern Ukraine has been a part of different countries. Before the First World War, the territory of Ukraine was a part of the Russian Empire and Austria-Hungary. Since the middle of the 19th century, six censuses have been conducted in Austria-Hungary, and only one in the Russian Empire. In the period between the First and Second World Wars, the territory of Ukraine was a part of five countries (Poland, Romania, Czechoslovakia, Hungary and the USSR). Now, according to the current administrative division, Ukraine consists of 25 oblasts (regions). The table 1 shows the data on population censuses on the modern territory of Ukraine conducted by different countries.

It should be noted that the general principles of modern censuses were formulated at the Eighth Session of the International Statistical Congress in 1872. Census programs and the procedure for conducting them are constantly being improved, but the principles laid down 150 years ago are still relevant today.

In the USSR demography and demographic processes were always considered in the political aspect. Apparently only the 1920 and 1926 censuses did not have any political influence. The 1920 census failed to be conducted throughout Ukraine, which was a part of the former Russian Empire, due to hostilities. In my opinion, the 1926 census was one of the best in the history of the USSR. This was due to the lack of political pressure on statistical bodies, scientific approach to the entire census process, open analysis of the quality of results and a large number of publications.

The next census was scheduled for 1933, but was postponed first to 1935 and later to 1936. The census finally took place in 1937. The reason for the postponement was the Soviet government's desire to hide the after-effects of the demographic catastrophe caused by collectivization and famine of 1932–1933. Despite postponing the census for three years, it could not but reflect the effects of the famine. The population of Ukraine in 1937 turned out to be 428 thousand less than in 1926. It was necessary either to publish the results of the census (but then they would have to be explained) or to

declare them incorrect. Eight months after the census, the Council of People's Commissars of the USSR adopted a resolution declaring the results of the census to be "defective," and the census itself methodologically incorrect. The materials of the development were classified, and the organizers of the census were repressed (shot or sent into exile). Ukrainian scientists[1] were also repressed. For many years after, the 1937 census was considered to be a classic example of a "defective census." The general results of this census were first published only in 1991. According to the modern experts, the 1937 census met the standards of conducting censuses.

Table 1. Population censuses on the territory of modern Ukraine

| Country that conducted the census | Number of censuses | Years | Number of oblasts within modern borders |
|---|---|---|---|
| Before the First World War | | | |
| Austria-Hungary | 6 | *1857, 1869*, 1880, 1890, 1900, 1910 | 5 |
| Russian Empire | 1 | 1897 | 20 |
| Between the First and Second World Wars | | | |
| Czechoslovakia | 2 | 1921, 1930 | 1* |
| Hungary | 1 | 1941 | |
| Poland | 2 | 1921, 1931 | 5 |
| Romania | 1 | 1930 | 1 |
| USSR (1917–1939) | 4 | 1920, 1926, 1937, 1939 | 18 (14**) |
| After the Second World War | | | |
| USSR (1945–1991) | 4 | 1959, 1970, 1979, 1989 | 25 |
| Ukraine | 1 | 2001 | 25 |

***Note****. * – Zakarpattya oblast; ** – 1920.*
*Source: compiled by the author.*

In 1939 a new census was conducted. Unlik**e** the "defective" 1937 census, the results of the 1939 census were recognized by the government. Before the war, only brief results of the census were published in the press; the main results were published only in 1992. Although abortion was banned in the USSR in late 1936 to compensate for the loss of population due to the famine of the 1930s, the 1939 census also showed the effects of famine. In order to conceal the effects of the famine, statistical bodies, under pressure from the authorities, resorted to the deliberately inflated data. Thus, the population of Ukraine was overstated by more than 800 thousand people (Rudnytskyi at al. 2015).

In the first years after the end of the Second World War, most European countries conducted population censuses. In the USSR, the first post-war census took place only 14 years later, in 1959. The main reason for this delay was the reluctance of the Soviet authorities to show the true extent of human losses, as well as financial difficulties.

In the 1959 and 1970 censuses, the Soviet government tried to "finally resolve" the issue of the Crimean Tatars. In 1944, the Crimean Tatars were deported from the Crimea to the Soviet republics of Central Asia, most of them to Uzbekistan. When publishing the ethnic composition of the population of the USSR and the republics, the data on the Crimean Tatars were absent. During the census, they were included in other ethnic groups, mostly Tatars. In response, the Crimean Tatars in 1964, 1971 and 1973–1974 conducted a "self-census of the Crimean Tatar people." This is a unique case in the history of censuses.

## 2 Censuses after 1991

Since gaining independence in 1991, only one census has been conducted in Ukraine, in 2001. The second census has not yet been conducted. A characteristic feature of population censuses in the years

[1] In 1938, the world's first specialized Institute for Demography of the Ukrainian Academy of Sciences was closed (since 1934 it had been called the Institute for Demography and Sanitary Statistics). Many employees were repressed, including its permanent director M.V. Ptoukha. The Institute was re-established in 2002 as the Institute for Demography and Social Studies. In 2007, the institute was named after M.V. Ptoukha.

of independence is their constant rescheduling. If the first census was postponed only once, the date of the second census has been changed five times (table 2).

Table 2. Planned and actual years of the population census of Ukraine

| Census order | Planned years | Year of the census |
|---|---|---|
| The first | 1999, 2001 | 2001 |
| The second | 2011, 2012, 2013, 2016, 2020, 2023 | ? |
| *The third* | *2030* | *???* |

*Source: compiled by the author.*

All postponements were justified by the lack of money to finance the work. The closest to success the State Statistics Service, which in Ukraine is responsible for conducting censuses, was in 2012. Tens of millions of forms and instructions were printed, and about two hundred thousand census staff were recruited and trained. All that remained was to carry out field work. But at the last moment the census was postponed to 2013.

The lack of a census for more than twenty years leads to growing discrepancies between population estimates and the actual population in a given area. The urgency of this problem increases with the reduction of the administrative level of the territory: state – oblast (region) – raion (district) – territorial community. Inter-budget transfers (grants, subsidies, subventions), investments in the development of a certain territory depend on the population and its age and sex structure. With the use of demographic indicators, plans for the development of administrative-territorial units are developed. In 2020, the administrative-territorial reform was carried out: instead of 490, 140 districts were formed, which, in turn, consist of 1,469 territorial communities. This required a significant number of calculations. In half of the cases, the new districts had territorial boundaries that did not coincide with the boundaries of the old districts. There were no territorial communities at all. That led to discrepancies between the data of statistical bodies and the data of district administrations and territorial communities.

The lack of the reliable basic information reduces the quality of demographic forecasts.

In Ukraine, the registration of the natural population movement is established at a high level. All births are registered, otherwise the child will not be able to "go through life". There may be problems with the registration of deaths in the remote rural areas, in the event of homicides or accidents, but they do not have a significant impact on the quality of demographic data. In Ukraine as in many countries, accounting for migration is a bigger problem. After the census, the level of shot-term labor migration in Ukraine has increased. Some migrant workers remained for permanent residence in other countries, but they are registered in Ukraine. At the beginning of the Russian-Ukrainian war in 2014–2015, IDPs (internally displaced persons) appeared in Ukraine, and some people from the zone of active hostilities left for the border countries (including the Russian Federation and Belarus). Due to the occupation of the part of Donbass, it is unknown how many of them have returned to Ukraine.

The full-scale invasion of the Russian Federation into Ukraine on February 24, 2022 brought the problem of IDPs and forced external migrants to a new level of quality. This makes the need to record demographic events and conduct a census an urgent state problem

Even before the full-scale invasion of Russia into Ukraine, the problem of conducting a census acquired a state character. This was also due to the fact that as a result of the administrative-territorial reform, the registration of migrants was transferred from the State Migration Service of Ukraine to the territorial communities. The territorial communities were not ready for this either organizationally, methodically or technically. Although there was a decision of the Cabinet of Ministers of Ukraine to conduct a census in 2023, few people believed it. The President of Ukraine had to confirm this date several times.

The Ministry of Digital Transformation of Ukraine got involved in the organization of the census and began to influence the methodology of the census, although the census does not belong to its functions. The influence was that the idea of conducting a census in 2023 using the data of the registers was being promoted. Simultaneously with the preparations for the 2023 census, the State Statistics Service of Ukraine began preparations for the 2030 census, which was proposed to be conducted only on the basis of registers. Differences in conducting population censuses according to the classical methodology and on the basis of registers are given in table 3.

Table 3. Comparison of the separate stages of population censuses according to the classical methodology and on the basis of registers

| | Population census according to the classical methodology | Population census based on registers |
|---|---|---|
| Census preparation | | |
| Development of the census program | List of questions permitted by law | Limited by available information in the registers and its completeness |
| Conducting a pilot census | One of the mandatory stages | Testing of technology |
| Mapping and address lists | Mandatory | Not required |
| Census tools | Development of forms and instructions | Instructions on how to form data base |
| Estimation of the current population according to the registered data | Mandatory | Not required |
| Recruitment and training of census staff | Mandatory | Not required |
| Conducting a census | | |
| Collection of information | Survey | Not required |
| Creating a database with primary information | Data entry, arithmetic and logical control | Automatic |

*Source: compiled by the author*

The stages of developing census results and disseminating the results do not differ. In the intermediate version (conducting a census using registers) the basis is the classical methodology.

With the assistance of the United Nations Fund for Population Activities (UNFPA), in December, 2020 a working group was set up at the State Statistic Service of Ukraine with the participation of international experts to develop a "Roadmap for Ukraine's transition to a census based on registers". It was planned to conduct such a census in 2030. A general conclusion of the group was that Ukraine is not ready for the 2030 census solely on the basis of registers.

The reason is the condition of the registers in Ukraine: the lack of the basic registers, incomplete registers, the lack of analysis of the consistency of the data from different registers, the lack of a single identifier. The group recommended to conduct a 2030 census using registers. At the same time, a Roadmap for Ukraine's transition to a census based on registers was developed. Due to the large number of legal, organizational, methodological and technological issues that need to be addressed in order to conduct a census based on registers, the recommendations do not contain a specific date.

Table 3 shows that the main differences between the census according to the classical methodology and on the basis of registers are in the first stages, namely: preparatory work and the formation of the database. When forming a database during the census according to the classical methodology, information is obtained directly from the population. According to Article 6 of the Law of Ukraine “On the All-Ukrainian Population Census” (Law 2000), census documentation is filled in on the basis of information received from respondents without its documentary confirmation. When conducting a census on the basis of registers, the database is formed using information from various registers according to the developed technology. Strictly speaking, a census based on registers is not a census. It combines existing information from different sources. Wherein the information in registers is the information legally (documentary) confirmed. Difference in the status of information (without documentary evidence and documented) is the main difference between censuses. This causes differences in the possibility to obtain and interpret the results (table 4).

The benefits of the register-based census are usually:

– reduction of spending per unit of population;

– quickness of the census;

– solving issues related to the refusal of the respondents to participate in the census.

However, the question arises, what information the users need: real or formal? In Ukraine, quite often the actual place of residence is different from the registered one. What information do local authorities need: the actual number of residents or formally registered? At present, issues concerning the ethnic composition of the population and language are important for Ukraine. This information

cannot be obtained from the registers. From the existing registers it is not possible to even formally determine the composition and structure of the family.

Table 4. Status of information obtained by two types of census

| Data collected during the Population Census (Article 5 of the Law) | Population Census according to the classical methodology | Population Census based on registers |
|---|---|---|
| Population category | Available and permanent | Legal |
| Sex | Correct | Correct |
| Age, date of birth | There may be inaccuracies | Legal |
| Place of birth | There may be inaccuracies | Legal |
| Ethnicity | Actual status | Impossible to determine |
| Linguistic features | Actual status | Impossible to determine |
| Citizenship | There may be inaccuracies | Legal |
| Education | There may be inaccuracies | Legal |
| Sources of livelihood | Actual status | Legal |
| Employment | Actual status | Legal |
| Migration activity | Actual status | Impossible to determine |
| Housing conditions (housing characteristics) | Actual status | Legal |
| Housing conditions (including family composition) | Actual status | Formal status |
| Composition and family relations of household members | Actual status | Formal status |
| Marital status | Actual status | Legal |

*Source: compiled by the author.*

# Conclusion

Given the real situation (Russian-Ukrainian war; problems with completeness and consistency of the registers; significant uncontrolled migration; the need to obtain information on the actual location of the population), in my opinion, the next census should be conducted exclusively according to the classical methodology. It is better to conduct the census 1–2 years after the end of the war, when migration processes have been stabilized.

In addition to traditional use, the data of the next census are important for the reconstruction of the demographic dynamics and determination of demographic losses. Reconstruction should consist of three stages: 2002–2014, 2015–2022, 2022 – the year of the census.

As for the procedure for conducting the next censuses, it is too early to talk about it.

## References

Law of Ukraine (2000) "On the All-Ukrainian Population Census" https://zakon.rada.gov.ua/laws/show/2058-14#Text

Rudnytskyi O., Levchuk N., O. Wolowyna, and P. Shevchuk (2015). 1932–34 Famine Losses within the Context of the Soviet Union. In: *Famines in European Economic History: The Last Great European Famines Reconsidered*, (eds. Curran D., L. Luciuk, and A. Newby), Routledge (Explorations in Economic History), New York, 192–222.

# Nowcasting Finnish real economic activity: a machine learning approach

**Paolo Fornaro[1] and Henri Luomaranta[2]**

[1] The Labour Institute for Economic Research
e-mail: paolo.fornaro@labore.fi

[2] Statistical Office, Country
e-mail: b.author@office.yy

## Abstract

We develop a nowcasting framework based on micro-level data in order to provide faster estimates of the Finnish monthly real economic activity indicator, the Trend Indicator of Output (TIO), and of quarterly GDP. In particular, we rely on firm-level turnovers, which are available shortly after the end of the reference month, to form our set of predictors. We rely on combinations of nowcasts obtained from a range of statistical models and machine learning methodologies which are able to handle high-dimensional information sets. The results of our pseudo-real-time analysis indicate that a simple nowcasts' combination based on these models provides faster estimates of the TIO and GDP, without increasing substantially the revision error. Finally, we examine the nowcasting accuracy obtained by relying on traffic data extracted from the Finnish Transport Agency website, and find that using machine learning techniques in combination with this big data source provides competetive predictions of real economic activity. The applications of these approaches of utilising surveyed microdata in nowcasting or imputation are discussed. There are multiple possibilities to extend this work in other statistical domains.

**Keywords:** Nowcasting, business surveys, machine learning

## References

Paolo Fornaro and Henri Luomaranta 2020. *Nowcasting Finnish real economic activity: a machine learning approach.* Empirical Economics 58(1):1-17

# The Population Census 2021 in Latvia

**B. Zukula[1] and J.Jukams[2]**

[1] Central Statistical Bureau of Latvia
e-mail: baiba.zukula@csp.gov.lv

[2] Central Statistical Bureau of Latvia
e-mail: janis.jukams@csp.gov.lv

## Abstract

A totally different approach from the traditional census is the register-based census that was developed by the Nordic countries in the 1970s. Denmark was the world's first country to conduct a fully register-based population and housing census in 1981. Under this approach there is no direct collection of data from the population, and the traditional enumeration is replaced by the use of administrative data held in various registers (population register, building/address register, social security register, etc.) through a matching process, making use of personal identification numbers. Information not available in registers is imputed. This approach permits the production of census data at a greatly reduced cost and with relatively limited manpower, once a good quality system of statistical registers has been established (UNECE, 2018).

In 2021, for the first time, a solely register-based census with a reference date of January 1, 2021 was conducted in Latvia.

First, for population estimation, a method developed in 2012 and based on logistic regression model was used to classify all residents in the Population register into two classes - de facto residents and de facto non-residents of Latvia (Ainārе et al., 2021). The dependent variable in the model was derived from Census 2011 data and is a binary variable with values 1 denoting de facto residents and value 0 denoting de facto non-residents respectively. Independent variables or predictors in the model are also binary variables describing individual's features (204 regressors in total) and were derived from several administrative data sources corresponding as close as possible to the year 2011. Population and Housing Census 2011 showed that usually resident population of Latvia accounted for 2 074.6 thousand at the beginning of 2011, which was 7% fewer people than in the Population register supervised by the Office of Citizenship and Migration Affairs (2 228.0 thousand). The model has been used for population estimation at CSB Latvia since 2012 and only in a few cases re-estimation of the model's parameter values has been done (e.g. in cases when regressors should be dropped from the model due to changes in administrative registers which (if changes in register occur frequently) can be a problem for supervised approach in general).

Next step, the availability, quality and reliability of different registers and administrative data sources was evaluated. Latvia has many high-quality, regularly updated registers and databases with a good coverage, such as the Office of Citizenship and Migration Affairs, the State Revenue Service, the State Social Insurance Agency and other state information systems, as well as other registers and databases where information on both individuals and dwellings is available. One major factor that facilitates the statistical use of administrative data records is the use of unified identification systems across different sources (Tønder, 2008). Single personal identifier (personal identification number) is used in all registers and databases. It allows data linkage to provide better data coverage, develop methods that are based on data from

several sources. Since 2012, work was started to improve cooperation with Latvia's largest administrative registers and to develop appropriate methodology for preparing all mandatory census indicators. Before deciding whether to use administrative registers in Censuses, it is necessary to develop methods for assessing the quality of registers, their metadata and data. The starting point for the quality assessment is the common statistical quality framework which has been developed at the level of the European Statistical System and the United Nations Statistical Commission (Dygaszewicz, 2020).

The necessary data for census variables in administrative data sources were identified at the level of persons and dwellings (addressing codes) to ensure the preparation of individual level indicators. This cooperation between institutions contributes to improving the quality of registers and databases, making the use of these data more efficient for the public.

During the preparation process it was acknowledged that the state's administrative registers alone cannot provide all the information required for the census as certain groups were missing or information was incomplete. Therefore, data from non-governmental institutions (e.g., artist unions) and private companies (e.g., water and sewerage service companies) were gathered.

In the end, information from 34 different registers and information from municipalities on demolished and uninhabitable residential buildings, as well as information from municipal companies on the provision of district heating, district water/sewerage (data from more than 70 companies) were used for census 2021. When conducting a Census, the essential features of a population and housing census defined by the International Conference of Statisticians (1853) and redefined and highlighted by the Conference of European Statisticians (2015) have to be taken into account (UNECE, 2018). The mentioned features must be ensured regardless of the census methodology used. It enables countries to provide internationally comparable data.

All five essential features of a Census were ensured in Census 2021 in Latvia, i.e.

| | |
|---|---|
| **Individual enumeration** | The principle of individual enumeration is a fundamental feature for any census of population. In the case of register-based censuses it is important that each census unit (i.e. individual or dwelling) has a special, uniquely identified record in the registers used. For Census 2021 the information on each person and each dwelling was recorded separately in the database, using individual personal and dwelling codes. It is possible to link individuals with dwellings (address code for both is used). |
| **Simultaneity** | The information on persons and dwellings obtained for Census 2021 from administrative data sources corresponds to a certain reference period or a specific point in time (critical moment). 1 January 2021 was used. |
| **Universality (within a precisely defined territory of a country)** | Census data must cover the whole country and all population groups. In Census 2021 Population Register (covering all individuals) and the State Land Service Real Estate Cadastre (State Immovable Property Cadastre) Information System (covering all dwellings) were used. |
| **Small area data** | Census 2021 provides geo-referenced data on the number of inhabitants and dwellings and different variables, as well as small subgroups of the population. Address coordinates are added to each person and dwelling. It allows to provide very detail data. |
| **Defined periodicity** | An advantage of a register-based census is the opportunity to conduct the Census more often than every ten year. Register data are permanently available and more regularly updated. It allowed to prepare test data already annually on 01.01. before Census reference date. |

Advantages of register-based census:

- Reduced face-to-face interviewing - during the period of Covid-19 restrictions (in 2020 and 2021), the face-to-face interviewing in data collection was stopped. In Latvia, Covid-19 restrictions didn't prevent the timely preparation of Census data in accordance with national and international legislation, that could be a challenge in case of traditional census.
- Reduced respondents' burden - the data and methodology developed will be used not only for the purposes of the Population and Housing census, but also in regular statistical sample surveys, thus reducing further the respondents' burden.
- Data are available faster - population and its indicators have been published already in the 2nd quarter of 2021, family and household indicators - in the 3rd quarter of 2021, housing indicators - in the 4th quarter of 2021, but variables on level of education and economic activity in the second quarter of 2022.
- Three times cheaper - in 2021, the traditional census would cost 10 mln euro, while using only administrative data sources, it costed around 2.9 mln euro.
- More accurate, more relevant data according to the classification (for example, occupation, industry) - in administrative data the employers indicates more precisely the employee's occupation. At the same time in the surveys the respondents use every day language to describe activity for which sometimes challenging to identify classification codes.

Disadvantages of register based census:

- Variables that are not available in administrative data sources - impossible to provide variables that are not in administrative data sources, for example, native language, language spoken at home, religious. However, this information can be acquired via surveys (for example, in Latvia questions on native language and language spoken at home were included in the international migration survey and will be collected in Adult education survey in 2022).
- Not fully covered information - even beside administrative data imputation methods are used, it is impossible to fully cover illegal employment and de facto place of residence of individuals within the country.

As mentioned before Nordic countries already at the end of 20th century proved that register based census is viable alternative for traditional census if country can develop and maintain good quality registers, further improve quality and expand the registration system. CSB Latvia will not return to traditional census. Our next task is to further cooperation with registers, identify new variables in surveys that can be replaced by administrative data sources. Another important aspect regarding population estimation using register/model based methodology is quality estimation - how do we ensure that our estimates are precise? Regular census coverage surveys are unavoidable in the future to ensure the quality of register/model based estimates beyond Census 2021.

Within the European Union many countries have taken the decision that 2021 was the last traditional census. Census information will be used as basis for register based systems that will allow to carry out register based censuses. Countries request that this approach to be taken into account when developing new regulation on population statistics and census. The variables and obligations mentioned have to be based on systems that are fully or will be fully register based in future. It means that only variables available in administrative data sources or created from information in administrative data sources will be available within the census.

**Keywords:** Register based census, census features, logistic regression model, supervised learning.

## References

I.Aināre, M.Liberts, B.Zukula, S.Šulca, J.Vaļkovska, B.Opermanis, A.Jurševskis, K.Lece, A.Ceriņa, J.Breidaks, J.Jukāms, R.Beināre (2021) Method used to produce population statistics. Methodological report, Central Statistical Bureau of Latvia, Riga, Latvia. Available at: https://stat.gov.lv/sites/default/files/Metadati/iedz_Metodologija_ENG.pdf

United Nations Economic Commission for Europe (UNECE) (2018) Guidelines on the use of registers and administrative data for population and housing censuses. New York and Geneva, United Nations. Available at: https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf

Dygaszewicz J. (2020) Transition from traditional census to combined and registers based census. Statistical Journal of the IAOS, vol. 36, no. 1, pp. 165-175, 2020. DOI: 10.3233/SJI-190566

Tønder J.K. (2008) The Register-based Statistical System. Preconditions and Processes. International Association for Official Statistics Conference Shanghai October 14 – 18, 2008. Available at: https://www.fao.org/3/I9360EN/i9360en.pdf

# HOUSEHOLDS AND DWELLINGS FOR REGISTER-BASED CENSUS: A GRAPH-BASED APPROACH

**H. Visk[1], V. Levenko[2], K. Lehto[3], E. Maasing[4], E.-M. Tiit[5]**

[1] Statistics Estonia
e-mail: helle.visk@stat.ee

[2] Statistics Estonia
e-mail: vassili.levenko@stat.ee

[3] Statistics Estonia
e-mail: kristi.lehto@stat.ee

[4] Statistics Estonia, Tallinn University of Technology
e-mail: ethel.maasing@stat.ee

[5] Statistics Estonia, University of Tartu
e-mail: etiit@ut.ee

## Abstract

The 2021 census was the first in Estonia to produce all EU-mandatory census characteristics solely relying on the administrative data. One of the greatest challenges was caused by the low, just 80% accuracy of the place of residence data in the Population Register (Äär, 2017; Sõstra *et al.*, 2019) and its impact on households.

Register-based censuses define household as a set of people living in the same dwelling. When using the place of residence from Population Register to determine households, the resulting household and family statistics suffers from heavy bias towards more lone parent families and less married and consensual union couples' families (Statistics Estonia, 2017).

To obtain better statistics on households and families, we have developed a graph-based method which uses input from multiple administrative registers. We consider the people and addresses as vertices of a graph. A connection between two persons (such as marriage, parenthood, care leave) or a person and a place (such as registered address, real estate ownership) form the edges of the graph. A household is viewed as a subgraph containing household members and their dwelling. Then, determining households and their dwellings equates to finding densely connected subgraphs, in other words, to community detection.

To find connections between people or people and places we used data from 17 administrative registers. Each edge in the graph was assigned a weight describing the probability of people living in the same household or a person living on an address. The probability models were fitted on existing household data from Estonian Social Survey and Labour Force Survey.

The new framework was used to compute place of residence in the census. In the presentation, we will give an overview of the evolution of the register-based households methodology in Statistics Estonia with the focus on graph-based approach.

**Keywords:** register-based census, households, graphs, place of residence, population register

## References

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

Sõstra, K. *et al.* (2019) *Leibkondade kontrolluuringu (LEKU) tulemuste analüüs*. Statistikaamet. Available at: https://www.stat.ee/sites/default/files/2020-11/Leibkondade%20kontrolluuringu%20tulemused%202018.pdf (Accessed: 9 May 2022).

Statistics Estonia (2017) *Estonia's first register-based pilot census*. Available at: https://www.stat.ee/sites/default/files/2020-07/Results%20of%20the%20first%20pilot%20census%202016%20%281%29.pdf (Accessed: 3 June 2022).

# ESTIMATION OF INCOME INEQUALITY INDICATORS

**O. Baškauskas**[1] **and D. Krapavickaitė**[2]

Vilnius Gediminas Technical University, Lithuania
e-mail: [1] ovidijusb11@gmail.com [2] danute.krapavickaite@gmail.com

## Abstract

Income inequality is observed in any country. There are many socio-demographic indicators characterizing population income. They are needed not only for the whole country, but also for its domains.

If a direct design-based estimate of the population parameter does not reach the accuracy required then the domain/area is called small. Otherwise it is regarded as large.

The small area estimates are not used in the work of Statistics Lithuania yet. This presentation is devoted to estimate two poverty indicators in the population and small areas:

– proportion of individuals "at risk of poverty" or at-risk-of-poverty rate;

– intensity of the poverty measured by poverty gap.

Following Guadarrama et all., 2014, these indicators are calculated using open data source of Statistics Lithuania. Sample sizes in the domains are not very small because not very detailed open data sets are used.

The poverty indicators are estimated using direct estimator, Fay-Heriot estimator, synthetic post-stratified estimator and sample size dependent composit estimator. Their mean squared errors are estimated and the results obtained show that the performance of the Fay-Heriot estimator is the best. It shows the highest accuracy for areas with the smallest domain sizes. Two estimators for the mean squared error (Rao et all., 2015) of the sample size dependent composit estimator are applied and their results are logically interdependent.

The results of this study are designed to Statistics Lithuania with the wishes to implement small area estimation methods in their statistical production.

**Keywords:** small area estimation, Fay-Heriot estimator, synthetic estimator, composit estimator, mean squared error.

## References

Guadarrama M.; Molina I.; Rao J. N. K. (2014) A comparison of small area estimation methods for poverty mapping. *Statistics in Transition*, **17**(1), pp. 41-66. 945-970.

Molina I., Marhuenda Y. (2020). Package 'sae', The Comprehensive R Archive Network, `https://cran.r-project.org/`

Rao J. N. K.; Molina I. (2015) *Small Area Estimation*, Wiley, Hoboken.

Statistics Lithuania. (2019) *Survey on income and living conditions.* Open data sources. `https://open-data-sets-ls-osp-sdg.hub.arcgis.com/` [accesed 2022-04-07].

# STATISTICAL ANALYSIS WITH MISSING DATA

**Yana Bondarenko**

Oles Honchar Dnipro National University, Ukraine
e-mail: yana.bondarenko@pm.me

## Abstract

Research of advanced techniques for processing multidimensional missing data is presented. The theoretical part of study is focused on the review of the different data imputation methods to handle missing data. The practical part of study is presented machine learning algorithms such as random forest, logistic regression, and nearest neighbor method to solve classification problem for data with artificial and natural missing values.

**Keywords:** missing value, imputation method, machine learning algorithm, accuracy

**Problem formulation.** Data matrix $X^{(n\times d)}$ is specified. Some of the values are not observed. It is necessary to restore data matrix for the application of machine learning algorithms to solve classification problems.

**Removing objects with missing values.** One of the simplest ways to solve problem of processing missing values is to remove the objects (rows of data matrix) that have missing values. This method is used when a small number of objects is not observed. The loss of information is the only drawback of this method. Removing features that have a small number of missing values is an alternative method.

**Imputation with a special value.** The second easiest method is to impute the missing data with a special predefined value (for example, zero or minus one). This approach does not reduce the sample size, but it can add to data matrix values that are significantly different from the real ones. It is appropriate to impute missing values with a value that does not occur in the data matrix (for example, minus one for non-negative values) for further application of machine learning algorithms based on decision trees and impute missing values with zero for algorithms that are sensitive to the features scale.

**Mode imputation.** The third easiest method is to impute the missing values of categorical feature by mode of the non-missing values of that feature.

**Mean imputation.** The fourth simplest method is to impute the missing values of numerical feature by mean of the non-missing values of that feature.

**SVD imputation.** Let's consider the method of replacing missing values using the singular value decomposition of data matrix. First, missing values are replaced by mean of the non-missing values for each feature in data matrix, then, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data.

Second, the decomposition of data matrix is found in the form

$$X^{(n\times d)} = U^{(n\times n)} S^{(n\times d)} V^{(d\times d)},$$

where $S^{(n\times d)}$ is a singular matrix (it is diagonal matrix with the roots of the eigenvalues of the matrix $X^{(n\times d)}\left(X^{(n\times d)}\right)^T$ in descending order on the main diagonal). The matrices $U^{(n\times n)}$, $V^{(d\times d)}$ are orthogonal and, in addition, the columns of the matrix $U^{(n\times n)}$ are the eigenvectors of the matrix $X^{(n\times d)}$, and the matrix $V^{(d\times d)}$ can be presented in the form $V^{(d\times d)} = \left(S^{(n\times d)}\right)^{-1}\left(U^{(n\times n)}\right)^T X^{(n\times d)}$.

Third, the first $r$ rows and columns are selected in matrix $S^{(n\times d)}$, and all remaining ones are deleted. The first $r$ most significant singular values are called principal components.

Fourth, data matrix can be restored having selected the first $r$ columns in the matrix $U^{(n\times n)}$, and the first $r$ rows in the matrix $V^{(d\times d)}$:

$$X_{approx}^{(n\times d)} = U^{(n\times r)} S^{(r\times r)} V^{(r\times d)}.$$

Fifth and finally, the values in the places of missing values in matrix $X^{(n\times d)}$ are replaced on the values obtained in the reconstructed matrix $X_{approx}^{(n\times d)}$.

Steps 2, 3, 4 can be repeated for a predefined number of iterations to improve the recovery of the data matrix or to use the quality criteria of matrix recovery by calculating the proximity to 1 for the coefficient of determination:

$$Q(r) = \frac{\sum_{k=1}^{r} \lambda_k}{\sum_{k=1}^{n} \lambda_k},$$

where $\lambda_k$ are the eigenvalues of the matrix $X^{(n\times d)}\left(X^{(n\times d)}\right)^{\mathrm{T}}$. The dependence of the coefficient of determination $Q(r)$ on the number of principal components $r$ allows to evaluate the efficiency of the method. At the end, the reconstructed values in the places of missing values in matrix $X^{(n\times d)}$ are replaced by the nearest unique values for each feature of matrix $X_{approx}^{(n\times d)}$ to preserve the nature of data.

**Nearest neighbor imputation algorithm.** A hypothesis about similar values of features for close objects is proposed. Thus, the missing values of the features for a certain object can be restored using the known values of the features of $k$ nearest neighbors of this object. Let's consider the method of replacing missing values using the nearest neighbor imputation algorithm.

First, the mask of data matrix $X^{(n\times d)}$ from the Boolean variable True (missing value) and False (no missing value) is created. And this mask is applied to find the number of missing values in each object of the data matrix.

Second, the mask of objects from Boolean variables True (missing value in features) and False (no missing value in features) is created. And this mask is applied to creation of matrix of objects $X^{full}$ with non-missing values.

Third, the mask of each object from Boolean variables True (missing values in features) and False (no missing values in features) is created. And this mask is applied to find features with non-missing values in each object in the data matrix.

Fourth, the distances between the objects of $X^{full}$ and the object of $X^{(n\times d)}$ is calculated (it should be noted that the square of the distance between two objects is equal to the sum of the squared distances between them for each feature with non-missing values).

Fifth, the distances are sorted according to the ascending order and the $k$ smallest are selected, hence the $k$ nearest neighbors for each object are found.

Sixth, missing values are replaced by mean for each feature of the $k$ nearest neighbors in each object in the data matrix.

At the end, the mean values in the places of missing values in matrix $X^{(n\times d)}$ are replaced by the nearest unique values for each feature of matrix $X^{(n\times d)}$ to preserve the nature of data.

It is appropriate to use algorithm if there is a large number of objects with non-missing values, otherwise, at first it is necessary to replace missing values by mean for each feature, and after that, the values in the places of missing values are replaced by the nearest unique values for each feature to preserve the nature of data, and finally, the entire data matrix is selected to be the matrix of objects $X^{full}$ with non-missing values.

**Random forest imputation algorithm.** First, missing values are replaced by mean of the non-missing values for each feature in data matrix. Second, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data. Third, prediction for each feature with missing values is made using the random forest algorithm (at the same time, training has been implemented on objects with non-missing values for this feature). Fourth, replacement of missing values in each feature is carried out using the prediction of decision trees composition obtained above. At the end, the values in the places of missing values in data matrix are replaced by the nearest unique values for each feature of matrix to preserve the nature of data.

**Linear regression imputation algorithm.** First, missing values are replaced by mean of the non-missing values for each feature in data matrix. Second, these mean values are replaced by the nearest unique values for each feature to preserve the nature of data. Third, prediction for each feature with missing values is made using linear regression algorithm (at the same time, training has been implemented on objects with non-missing values for this feature). Fourth, replacement of missing values in each feature is carried out using the prediction with linear regression. At the end, the values in the places of missing values in data matrix are replaced by the nearest unique values for each feature of matrix to preserve the nature of data.

**k-means imputation algorithm.** A hypothesis about similar values of features for close objects is proposed. Thus, the missing values of the features for certain object can be restored using the known values of the center of cluster, which owns the object with missing values.

**Initial data and specific features of implementation methodology.** Six different data sets were used to compare the quality of missing data replacement. Three data sets with complete values and three data sets with natural missing values were studied. Machine learning algorithms were applied to solve classification problems.

Complete data sets were used to estimate the performance of machine learning algorithms by selecting a different proportion of missing data. In addition, information about true values of artificial missing value allows to compare the recovered data directly. Missing data were created artificially for complete data according to the following scheme: 1) subset of one fourth of the most important features is selected with random forest algorithm (it should be reminded that random forest is able to estimate the importance of features based on the frequency of each feature during decision tree construction). This subset of features is used in all experiments; 2) missing value with a certain probability is created for each value from the subset of selected features so that the proportion of all missing values was the same as the given one.

Complete data sets (AI4I 2020 Predictive Maintenance Dataset Data Set, Banknote Authentication Data Set, Car Evaluation Data Set), as well as data sets with natural missing values (Cargo 2000 Freight Tracking and Tracing Data Set, Cervical cancer (Risk Factors) Data Set, HCC Survival Data Set) were selected from the UCI Machine Learning Repository.

The following parameters for imputation algorithms were used. Imputation with a special value: missing values were replaced with minus one when using a random forest algorithm, and missing values were replaced with zero when using a logistic regression algorithm. SVD imputation: the rank of the matrix $X_{approx}^{(n\times d)}$ is half the number of features, maximum number of iterations is 10. Nearest neighbor imputation algorithm: number of nearest neighbors is 5, space metric is $L_2$. Random forest imputation algorithm: number of decision trees is 10, maximum number of iterations is 3. Linear regression imputation algorithm: maximum number of iterations is 3. k-means imputation algorithm: number of clusters is 8, maximum number of iterations is 3.

**Results.** Data sets with a different proportion (from 2,5% to 15% with a step 2,5%) of missing data were created in one fourth of the most important features selected with random forest algorithm. Accuracy and RMSE were calculated according to 10-fold stratified cross-validation sample.

Dependences of the classification accuracy of recovered data on the proportion of missing values are presented in Figure 1a, Figure 1b, Figure 1c.

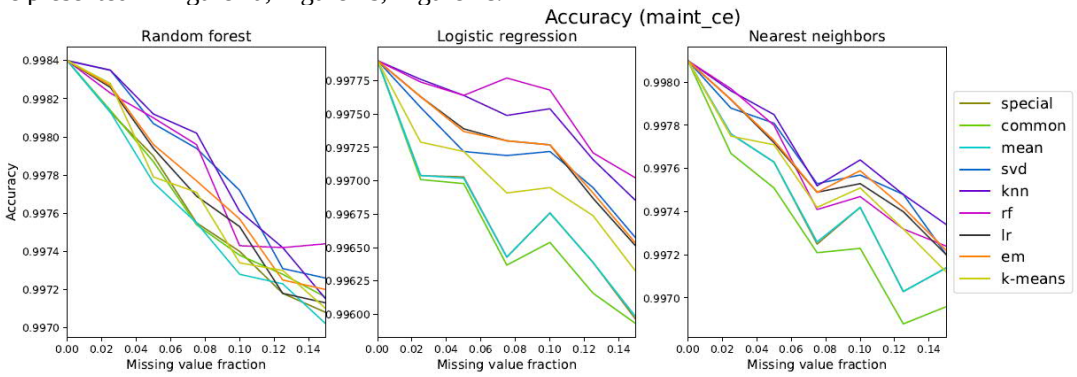


Fig. 1a. Dependence of the classification accuracy of recovered data
on the proportion of missing values (AI4I 2020 Predictive Maintenance Dataset Data Set)

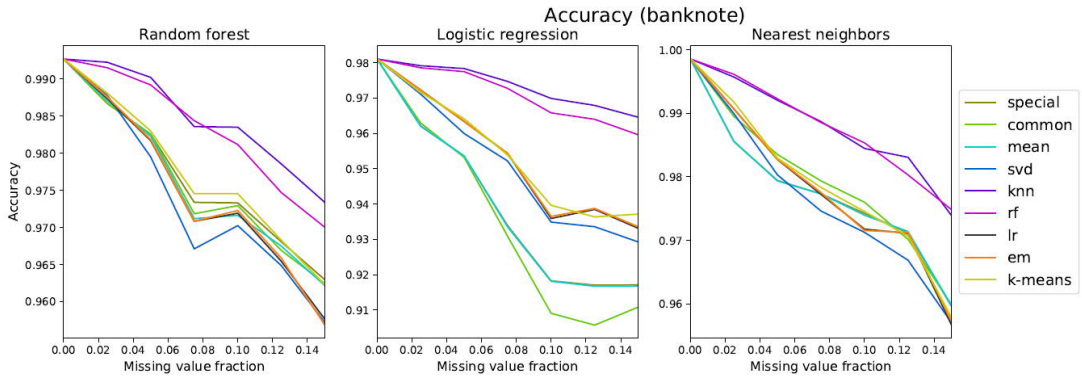


Fig. 1b. Dependence of the classification accuracy of recovered data on the proportion of missing values (Banknote Authentication Data Set)

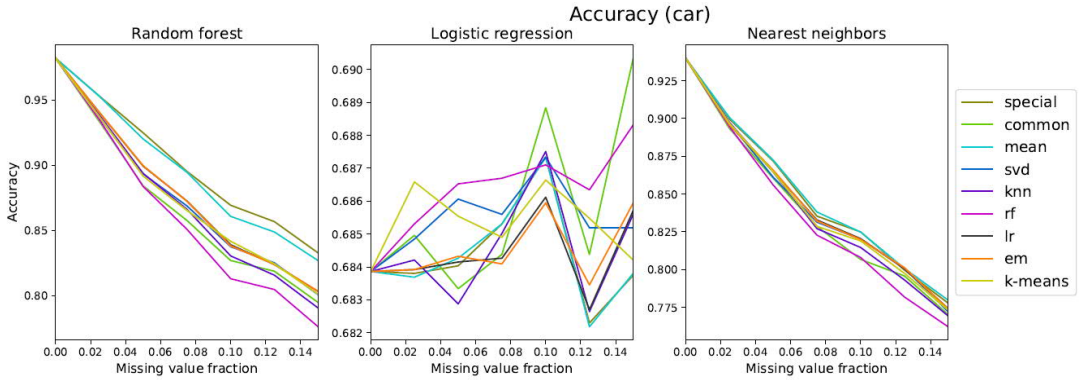


Fig. 1c. Dependence of the classification accuracy of recovered data on the proportion of missing values (Car Evaluation Data Set)

Dependences of the RMSE between recovered and real data on the proportion of missing values are presented in Figure 2.

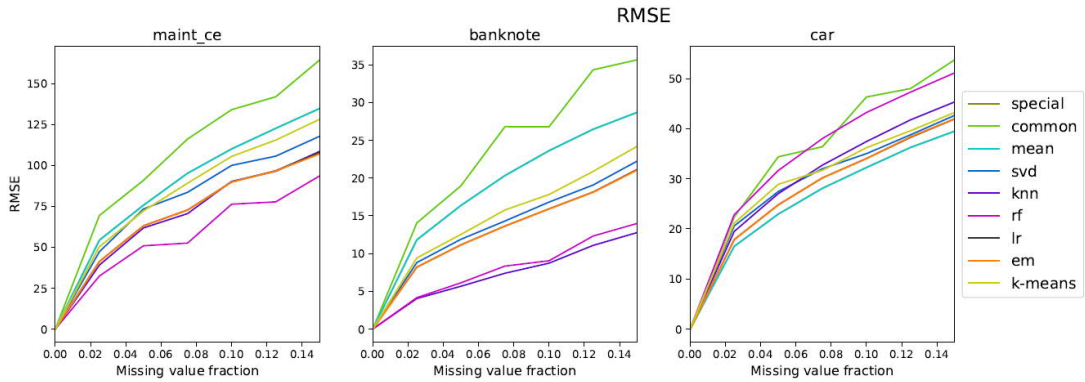


Fig. 2. Dependence of the RMSE between recovered and real data on the proportion of missing values

Performance outcomes for data with natural missing values are shown in Table 1. The best and closest results in each column are highlighted in bold.

Experiments have shown that there is no universal method for missing value replacement, which would be superior in accuracy to all others. Simple imputation methods (such as mode imputation,

mean imputation, imputation with a special value) have demonstrated performance comparable to advanced imputation methods (such as k nearest neighbors, random forest, linear regression, k means) in case of data with natural values.

Table 1. Classification accuracy for data sets with natural missing values

| Datasets | Cargo 2000 | | | Cervical cancer | | | HCC Survival | | |
|---|---|---|---|---|---|---|---|---|---|
| Methods | RF | LR | KNN | RF | LR | KNN | RF | LR | KNN |
| special | **0.9997** | 0.3555 | 0.5630 | 0.9953 | 0.9918 | 0.9918 | 0.7327 | **0.7580** | **0.6669** |
| mean | **0.9997** | 0.3555 | 0.5630 | 0.9953 | 0.9918 | 0.9918 | 0.6900 | **0.7580** | **0.6669** |
| SVD | 0.9903 | 0.6127 | 0.5653 | **0.9965** | 0.9918 | 0.9918 | 0.7323 | 0.7040 | 0.6415 |
| KNN | **0.9997** | 0.5830 | 0.5721 | 0.9930 | **0.9930** | 0.9918 | **0.7463** | 0.7394 | 0.6724 |
| RF | 0.9741 | 0.7813 | 0.5546 | 0.9941 | 0.9918 | 0.9918 | 0.7084 | 0.6970 | 0.6591 |
| LR | 0.9974 | 0.8092 | 0.8488 | 0.9953 | 0.9918 | 0.9918 | 0.7029 | 0.7150 | 0.6661 |
| k-means | 0.9997 | 0.5526 | 0.5052 | 0.9964 | 0.9918 | 0.9918 | 0.7455 | 0.7518 | 0.6536 |

Selection of the imputation method may depend on the types of features with missing values, on the number of objects with missing values, and on the cause of missing values. Each problem requires an individual approach for imputation missing values.

## References

Bache, K. and Lichman, M. (2013) UCI Machine Learning Repository. University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml

Bishop, C.M. (2007) *Pattern Recognition and Machine Learning*. Springer, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Kayumov E. (2016) Imputation methods for missing values. https://github.com/emilkayumov/missing-value

Little, R. J. A., Rubin, D. B (2002) *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge.

VanderPlas J. (2016) *Python Data Science Handbook. Essential Tools for Working with Data*. O'Reilly Media, Inc.

# ORGANISATION AND REALISATION OF SURVEY BY DIFFERENT STAKEHOLDERS (ENTREPRENEURS, EDUCATORS, STUDENTS AND PUBLIC ADMINISTRATORS) ON WORK-BASED LEARNING

**Ilze Brante[1] and Biruta Sloka[2]**

[1]University of Latvia, Latvia
e-mail: ilze.brante@gmail.com

[2]University of Latvia, Latvia
e-mail: Biruta.Sloka@lu.lv

## Abstract

Latvia has accumulated extensive experience on work-based learning. There are several countries (Albania, Armenia, Ukraine, and several more) who are interested to use the experience of Latvia. Therefore there is conducted research on views of different stakeholders involved in work-based learning: entrepreneurs and entrepreneur's organisations, educators, public administrators and students – all are extremely important for successful realisation of work-based learning. Now is development of the questionnaire for all those stakeholders to get advanced situation analysis on advantages, challenges and possible problems on work-based learning. The aim of this report is to investigate the best possible solution for questionnaires design (four different – for all stakeholders) and find the best way to determine the populations (four different) and design the sample to get representative data taking into account that part of questionnaires will be filled by respondents – paper versions and part of the will answer the survey questions using survey platform *QuestionPro.* For questionnaire design it was studied findings reflected in serious sources (Lohr, 2019; Sapsford, 2007; Bryman, 2012; Greenlaw, Brown-Welty, 2009) and it was decided to use 1-10 point scale for evaluation of several aspects by respondents. The designed questionnaires will be tested in pilot survey to find the best possible wording for each questionnaire. Obtained data in the survey are planned to analyse by different statistical indicators: descriptive statistics (indicators of central tendency or location, indicators of variability), cross-tabulations, is planned to test statistical hypotheses with t-test, chi-square test, is planned to perform correlation analysis and factor analysis.

**Keywords:** Questionnaire for different stakeholders; population, sample, evaluation scale

## References:

Bryman, A. (2012). *Social Research Methods***,** 4th edit, Oxford University Press, 766 p.

Greenlaw C., Brown-Welty S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review* 33, 464-480.

Lohr, S.L. (2019). *Sampling: Design and Analysis***,** 2nd edit., Boca Raton, CRC Press, 596 p.

Sapsford**,** R. (2007). *Survey Research***,** 2nd edit., Sage Publications, 276 p.

# NON-PROBABILITY SAMPLE INTEGRATION IN THE SURVEY OF LITHUANIAN CENSUS

**I. Burakauskaitė[1] and A. Čiginas[2]**

[1] Statistics Lithuania, Vilnius, Lithuania
e-mail: ieva.burakauskaite@stat.gov.lt

[2] Statistics Lithuania and Vilnius University, Vilnius, Lithuania
e-mail: andrius.ciginas@stat.gov.lt

## Abstract

The sample of the Statistical survey on population by ethnicity, native language and religion 2021 consists of the voluntary sample and the probability sample drawn from the rest of the census population. A natural post-stratified calibrated estimator tends to underestimate minor religions and other small proportions of interest. Alternatively, to correct the bias of estimates based on the non-probability sample, we evaluate the propensity scores for individuals using the sociodemographic and the previous census data. We show that the combination of the corrected estimates with the calibrated ones improves the estimation accuracy for minor religions.

**Keywords:** population census, missing at random, propensity score, variance estimation.

# SEQUENTIAL RESOURCE ALLOCATION UNDER MULTI-ARMED BANDIT MODEL WITH ONLINE CLUSTERING AS SIDE INFORMATION

**A. Dzhoha[1] and I. Rozora[2]**

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: andrew.djoga@gmail.com

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: irozora@knu.ua

## Abstract

We consider the sequential resource allocation problem under the multi-armed bandit model in the non-stationary stochastic environment. The stochastic multi-armed bandit problem is a classic example of the exploration-exploitation dilemma which is originally presented by Thompson (1933) in the context of clinical trials and later formalized by Robbins (1952). It's a sequential problem defined by a set of actions where at each step, an action is selected, and then a stochastic environment reveals a reward. The goal is to maximize the total reward obtained in a sequence after all steps. Motivated by many real applications, where information can naturally be grouped, we consider a variation of the contextual multi-armed bandit (Bubeck & Cesa-Bianchi 2012) with online clustering representing side information. We assume a stochastic environment, in which the reward of each action conditioned on a cluster follows a Bernoulli distribution with unknown parameters. Additionally, we assume that the nature of the problem changes over time and the clusters drift incrementally making the reward process non-stationary. In this setting, we propose a new algorithm based on a two-stage approach. The first stage is a sequential modification of the traditional k-means clustering algorithm (Duda et al. 2001), in which the algorithm deals with the continuous data stream and acts on a subset of data rather than in a single batch. In the second stage, we incorporate the current information about clusters into the Thompson Sampling algorithm, which is one of the stochastic bandit policies. We introduce a discounting mechanism to track changes in the underlying reward and account for a potential cluster misclassification. We provide the regret analysis of edge cases with supporting numerical experiments for this algorithm.

**Keywords:** multi-armed bandit problem, non-stationary stochastic environment, online clustering.

## References

Bubeck S., Cesa-Bianchi N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122.

Duda R. O., Hart P. E. and Stork D. G. (2001) *Pattern Classification.* John Wiley & Sons.

Robbins, H. (1952) Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, **58**(5), 527–535.

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

# ORGANISATION OF SURVEY ON FINANCIAL LITERACY ASPECTS RELATED TO VOLUNTARY PENSION SAVINGS AND CHALLENGES IN THE REALISATION OF THE SURVEY

**Evija Dundure[1] and Biruta Sloka[2]**

[1] University of Latvia, Latvia
e-mail: dundure.evija@gmail.com

[2] University of Latvia, Latvia
e-mail: biruta.sloka@lu.lv

## Abstract

Recently, people's own decisions have become increasingly important in shaping pension revenues, which is the basis for in-depth and long-term policy-making and academic research of all factors influencing pension planning, as well as data analysis. One of the very important aspects is the willingness of the population to make voluntary contributions to the pension savings, so it is important to ascertain the opinion of the population about their voluntary contribution plans to the future pension. To obtain information on young people's understanding of the readiness to make contributions to pension plans, a survey was conducted. The challenge for this survey is a two-year pandemic that has transformed the views of part of people, as well as the uncertainties caused by the war in Ukraine leading to rising inflation and fluctuations in the financial markets.

The aim of this study is to determine the possibilities of the development factors of financial planning for retirement and to assess their impact on the behavior of the population to answer the question of how to promote better pension planning, which will increase the accumulation of funds for the retirement period.

Tasks of this research is: evaluation of ways on preparation of questions in survey on willingness to contribute for voluntary pension savings on reflecting people views on their readiness and financial literacy aspects; evaluation of ethical aspects on investigation of financial literacy level in the survey; evaluation of ways of realization of the survey and analyses obtained data in survey realization.

The results of this study point out that conducting a survey on future financial stability plans is becoming increasingly problematic in the light of developments in the real political and economic situation on the one hand and limited guarantees of financial stability for voluntary pension savings on the other. The survey has demonstrated that majority of young people in Latvia do not think about future pension savings, results vary depending on education level and employment and has indicated that future challenge for the research is to provide a complete picture of the aspects of pension planning in a particular country.

**Keywords:** survey creation, survey realization, voluntary savings

# USING ADMINISTRATIVE DATA TO CLARIFY AND ADJUST ECONOMIC SURVEY DATA: THE CASE OF LATVIAN HFCS

**Andris Fisenko**[1]

[1] Bank of Latvia, Latvia
e-mail: Andris.Fisenko@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 in all Euro Area countries, as well as some other European Union Member States by the European Central Bank and the national central banks of these countries. The HFCS is conducted at the national level.

In 2020, the Latvian HFCS was conducted in a close cooperation by the Bank of Latvia and the Central Statistical Bureau of Latvia (CSB) already for the third time. CSB ensured the collection of the HFCS data and the adding of respondents' data from several administrative data sources to the survey data. Data editing and imputation was done at the Bank of Latvia.

HFCS is a complex survey, the response level is around 50%. In addition, for many important variables high is also item non-response. Many households do not report their assets, or their report is incomplete. Therefore, editing of faulty or missing survey data is an extremely important stage for improvement of the HFCS micro-data set. Since Bank of Latvia receives from CSB already anonymised HFCS data recontacts with respondents are impossible. For editing of HFCS data we use administrative data that have been added to HFCS data.

The presentation is summary the findings from the experience obtained from editing of the Latvian HFCS data based on administrative data.

## References

Household Finance and Consumption Network (2020) The Household Finance and Consumption Survey: Methodological report for the 2017 wave. *ECB Statistical Paper Series*, 17.

Fisenko, A., Lapiņš, J. (2018) Use of Register Data in Latvian Household Finance and Consumption Survey. *Workshop of the Baltic-Nordic-Ukrainian Network on Survey Statistics 2018*, Central Statistical Bureau of Latvia, Riga, pp.28-29

# QUALITATIVE SURVEY DATA

**K. Lagus[1] and M. Valaste[2]**

[1] University of Helsinki, Finland
e-mail: krista.lagus@helsinki.fi

[2] Universtity of Helsinki, Finland
e-mail: maria.valaste@helsinki.fi

## Abstract

The lack of easy-to-use analysis tools for Survey open text questions poses a challenge for researchers. Survey answers have rich structural and statistical contexts that sometimes extend to population representativeness. The objective here is to design a shared, easy-to-use and statistically sound pipeline for analysing the qualitative text data in surveys and for linking it with central aspects and concepts of the survey data set.

We work simultaneously with several pilot surveys: Tech in eldercare, Loneliness, Child barometers, Faculty social wellbeing, and Everyday life in Varkaus. Two data sets were obtained from the Finnish Social Science Data Archive (FSD), two from research projects in progress, and one is yet to be obtained. Data set sizes vary from below 300 to over 4000 text answers. Various analysis approaches will be explored on these data sets, in order to design the pipeline. We also examine interfaces and data permission processes.

**Keywords:** Textual qualitative survey data

## References

Saari, Juho (Tampere University) & Helsingin Sanomat & Kauhanen, Jussi (University of Eastern Finland) & Karhunen, Leila (University of Eastern Finland) & Lagus, Krista (Aalto University) & Kainulainen, Sakari (Diaconia University of Applied Sciences) & Pantzar, Mika (University of Helsinki) & Erola, Jani (University of Turku) & Junttila, Niina (University of Turku) & Müller, Kiti (Finnish Institute of Occupational Health) & Huhta, Jaana (Finnish Institute of Occupational Health): Helsingin Sanomat Loneliness Survey 2014 [dataset]. Version 1.0 (2020-09-01). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3360

The Office of Ombudsman for Children: Child Barometer 2016 [dataset]. Version 1.0 (2016-12-09). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3134

Office of Ombudsman for Children: Child Barometer 2018 [dataset]. Version 1.0 (2020-02-18). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3307

Office of Ombudsman for Children: Child Barometer 2020 [dataset]. Version 1.0 (2021-05-26). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3497

# CROWDSOURCING SOCIAL WELLBEING NON-PROBABILITY SURVEY

**K. Laine[1], M. Litova[2] and T. Oikarinen[3]**

[1] University of Helsinki, Finland
e-mail: katja.laine@helsinki.fi

[2] University of Helsinki, Finland
e-mail: maria.litova@helsinki.fi

[2] University of Helsinki, Finland
e-mail: tuukka.oikarinen@helsinki.fi

## Abstract

In Spring 2022, The Faculty of Social Sciences at the University of Helsinki piloted a crowdsourcing project to improve social wellbeing at the faculty. The project was originally started for obtaining feedback on the draft of the Code of Conduct, but it was then expanded for open dialogue about equality and social wellbeing at the faculty. The platform has three parts in general. There are two conversation areas, another for commenting on the draft of the Code of Conduct, and the other for general conversation about the topic, and a survey. Both survey and discussion areas data were collected in the *otakantaa.fi* platform.

The survey consists of three question groups: questions on background information, closed questions related to subjective experiences of social wellbeing at the faculty and open-ended text questions covering the topics of social interaction, the role of faculty in wellbeing and equality maintaining and other. The answers to the open-ended questions are presented in Finnish and English.

Survey was advertised on campus, in university events, via email and in social media. The platform was open-access and available for anonymous participation. This led to a worry that due to the sensitive and polarizing nature of the topic the platform would be a target for harassment, inappropriate behaviour or trolling and the survey would have to be closed. Eventually the challenge turned out rather to be gathering sufficient numbers of participants for a meaningful statistical analysis.

**Keywords:** social wellbeing, non-probability survey.

## References

The equality project of the Faculty of Social Sciences. Available at: https://www.otakantaa.fi/fi/hankkeet/740/

FIN-CLARIAH. Qualitative survey data. Available at: https://www.kielipankki.fi/organization/fin-clariah/fin-clariah-2022-06-03/

# MIXED-MODE CENSUS SURVEY IN ESTONIA

**K. Lehto[1] and I. Traat[2]**

[1] Statistics Estonia
e-mail: kristi.lehto@stat.ee

[2] University of Tartu
e-mail: imbi.traat@ut.ee

## Abstract

In 2021 census EU-mandatory census characteristics were collected from administrative data, however the purpose of the sample survey was to collect information on people living in Estonia that is not available in the registers (religious affiliation, knowledge of languages and dialects, existence of a long-term illness or health problem and health-related limitations on daily activities).

The survey design was worked out with collaboration between the Tartu University and Statistics Estonia. The sample design is stratified systematic sampling from dwellings. Stratification is made by local government units. First, people had the opportunity to respond to an online questionnaire during a specific period (CAWI). Then, enumerators received a list with dwellings from where residents did not answer to the questionnaire (CATI/CAPI).

The sample included approximately 40,000 dwellings (around 30,000 of these inhabited), i.e. around 60,000 persons for whom participation in the population and housing census was mandatory according to the law. In CAWI mode all those who wished could respond voluntarily even outside the sample.

CAWI respondents are different from CATI/CAPI respondents. They are younger, healthier, less religious and know more foreign languages based on Census 2011. In order to obtain an unbiased estimates, it is necessary to skillfully combine the data of different modes. The estimate extends the proportion of the surveyed characteristic found in CAWI respondents to CAWI population, and the proportion found in CATI/CAPI respondents to the rest of population.

First results based on Census survey will be published in November 2022.

**Keywords:** census survey, mixed-mode, survey with voluntary part

# Model calibration and MRP methods for small area estimation: an empirical comparison

**Risto Lehtonen**[1] **and Ari Veijanen**[2]

[1] University of Helsinki, Finland
e-mail: risto.lehtonen@helsinki.fi

[2] University of Helsinki, Finland
e-mail: ariveijanen@gmail.com

## Abstract

Multilevel regression and post-stratification (MRP) of Gelman and Little (1997) is widely used for small area estimation with probability and non-probability data in public polls and political sciences, as well as in social and health surveys. The properties of the method are rarely discussed from the design-based inference point of view in the literature; Si (2021) provides a recent exception. We investigate with design-based simulation experiments the finite population properties (bias and accuracy) of MRP for the estimation of proportions of a binary variable for population subgroups or domains (small or large). Our focus is in the capacity of MRP to account for unequal probability sampling. We compare MRP with model calibration (MC) of Wu and Sitter (2001); see Lehtonen and Veijanen (2019) for MC in domain estimation. The synthetic (SYN) estimator acts as another reference.

As a model-assisted method, the MC estimator is design consistent irrespective of the correctness of the model. Model-based MRP and SYN estimators can be severely biased if the sampling information is ignored. Our Monte Carlo experiments showed that the bias in small domains does not necessarily vanish as the domain sample size increased, and the bias can become the dominating component of MSE. However, bias can be successfully reduced by incorporating sampling information into the poststratification cell structure, which indicates design consistency. In small domains, MRP can be more accurate than MC when strong auxiliary information is supplied to the model. The difference in accuracy between MRP and MC reduced as the domain sample size increased. The synthetic estimator was less accurate and more severely biased than MRP. Our limited empirical exercise recommends further studies on the limitations and potentials of MRP in relation to other SAE methods.

**Keywords:** Design consistency; accuracy; Monte Carlo simulation

## References

Gelman A. and Little T.C. (1997) Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, **23**, 127–35.

Lehtonen R. and Veijanen A. (2019) Hybrid calibration methods for small domain estimation. *Statistica & Applicazioni*, **XVII**, 2, 201–235.

Si Y. (2021) On the use of auxiliary variables in multilevel regression and poststratication. https://arxiv.org/abs/2011.00360v2

Wu C. and Sitter R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *JASA*, **96**, 185–193.

# THE RELATIONSHIP BETWEEN THE NUMBER OF REMINDERS AND THE PROPORTION OF FULL RESPONSES IN ONLINE SURVEYS

**R. Moskotina[1] and M. Sydorov[2]**

[1] Taras Shevchenko National University of Kyiv, Ukraine
email: rmoskotina@ukr.net

[2] Taras Shevchenko National University of Kyiv, Ukraine
email: myksyd@knu.ua

## Abstract

Using the reminders is the approach to increase the survey response rate and the number of participants in online surveys. But not all participants complete the survey. So it is important not only to involve respondents in the survey but motivate participants to complete it and reduce item nonresponse. For this purpose reminders also can be used. We would like to find out how the number of reminders could correlate with the proportion of full responses in online surveys. To do this we use data of the monitoring survey UNiDOS (17th wave, November-December 2021). The survey was conducted separately for 1st year students and 2nd + year students (2-4 years of bachelor, 1-5 years of master degree) using LimeSurvey shell. Students received survey invitation by e-mails (5166 invitations for 1st year students and 15244 invitations for 2nd + year students). Then the respondents who did not pass the survey or incomplete filled in the questionnaire received reminders. The maximum number of reminders was 5. There are 31,9% of full responses from 1st year students and 15,5% of full responses from 2nd + year students.

***Table 1.*** *The proportion of full responses depending on number of reminders, cumulative percentage*

| | **1st year students** | **2nd + year students** |
|---|---|---|
| After invitation | 7,6% | 4,1% |
| After first reminder | 17,5% | 10,4% |
| After second reminder | 25,3% | 14,1% |
| After third reminder | 28,1% | 15,0% |
| After fourth reminder | 31,9% | 15,5% |
| After fifth reminder | 31,9% | 15,5% |

As we can see from Table 1, 7,6% of full responses from 1st year students and 4,1% of full responses from 2nd + year students are received without reminders. The first reminder increases the percentage of full responses from 1st year students and 2nd + year students by 9,9% and 6,3% respectively. The second reminder increases the proportion of full responses from 1st year students and 2nd + year students by 7,8% and 3,7% respectively. The third reminder raises the proportion of full responses from 1st year students and 2nd + year students by 2,8% and 0,9% respectively. The fourth reminder raises the percentage of full responses from 1st year students and 2nd + year students by 3,8% and 0,5% respectively. And the fifth reminder no longer increases the proportion of full responses.

Each subsequent reminder tends to reduce the increase of full responses. The 2nd + year students were less motivated to complete the survey than 1st year students. The latter more often completed the survey without reminder, a little less often ignored the reminders. Also the proportion of full responses from 1st year students is more than the proportion of full responses from 2nd + year students. Thus, on the one hand reminders really increase the proportion of full responses in online surveys. On the other hand more motivated respondents respond to reminders somewhat better than less motivated ones.

**Keywords:** reminders, online surveys, full responses.

## References

R Core Team (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. Retrieved 22/06/2022 from https://www.R-project.org/

*Research reports*. Retrieved 22/06/2022 from http://unidos.univ.kiev.ua/?q=en/node/10

# USAGE OF NON-PROBABILITY SAMPLE AND SCRAPED DATA TO ESTIMATE PROPORTIONS

**V. Nekrašaitė-Liegė[1,2], A. Čiginas[1,3] and D. Krapavickaitė[2]**

[1] Statistics Lithuania, Lithuania
e-mail: Vilma.Nekrasaite-Liege@stat.gov.lt, Andrius.Ciginas@stat.gov.lt

[2] Vilnius Gediminas Technical University, Lithuania
e-mail: Vilma.Nekrasaite-Liege@vilniustech.lt, Danute.Krapavickaite@vilniustech.lt

[3] Vilnius University, Lithuania
e-mail: Andrius.Ciginas@mif.vu.lt

## Abstract

An increasing amount of data sources suggests a task to integrate them with the ordinary data sources used in official statistics. One of the problems under the study at Statistics Lithuania is to revise some indicators and to find out if there is room for their accuracy improvement using data from additional sources. The proportion of companies possessing the websites is one such indicator. Traditionally it is estimated using the data of the Information and Communication Technology sample survey.

Information about enterprise website possession is provided also by a private company. However, this data source is updated on a voluntary basis and has some drawbacks: it does not cover all the population, thus the estimator based on this data source should be biased (Tam and Kim, 2018).

Another way to create a list of enterprises owing the websites is to do it by web scrapping (ESSnet Big Data I, ESSnet Big Data II). Following a common methodology, ten potential URLs are found for each enterprise applying a search engine to the population. A logistic regression model is used to estimate the probability, that the selected URL is a website of the particular enterprise. If this probability reaches the fixed threshold, then a conclusion, that the enterprise owns the website, is made. Otherwise, the conclusion is opposite. However, it is known from other research sources, that the accuracy of such an enterprise classification is around 59-89 percent truthful and depends on a search engine, training sample, etc.

Therefore, it may seem that there is no possibility of renouncing the collection of the data on websites through the ICT survey, however, the combination of different sources may lead to more efficient estimators. See Beaumont (2020), Kim and Tam (2021) and Rao (2021) among others.

In this research, the number of methods to integrate auxiliary data obtained from alternative sources with the survey data for bias adjustment is examined. The integration leads to more efficient estimators in comparison with the estimators based only on the survey data. The accuracy measures of the estimators considered are evaluated.

**Keywords:** Big data, coverage bias, post-stratification, calibration weighting, accuracy estimation.

## References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* **46**(1), 1–28.

ESSnet Big Data I. WP2 led by Monica Scannapieco/ISTAT (OBEC) https://ec.europa.eu/eurostat/cros/content/wp2-webscraping-enterprise-characteristics_en

ESSnet Big Data II. WPC led by Galia Stateva/BNSI (OBEC) https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en

Kim, J.-K. and Tam S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review* **89**(2), 382–401.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya* **B 83** (1), 242–272.

Tam, S.-M. and Kim J.-K. (2018). Big data, selection bias and ethics – an official statistician's perspective. *Statistical Journal of the IAOS* **34**, 577–588.

# CHALLENGES AND SOLUTIONS TO MAINTAINING SURVEY RESPONSE RATES IN SOCIAL STATISTICS

**S. Purona-Sida**[1]

[1] University of Latvia
Central Statistical Bureau of Latvia
e-mail: Sigita.purona-sida@csp.gov.lv

## Abstract

In Latvia, since 2017, when data collection was also introduced via the Internet, the social statistics provide respondents with the possibility to reply in person, by phone, by internet and by completing a paper questionnaire (only HBS).

By 2020, the most significant part of the responses were collected in face-to-face interviews and slightly less than 10% in online interviews. Latvia, like other countries of the European Union, faced the challenge of keeping response rate of the surveys for several decades (Särndal 2005, Tourangeau 2013), while the measures to limit the covid-19 pandemic provided for the complete cessation of face-to-face interviews for nearly two years.

In order not to stop collecting social statistical data and to ensure the volume and structure of data that meets the quality requirements, the Central Statistical Bureau of Latvia (hereinafter - CSB) took a number of measures to organise surveys and maintain the response rates.

The measures included a complex approach and sustainable solutions. Primarily, in cooperation with several Latvian administrative data holders (Office of Citizenship and Migration Affairs, Road Traffic Safety Directorate and State Revenue Service), an agreement was reached on obtaining the contact information of respondents – phone numbers and e-mails. Secondly, the CSB refocused the interviewers service by providing operationally a load redistribution between face-to-face interviewers and telephone interviewers. A training and adapted working tools were performed.

As a result, around 80% of phone numbers and emails were reached to contact respondents. None of the social surveys delayed the deadlines due to field works. However, some respondents did not respond to an invitation to reply via the Internet and some of the contact information failed to obtain. In the surveys where this was possible (the person samples), the CSB redesigned sample from two stage to a one stage sample design and recalculate sample volumes, as a lower number of respondents is sufficient for one stage than for two stage sample. The CSB carried out the Adaptive data Collection approach. The “adaptive” meaning is flexible and active control of collection results, ensuring that the answers in certain groups are well balanced and representative to the population. Thus, it was ensured that respondents to the unrepresented age/sex/groups were surveyed.

The third group refers to communication with respondents, particularly groups that are more difficult to reach – rural residents, young people and men (Woronkowicz 2020). Through the funding available in the grants granted by Eurstat, EU-SILC, LFS, EU-GBV, AES surveys have developed and implemented measures: development of a branding for each survey; communication with the respondent: letters, reminders, booklets; content for home page, social networks, newspapers, TV and radio – inforgraphs, radio news, cartoons; participation in Instagram discussions, broadcasts and podcasts and test for the main news home page delfi.lv.

The pandemic crisis has shed light on that it is possible to find sustainable and cost-effective solutions to maintain response rates in social surveys. The complex approach and operational action have not

only allowed all social surveys to be conducted within the planned deadlines, but even increased the survey response rates.

**Keywords:** response level, social surveys

## References

Särndal C.E., Lundström S (2005) *Estimation in Survey with Nonresponse*. John Wiley & Sons, Ltd, Print ISBN:9780470011331

Tourangeau, R., and Plewes, T.J. editors (2013) *Nonresponse in Social Science Surveys: A Research Agenda*. National Academy of Sciences, Washington.

Woronkowicz J., Hale, J.S. et.al. (2020) 'You're Not My Friend': Communication Style, Sponsor Salience, and Gender in Recruitment Messaging *Survey practice*, Vol 13, Issue, 2020, June 16, 2020, https://doi.org/10.29115/SP-2020-0005

# TEACHING ASPECTS ON SURVEYS, ON QUESTIONNAIRE DESIGN, ON PILOT SURVEY, ON SAMPLE SELECTION AND DATA COLLECTION USING QUESTIONPRO AND OBTAINED DATA ANALYSIS WITH SPSS

**Biruta Sloka**[1]

[1]University of Latvia, Latvia
e-mail: Biruta.Sloka@lu.lv

## Abstract

There are requirements for students on all study levels (bachelor, master and doctoral) to have empirical research and very often students use survey designed by themselves and developed by them questionnaire for their empirical research. The aim of this report is share the practical experience in teaching aspects on surveys and questionnaire design, to discuss different aspects on best possible and reliable information source for studies. It is very good that Association of Statisticians of Latvia organise readings (lasījumi) on important aspects held by different specialists and the information of this reading is available on association wepage (Association of Statisticians of Latvia (2022). This really is an inspiring source to get information on latest developments also in survey organisation (Bank of Latvia, 2022; Official Statistics Portal of Latvia, 2022). It is good source for teaching staff and for students. For students there are used also other sources, also platform *Sage Research Methods*, available for students and staff of University of Latvia with registration of their student/staff ID. On the platform there are many sources on different aspects for surveys, including questionnaire design, sample creation, data analysis, etc. For students on different aspects of survey organisation and analysis aspects there are recommended several sources for individual studies (Lohr, 2019; Sapsford, 2007; Bryman, 2012; Greenlaw, Brown-Welty, 2009). It is good that students for their surveys can use survey platform *QuestionPro*, on classes there are discussed aspects on different evaluation scales use and application of SPSS for data analysis with different statistical indicators: descriptive statistics (indicators of central tendency or location, indicators of variability), cross-tabulations, use of tests of statistical hypotheses with t-test, chi-square test, is planned to perform correlation analysis and factor analysis. For those exercises there are very useful datasets of Labour Force Survey, EU-SILC, and other available at Official Statistics Portal of Latvia. Tasks in testes for students are given from those datasets.

**Keywords:** Questionnaire design, population, sample, evaluation scale, metadata by Official Statistics Portal of Latvia

## References:

Association of Statisticians of Latvia (2022). Readings (Lasījumi), available https://www.statistikuasociacija.lv/?cat=9 [accessed 17.06.2022]

Bank of Latvia (2022). Household and Finance Consumption Survey, available https://www.bank.lv/en/statistics/stat-data/hfcs [accessed 20.06.2022]

Bryman, A. (2012). *Social Research Methods***,** 4th edit, Oxford University Press, 766 p.

Greenlaw C., Brown-Welty S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review* 33, 464-480.

Lohr, S.L. (2019). *Sampling: Design and Analysis*, 2nd edit., Boca Raton, CRC Press, 596 p.

Official Statistics Portal of Republic of Latvia (2022). Use of ICT in households, metadata, available at https://stat.gov.lv/en/metadata/5864-use-ict-households#stat_pres [accessed 20.06.2022]

Sage Research Methods (2022). Database – platform available for students and staff of University of Latvia (with registration indicating student/staff ID)

Sapsford, R. (2007). *Survey Research*, 2nd edit., Sage Publications, 276 p.

# COMPILATION OF ACTIVITY STATUS AND EMPLOYMENT VARIABLES IN ESTONIAN REGISTER-BASED CENSUS

**Kaja Sõstra**[1]

[1] Statistics Estonia
e-mail: kaja.sostra@stat.ee

## Abstract

Population and housing census includes several variables related with person's activity status and employment: current activity status, status in employment, occupation, industry, location of place of work. Activity status is the most challenging census characteristic because it changes frequently over time, and there is no register comprising updated information on all activity status components. Several databases should be used in the activity status algorithm. Concepts and definitions in administrative register often differ from statistical definitions. Therefore, the compilation algorithm includes complicated linking procedures, numerous logical checks and rules for adjusting register information to the statistical needs.

Pursuant to the Census Regulation (Regulation (EU) 2017/543), current activity status of persons aged 15 and over has the following breakdowns: employed, unemployed, pension or capital income recipients, students and others. Main data sources of the activity status and other employment variables are Employment Register, Tax declarations in Register of Taxable Persons, Register of Persons Registered as Unemployed and Job-Seekers, Social Services and Benefits Registry, Estonian Education Information System etc.

Firstly, the algorithm was elaborated for the first pilot census (Muusikus, Lehto, 2018). Separate lists of employed, unemployed, pensioners and students are prepared using different registers. Activity status lists are linked with the list of permanent residents compiled based on the residency index methodology (Maasing et al, 2017). Each person is ascribed one activity status according to the order of priority of activity statuses established in the Census Regulation. Other employment variables are compiled for employed persons. Process of further development and adaptation of the algorithms for register-based population and housing census 2021 will be presented.

**Keywords:** register-based census, current activity status.

## References

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.

Muusikus, M. Lehto, K. (2018) Compilation of activity status based on register data. *Eesti Statistika kvartalikiri* 1/18, *Quarterly Bulletin of Statistics Estonia*, 12-20.

Maasing, E., Tiit, E.-M., & Vähi, M. (2017). Residency index – a tool for measuring the population size. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 21 (1), pp. 129−139

# STATVILLAGE HOMEWORK

**Donatas Šlevinskas**

[1] Vilnius Gediminas Technical University, Lithuania
e-mail: donatas.slevinskas@stud.vilniustech.lt

## Abstract

During the course of sampling methods I was given a homework about hypothetical village in Canada called StatVillage. Here we try to estimate various parameters of the population variable using different sampling and estimation techniques (ex. stratified random sampling, poststratification, single-stage cluster sampling, ratio estimator). The main task is to get estimates with a low variance. In addition, missing values of numerical variables are filled in using regression trees.

**Keywords:** stratified random sampling, ratio estimator, regression trees.

## References

Krapavickaitė, D. and Plikusas, A. (2005) *Imčių teorijos pagrindai*. Technika, Vilnius.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013*) An Introduction to Statistical Learning: with Applications in R.* Springer, New York.

Schwarz, C. J. (1997) StatVillage 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(**2**).** http://jse.amstat.org/v5n2/schwarz.supp/index.html

# REGISTER-BASED POPULATION AND HOUSING CENSUS IN LITHUANIA

**Milda Šličkutė-Šeštokienė**[1]

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

## Abstract

Statistics Lithuania, together with all EU countries, are constantly moving towards higher usage of administrative data sources while diminishing the reliance on traditional statistical data collection methods. Population and Housing Census is one of examples which moved from traditional approach in 2001 towards combined approach in 2011 and into completely Register-based approach in 2021. This transition was challenging but also very valuable experience which resulted in higher quality of the results with significant lower budget.

The main aspects of transition towards register-based census will be presented showing the sources and methods of the register-based census as well as the key results of Census 2021.

**Keywords:** census, register-based, administrative data sources.

## References

United Nations Economic Commission for Europe (UNECE) (2018) *Guidelines on the use of registers and administrative data for population and housing censuses*. New York and Geneva, United Nations. Available at: https://unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20184.pdf

Britt Wallgren and Anders Wallgren, (2014) *Register-based Statistics: Statistical Methods for Administrative Data Title of the Book*. Wiley Series in Survey Methodology, City.

# CHALLENGES FOR REAL SURVEY DEVELOPMENT AND ORGANISATION

**Rita Vanaga[1] and Biruta Sloka[2]**

[1]University of Latvia, Latvia
e-mail: ritava@inbox.lv

[2]University of Latvia, Latvia
e-mail: Biruta.Sloka@lu.lv

## Abstract

There are several financing models of finance regulators and many of them advantages and have problems. Therefore it was conducted research to investigate the best possible financing model could be applied for Latvia. The aim of this report is to share the experience and exchange opinions on survey organisation and questionnaire design for the survey. Population has been defined as close community having relevant education and experience (314 units) – real participants in finance market. It was decided to invite every third from the list, invitation was done with personal approach, it was three times reminder after two weeks for those who have not responded yet. It was received 95 responses. For questionnaire design it was studied experience of many countries around the globe and it was decided to use 1-10 point scale for evaluation of several aspects by respondents. The designed survey has been tested in pilot survey and several questions (3 from 22) were adjusted as result of pilot survey. Obtained data were analysed by different statistical indicators: descriptive statistics, cross-tabulations, was tested statistical hypotheses with t-test, chi-square test, correlation analysis and factor analysis.

**Keywords:** Questionnaire development and testing; population, sample, survey data collection

## References:

Bryman, A. (2012). *Social Research Methods***,** 4th edit, Oxford University Press, 766 p.

Greenlaw C., Brown-Welty S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review* 33, 464-480.

Lohr, S.L. (2019). *Sampling: Design and Analysis***,** 2nd edit., Boca Raton, CRC Press, 596 p.

Sapsford**,** R. (2007). *Survey Research***,** 2nd edit., Sage Publications, 276 p.

Saeima – Parliament of Republic of Latvia (2000). Law on the Financial and Capital Market Commission, accepted 01.06.2000.

Supreme Court of the Republic of Latvia (2019). Summary of Court Practice in Cases of the Financial and Capital Market Commission (2006-2018), [Online] Available at http://www.at.gov.lv/en/judikatura/tiesu-prakses-apkopojumi/administrativas-tiesibas, [Accessed 20.06.2022].

Supreme Court of the Republic of Latvia (2015). Decision of 12 February 2015 in case no. Item 6 of SKA-563/2015 (A43017813).